Nonconvex Optimization and Model Representation with Applications in Control Theory and Machine Learning

Yue Sun

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Maryam Fazel, Chair

Mehran Mesbahi

Sewoong Oh

Program Authorized to Offer Degree:
Electrical Engineering

University of Washington

Abstract

Nonconvex Optimization and Model Representation with Applications in Control Theory and Machine Learning

Yue Sun

Chair of the Supervisory Committee:
Maryam Fazel
Electrical and Computer Engineering

In control and machine learning, the primary goal is to learn the models that make predictions or decisions and act in the world. This thesis covers two important aspects for control theory and machine learning: the model structure that allows low training and generalization error with few samples (i.e., low sample complexity), and convergence guarantees for first-order optimization algorithms for nonconvex optimization.

If the model and the training algorithm apply the knowledge of the structure of data (such as sparsity, low-rankness, etc.), the model can be learned with low sample complexity. We present two results, the Hankel nuclear norm regularization method for learning a low order system, and the overparameterized representation for linear meta-learning.

We study dynamical system identification in the first result. We assume the true system order is low. A low system order means that the state can be represented by a low dimensional

vector, and the system corresponds to a low rank Hankel matrix. The low-rankness is known to be encouraged by nuclear norm regularized estimator in matrix completion theory. We apply a nuclear norm regularized estimator for Hankel matrix, and show that it requires fewer samples than the ordinary least squares estimator.

We study linear meta-learning in the second part. The meta-learning algorithm contains two steps: learning a large model in representation learning stage, and fine tuning the model in few-shot learning stage. The few-shot dataset contains few samples, and to avoid overfitting, we need a fine-tuning algorithm that uses the information from representation learning. We generalize the subspace-based model in prior arts to Gaussian model, and describe the overparameterized meta-learning procedure. We show that the feature-task alignment reduces the sample complexity in representation learning, and the optimal task representation is overparameterized.

First order optimization methods such as gradient based method, is widely used in machine learning thanks to its simplicity for implementation and fast convergence. However, the objective function in machine learning can be nonconvex, and the first order method has only the theoretical guarantee that it converges to a stationary point, rather than a local/global minimum. We dive into more refined analysis of the convergence guarantee, and present two results, the convergence of perturbed gradient descent approach to a local minimum on Riemannian manifold, and a unified global convergence result of policy gradient descent for linear system control problems.

We study how Riemannian gradient converges to an approximate local minimum in the first part. While it is well-known that the perturbed gradient descent escapes saddle points in Euclidean space, less is known about the concrete convergence rate when we apply Riemannian gradient descent on the manifold. In the first result, we show that the perturbed Riemannian

gradient descent converges to an approximate local minimum and reveal the relation between convergence rate and the manifold curvature.

We study the policy gradient descent applied in control in the second part. Many control problems are revisited under the context of the recent boom in reinforcement learning (RL), however, there is a gap between the RL and control methodology: The policy gradient in RL applies first-order method on nonconvex landscape, and it is hard to show they converge to global minimum, while control theory invents reparameterization that makes the problem convex and they are proven to find the globally optimal controller in polynomial time. Targeting on interpreting the success of the nonconvex method, in the second result, we connect the nonconvex policy gradient descent applied for a collection of control problems with their convex parameterization, and propose a unified proof for the global convergence of policy gradient descent.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Chapter 1

INTRODUCTION

In the last few decades, we have witnessed the power of machine learning models which extracts the useful information from the data, and accomplishes variant difficult tasks based on the learnt information. Machine learning and control (also revisited by reinforcement learning) both aim to learn models that predict and act on the real world. A machine learning model can be as simple as a linear map, which is trained by solving a linear regression problem on features and labels. In recent years, more complicated models are investigated. These models perform remarkably well in many applications, such as robotics, image classification, objective detection, machine translation, recommendation systems, etc. Although these models behave well in practice, we do not have a good theoretical understanding of these methods. In this work, we aim to study two components raised in learning problems:

1. What is a good formulation of the machine learning model, so that it is trained with low training-generalization error with few samples.
2. How to train the machine learning model in a computationally efficient way with convergence guarantee.

The first challenge means that, it is important to define the correct model for machine learning tasks, and see how the structure of models helps learn with few data, i.e., improves the statistical rates. In the system identification and representation learning applications below, we hope to learn "simple" models that represents the real-world tasks. Both of them involve low rank structure of models. We propose that, the low rank structure enables the Hankel nuclear norm regularized algorithm to learn a low-order system on few data with small training and generalization error, and a proper representation of meta-learning model leads to optimal generalization guarantee compared with the SVD based method, which is proven suboptimal.

Besides that, we are interested in the convergence guarantee of optimization methods for training machine learning models. The convergence to global optimum is well studied only for convex optimization, whereas during the recent trend in machine learning, gradient based algorithms are applied for solving non-convex optimization problems, and they empirically learn good models. Thus we are interested in studying the theoretical convergence guarantee of first order algorithms for nonconvex optimization. We study the convergence of Riemannian gradient descent and show the relation of convergence rate and curvature constants, and the global optimality for a special family of nonconvex optimization problem in control theory.

This chapter is a brief introduction of the following chapters, and full introduction and literature review of each theme will be specified in the corresponding chapters.

- Chapter 2 investigates the second order convergence guarantee of gradient based method on Riemannian manifolds. Thanks to recent boom of machine learning, gradient based methods applied to nonconvex problem empirically perform remarkably well. One line among them is optimizing strict saddle functions (Ge et al., 2015; Jin et al., 2017a), where we can find an approximate local minimum in polynomial time. Another line of work is to study the optimization method on a manifold (Absil et al., 2009b), which is generally a nonconvex optimization problem if we trivially treat the manifold as a constraint. One can combine the geometric structure of the manifold and the convex optimization algorithms to obtain the convergence guarantee of the Riemannian gradient method, which is a gradient based method implemented on manifolds. However, the convergence to local minimum for nonconvex optimization problem on manifolds is less studied (convexity is not well defined on manifolds) before. We investigate the convergence rate to an approximate local minimum on an Riemannian manifold, and relate the rate with the curvature constants of the function and the manifold.
- Chapter 3 studies the convergence guarantee of policy gradient descent method for control problems. There is a recent boom in reinforcement learning that revisits the control problems. However, we see a difference in their philosophy that, control theory

has traditionally relied more on building physical models whereas machine learning relies more on data-driven methods. The policy gradient descent algorithm in the RL domain is typically used in the policy space where the costs are usually nonconvex (Fazel et al., 2018). Previous papers in control theory study the convex optimization theory with linear matrix inequalities (LMI) or semidefinite programming (SDP) (Dullerud & Paganini, 2013; Stengel, 1994; Rawlings et al., 2017; Boyd et al., 1994). Motivated by recent papers (Fazel et al., 2018; Mohammadi et al., 2019a; Bu et al., 2019a,b; Furieri et al., 2020; Zhang et al., 2020; Jansch-Porto et al., 2020) that directly study the nonconvex landscape of the linear quadratic regulator (LQR), LQR for Markov jump linear systems(MJLS), robust control and decentralized control problems, we propose an explanation that connects convex analysis in control theory with analysis of policy gradient method, and generalize nonconvex analysis to a broad range of optimal control problems. We show the generality of this idea by covering the results of the aforementioned papers on nonconvex landscape into a single unifying theorem. In summary, we build a bridge between the two methodologies and show the theoretical tools from control theory can help explaining the empirical success of nonconvex methods of reinforcement learning.

- Chapter 4 studies the system identification problem, which belongs to the model based method more usually used before. Previous works such as Oymak & Ozay (2018); Sarkar et al. (2019) use unregularized least squares method to regress the input-output map, however if we do not know the dimension of state space, the train-validation step (in order to find the state dimension) is required and not easy to implement. We study the Hankel nuclear norm regularized formulation, which encourages the simplicity of a linear system by the low-order property, and it reduces the sample complexity requirement while the statistical rate of error is preserved. We propose the statistical property, and a practical training-validation algorithm that tunes the regularizer efficiently.
- Chapter 5 studies the role of overparametrization and dimension reduction in representation learning. It aims to retrieve the principle features of the tasks, which are often

low-dimensional, from limited data available for related tasks. We consider a setup where task features follow a Gaussian distribution in the high dimensional space, whose covariance spectrum has a decaying pattern so they are approximately low dimensional. As mentioned in Kong et al. (2020b,a); Tripuraneni et al. (2020); Du et al. (2020), a low-rank approximation step such as k-SVD is commonly used to retrieve the low dimensional space. We are interested in overparameterized meta-learning. We show that learning large representations by letting directions weighted by their relative importance, although leading to an ill-posed overparameterized problem, can result in the optimal generalization error compared to low dimensional representations. Furthermore, the findings reveal a double descent phenomena when varying the representation dimension, which is typically observed in practical meta-learning.

Chapter 2

FIRST-ORDER METHOD FOR NONCONVEX OPTIMIZATION ON RIEMANNIAN MANIFOLDS

In this chapter, we investigate the convergence to a local minimum using the first order optimization algorithms. It is known that, for solving an unconstrained optimization problem in Euclidean space, the perturbed gradient descent algorithm converges to an approximate local minimum in polynomial time. We analyze the first order optimization algorithm on Riemannian manifold, and show that perturbed Riemannian gradient descent provably converges to an approximate local minimum. We give the concrete convergence rate of perturbed Riemannian gradient descent, which reveals the role of the manifold curvature with respect to the rate.

This work is published as Sun et al. (2019).

2.1 Introduction

We consider minimizing a non-convex smooth function on a smooth manifold $\mathcal{M}$,

$$\min_{x\in\mathcal{M}} \quad f(x), \tag{2.1}$$

where $\mathcal{M}$ is a d-dimensional smooth manifold[1], and f is twice differentiable. We assume the Hessian is ρ-Lipschitz. This framework includes a wide range of fundamental problems (often non-convex), such as PCA (Edelman et al., 1998), dictionary learning (Sun et al., 2017), low rank matrix completion (Boumal & Absil, 2011), and tensor factorization (Ishteva et al., 2011). Finding the global minimum of a nonconvex function is in general NP-hard; our goal

[1] Here d is the dimension of the manifold itself; we do not consider $\mathcal{M}$ as a submanifold of a higher dimensional space. For instance, if $\mathcal{M}$ is a 2-dimensional sphere embedded in $\mathbb{R}^3$, its dimension is $d = 2$.

is to find an approximate second order stationary point with first order optimization methods. We are interested in first-order methods as they are extremely prevalent in machine learning, partly because computing Hessians is often too costly. It is important to understand how first-order methods work when applied to nonconvex problems, and there has been recent interest on this topic since (Ge et al., 2015), as reviewed below.

In the Euclidean space, it is known that with random initialization, gradient descent avoids saddle points asymptotically (Pemantle, 1990; Lee et al., 2016). Lee et al. (2017, §5.5) show that the result above is also true on smooth manifolds, although the result is expressed in terms of nonstandard manifold smoothness measures. Importantly, these works do not give quantitative rates for the algorithm's behavior near saddle points.

Du et al. (2017) shows gradient descent can be *exponentially slow* in the presence of saddle points. To alleviate this phenomenon, if we define $(\epsilon, -\sqrt{\rho\epsilon})$ local minimum as x satisfying $\|\nabla f(x)\| \leq \epsilon$, $\lambda_{\min}\nabla^2 f(x) \geq -\sqrt{\rho\epsilon}$, it is shown that for a β-gradient Lipschitz, ρ-Hessian Lipschitz function, cubic regularization (Carmon & Duchi, 2017) and perturbed gradient descent (Ge et al., 2015; Jin et al., 2017a) converges to $(\epsilon, -\sqrt{\rho\epsilon})$ local minimum in polynomial time, and momentum based method accelerates the convergence (Jin et al., 2017b). We know much less about inequality constraints: Nouiehed et al. (2018) and Mokhtari et al. (2018) discuss second order convergence for general inequality-constrained problems, where they need an NP-hard subproblem (checking the co-positivity of a matrix) to admit a polynomial time approximation algorithm. However such an approximation exists only under very restrictive assumptions. Avdiukhin et al. (2019); Lu et al. (2019b,a) show that, when the negative curvature direction of saddle points always coordinate well with the nonlinear constraints (we omit the exact definitions in the papers), the perturbed projected gradient descent algorithm always converges to an approximate second order minimum. But it is unknown whether this assumption applies to the loss landscape of any well known applications.

An orthogonal line of work is optimization on Riemannian manifolds. Absil et al. (2009a) provides comprehensive background, showing how algorithms such as gradient descent, Newton and trust region methods can be implemented on Riemannian manifolds, together with asymptotic convergence guarantees to first order stationary points. Zhang & Sra (2016) provide global convergence guarantees for first order methods when optimizing geodesically convex functions. Bonnabel (2013) obtains the first asymptotic convergence result for stochastic gradient descent in this setting, which is further extended by Tripuraneni et al. (2018); Zhang et al. (2016); Khuzani & Li (2017). If the problem is non-convex, or the Riemannian Hessian is not positive definite, one can use second order methods to escape from saddle points. Boumal et al. (2016a) shows that Riemannian trust region method converges to a second order stationary point in polynomial time (Kasai & Mishra, 2018; Hu et al., 2018; Zhang & Zhang, 2018). But this method requires a Hessian oracle, whose complexity is d times more than computing gradient. In Euclidean space, trust region subproblem can be sometimes solved via a Hessian-vector product oracle, whose complexity is about the same as computing gradients. Agarwal et al. (2018) discusses its implementation on Riemannian manifolds, but not clear about the complexity and sensitivity of Hessian vector product oracle on manifold.

The study of the convergence of gradient descent for non-convex Riemannian problems is previously done only in the Euclidean space by modeling the manifold with equality constraints. Ge et al. (2015, Appendix B) proves that stochastic projected gradient descent methods converge to second order stationary points in polynomial time (here the analysis is not geometric, and depends on the algebraic representation of the equality constraints). Sun & Fazel (2018) proves perturbed projected gradient descent converges with a comparable rate to the unconstrained setting (Jin et al., 2017a) (polylog in dimension). The paper applies projections from the ambient Euclidean space to the manifold and analyzes the iterations under the Euclidean metric. This approach loses the geometric perspective enabled by Riemannian optimization, and cannot explain convergence rates in terms of inherent quantities such as the sectional curvature of the manifold.

Criscitiello & Boumal (2019) gives a similar convergence analysis to our result for a related perturbed Riemannian gradient method. We point out a few differences:

1. Criscitiello & Boumal (2019) assumes Lipschitzness on the pullback map $f \circ \mathrm{Retr}$. While

this makes the analysis simpler, it lumps the properties of the function and the manifold together, and the role of the manifold's curvature is not explicit. In contrast, our rates are expressed in terms of the function's smoothness parameters and the sectional curvature of the manifold separately, capturing the geometry more clearly.

2. The algorithm in Criscitiello & Boumal (2019) uses two types of iterates (some on the manifold but some taken on a tangent space), whereas all our algorithm steps are directly on the manifold, which is more natural.
3. To connect our iterations with intrinsic parameters of the manifold, we use the exponential map instead of the retraction map used in Criscitiello & Boumal (2019).

There are recent works analyzing other algorithms for escaping from saddle points on manifolds, such as cubic regularization (Agarwal et al., 2018), stochastic gradient descent (Durmus et al., 2020), stochastic variance reduced cubic regularization (Zhang & Tajbakhsh, 2020), etc.

Contributions. We provide convergence guarantees for perturbed first order Riemannian optimization methods to second-order stationary points (local minimum). We prove that as long as the function is appropriately smooth and the manifold has bounded sectional curvature, a perturbed Riemannian gradient descent algorithm escapes (an approximate) saddle points with a rate of $1/\epsilon^2$, a polylog dependence on the dimension of the manifold (hence almost dimension-free), and a polynomial dependence on the smoothness and curvature parameters. This is the first result showing such a rate for Riemannian optimization, and the first to relate the rate to geometric parameters of the manifold.

Despite analogies with the unconstrained (Euclidean) analysis and with the Riemannian optimization literature, the technical challenge in our proof is more than simply combining two lines of work: we need to analyze the interaction between the first-order method and the second order structure of the manifold to obtain second-order convergence guarantees. Unlike in Euclidean space, the curvature affects the Taylor approximation of gradient steps. On the other hand, unlike in the local rate analysis in first-order Riemannian optimization, our second-order analysis requires more refined properties of the manifold structure (whereas in prior works on first order convergence, the linear approximation of the manifold is enough for a local convergence rate proof, see Lemma 1). The works studying second order algorithms such as (Boumal et al., 2016a) use second order oracles (Hessian evaluation).

2.2 *Notation and background*

We consider a complete[2], smooth, d dimensional Riemannian manifold $(\mathcal{M}, \mathfrak{g})$, equipped with a Riemannian metric $\mathfrak{g}$, and we denote by $\mathcal{T}_x\mathcal{M}$ its tangent space at $x \in \mathcal{M}$ (which is a vector space of dimension d). We also denote by $\mathbb{B}_x(r) = \{v \in \mathcal{T}_x\mathcal{M}, \|v\| \leq r\}$ the ball of radius r in $\mathcal{T}_x\mathcal{M}$ centered at 0. At any point $x \in \mathcal{M}$, the metric $\mathfrak{g}$ induces a natural inner product on the tangent space denoted by $\langle \cdot, \cdot \rangle : \mathcal{T}_x\mathcal{M} \times \mathcal{T}_x\mathcal{M} \to \mathbb{R}$. We denote the Levi-Civita connection as ∇ (Absil et al., 2009a, Theorem 5.3.1). The Riemannian curvature tensor is denoted by $R(x)[u, v]$ where $x \in \mathcal{M}$, $u, v \in \mathcal{T}_x\mathcal{M}$ and is defined in terms of the connection ∇ (Absil et al., 2009a, Theorem 5.3.1). The sectional curvature $K(x)[u, v]$ for $x \in \mathcal{M}$ and $u, v \in \mathcal{T}_x\mathcal{M}$ is then defined in Lee (1997, Prop. 8.8).

$$K(x)[u, v] = \frac{\langle R(x)[u, v]u, v\rangle}{\langle u, u\rangle\langle v, v\rangle - \langle u, v\rangle^2}, \ x \in \mathcal{M}, \ u, v \in \mathcal{T}_x\mathcal{M}.$$

Denote the distance (induced by the Riemannian metric) between two points in $\mathcal{M}$ by $d(x, y)$. A geodesic $\gamma : \mathbb{R} \to \mathcal{M}$ is a constant speed curve whose length is equal to $d(x, y)$. It is the shortest path on manifold linking x and y. $\gamma_{x\to y}$ denotes the geodesic from x to y (thus $\gamma_{x\to y}(0) = x$ and $\gamma_{x\to y}(1) = y$).

The exponential map $\mathrm{Exp}_x(v)$ maps $v \in \mathcal{T}_x\mathcal{M}$ to $y \in \mathcal{M}$ such that there exists a geodesic γ with $\gamma(0) = x$, $\gamma(1) = y$ and $\frac{d}{dt}\gamma(0) = v$. The injectivity radius at point $x \in \mathcal{M}$ is the maximal radius r for which the exponential map is a diffeomorphism on $\mathbb{B}_x(r) \subset \mathcal{T}_x\mathcal{M}$. We denote the injectivity radius of the manifold by $\mathfrak{I}$. Since the manifold is complete, we have $\mathfrak{I} > 0$. When $x, y \in \mathcal{M}$ satisfies $d(x, y) \leq \mathfrak{I}$, the exponential map admits an inverse

[2] Since our results are local, completeness is not necessary and our results can be easily generalized, with extra assumptions on the injectivity radius.

$\mathrm{Exp}_x^{-1}(y)$, which satisfies $d(x, y) = \|\mathrm{Exp}_x^{-1}(y)\|$. Parallel translation Γ_x^y denotes a the map which transports $v \in \mathcal{T}_x\mathcal{M}$ to $\Gamma_x^y v \in \mathcal{T}_y\mathcal{M}$ along $\gamma_{x\to y}$ such that the vector stays constant by satisfying a zero-acceleration condition (Lee, 1997, Eq(4.13)).

For a smooth function $f : \mathcal{M} \to \mathbb{R}$, $\mathrm{grad} f(x) \in \mathcal{T}_x\mathcal{M}$ denotes the Riemannian gradient of f at $x \in \mathcal{M}$, which satisfies $\frac{d}{dt} f(\gamma(t)) = \langle \gamma'(t), \mathrm{grad} f(x)\rangle$ (see Absil et al., 2009a, Sec 3.5.1 and (3.31)). The Hessian of f is defined jointly with the Riemannian structure of the manifold. The (directional) Hessian at x in direction ξ_x is denoted by $H(x)[\xi_x] := \nabla_{\xi_x} \mathrm{grad} f$, and we denote $H(x)[u, v] := \langle u, H(x)[v]\rangle$. We call $x \in \mathcal{M}$ an $(\epsilon, -\sqrt{\rho\epsilon})$ saddle point when $\|\nabla f(x)\| \le \epsilon$ and $\lambda_{\min}(H(x)) \le -\sqrt{\rho\epsilon}$. Do Carmo (2016) and Lee (1997) provide a thorough review on these important concepts of Riemannian geometry covering the above definitions.

2.3 Perturbed Riemannian gradient algorithm

Our main Algorithm 1 runs as follows:

1. Check the norm of the gradient: If it is large, do one step of Riemannian gradient descent, and the function value decreases.
2. If the norm of gradient is small, it's either an approximate saddle point or a local minimum. Perturb the variable by adding an appropriate level of noise in its tangent space, map it back to the manifold and run a few iterations.
   (a) If the function value decreases, the iterates are escaping from the approximate saddle point (and the algorithm continues)
   (b) If the function value does not decrease, then it is an approximate local minimum (the algorithm terminates).

Algorithm 1 relies on the manifold's exponential map, and is useful for cases where this map is easy to compute (true for many common manifolds). We refer readers to Lee (1997, pp. 81-86) for the exponential map of sphere and hyperbolic manifolds, and Absil et al. (2009a, Example 5.4.2, 5.4.3) for the Stiefel and Grassmann manifolds. If the exponential map is not

Algorithm 1 Perturbed Riemannian gradient algorithm

```
Require: Initial point $x_0 \in \mathcal{M}$, parameters $\beta, \rho, K, \mathfrak{I}$, accuracy $\epsilon$, probability of success $\delta$
  (parameters defined in Assumptions 1, 2, 3 and assumption of Theorem 1).
  Set constants: $\hat{c} \geq 4$, $C := C(K, \beta, \rho)$ (defined in Lemma 2 and proof of Lemma 8)
      and $\sqrt{c_{\max}} \leq \frac{1}{56\hat{c}^2}$, $r = \frac{\sqrt{c_{\max}}}{\chi^2}\epsilon$, $\chi = 3\max\{\log(\frac{d\beta(f(x_0)-f^*)}{\hat{c}\epsilon^2\delta}), 4\}$.
  Set threshold values: $f_{\text{thres}} = \frac{c_{\max}}{\chi^3}\sqrt{\frac{\epsilon^3}{\rho}}$, $g_{\text{thres}} = \frac{\sqrt{c_{\max}}}{\chi^2}\epsilon$, $t_{\text{thres}} = \frac{\chi}{c_{\max}}\frac{\beta}{\sqrt{\rho\epsilon}}$, $t_{\text{noise}} = -t_{\text{thres}} - 1$.
  Set stepsize: $\eta = \frac{c_{\max}}{\beta}$.
  while 1 do
      if $\|\text{grad}f(x_t)\| \leq g_{\text{thres}}$ and $t - t_{\text{noise}} > t_{\text{thres}}$ then
          $t_{\text{noise}} \leftarrow t$, $\tilde{x}_t \leftarrow x_t$, $x_t \leftarrow \text{Exp}_{x_t}(\xi_t)$, $\xi_t$ uniformly sampled from $\mathbb{B}_{x_t}(r) \subset \mathcal{T}_x\mathcal{M}$.
      if $t - t_{\text{noise}} = t_{\text{thres}}$ and $f(x_t) - f(\tilde{x}_{t_{\text{noise}}}) > -f_{\text{thres}}$ then
          Return $\tilde{x}_{t_{\text{noise}}}$
      $x_{t+1}+ \leftarrow \text{Exp}_{x_t}(-\min\{\eta, \frac{\mathfrak{I}}{\|\text{grad}f(x_t)\|}\}\text{grad}f(x_t))$.
      $t \leftarrow t + 1$.
```

computable, the algorithm can use a retraction[3] instead, however our current analysis only covers the case of the exponential map. In Figure 2.1[4], we illustrate a function with saddle point on sphere, and plot the trajectory of Algorithm 1 when it is initialized at a saddle point.

2.4 Main theorem: escape rate for perturbed Riemannian gradient descent

We now turn to our main results, beginning with our assumptions and a statement of our main theorem. We then develop a brief proof sketch.

Our main result involves two conditions on function f and one on the curvature of the manifold $\mathcal{M}$.

Assumption 1 (Lipschitz gradient)**.** *There is a finite constant β such that*

$$\|\text{grad}f(y) - \Gamma_x^y \text{grad}f(x)\| \leq \beta d(x, y) \quad \textit{for all } x, y \in \mathcal{M}.$$

[3] A retraction is a first-order approximation of the exponential map which is often easier to compute.

[4] Codes for generating figures are available at `https://sunyue93.github.io/code.zip`.

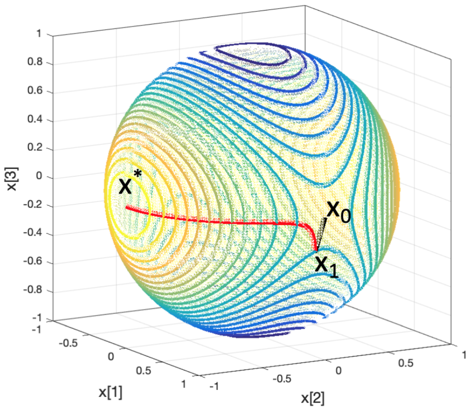


Figure 2.1: Function f with saddle point on a sphere. $f(x) = x_1^2 - x_2^2 + 4x_3^2$. We plot the contour of this function on unit sphere. Algorithm 1 initializes at $x_0 = [1, 0, 0]$ (a saddle point), perturbs it towards x_1 and runs Riemannian gradient descent, and terminates at $x^* = [0, -1, 0]$ (a local minimum). We amplify the first iteration to make saddle perturbation visible.

Assumption 2 (Lipschitz Hessian)**.** *There is a finite constant ρ such that*

$$\|H(y) - \Gamma_x^y H(x) \Gamma_y^x\|_2 \leq \rho d(x, y) \quad \textit{for all } x, y \in \mathcal{M}.$$

Assumption 3 (Bounded sectional curvature)**.** *There is a finite constant K such that*

$$|K(x)[u, v]| \leq K \quad \textit{for all } x \in \mathcal{M} \textit{ and } u, v \in \mathcal{T}_x\mathcal{M}$$

K is an intrinsic parameter of the manifold capturing the curvature. We list a few examples here: (i) A sphere of radius R has a constant sectional curvature $K = 1/R^2$ (Lee, 1997, Theorem 1.9). If the radius is bigger, K is smaller which means the sphere is less curved; (ii) A hyper-bolic space H_R^n of radius R has $K = -1/R^2$ (Lee, 1997, Theorem 1.9); (iii) For sectional curvature of the Stiefel and the Grasmann manifolds, we refer readers to Rapcsák (2008, Section 5) and Wong (1968), respectively.

Note that the constant K is not directly related to the RLICQ parameter R defined by

Ge et al. (2015) which first requires describing the manifold by equality constraints. Different representations of the same manifold could lead to different curvature bounds, while sectional curvature is an intrinsic property of manifold. If the manifold is a sphere $\sum_{i=1}^{d+1} x_i^2 = R^2$, then $K = 1/R^2$, but generally there is no simple connection. The smoothness parameters are natural compared to some quantity from complicated compositions (Lee et al., 2017, Section 5.5) or pullback (Zhang & Zhang, 2018; Criscitiello & Boumal, 2019). With these assumptions, the main result of this work is the following:

Theorem 1. *Under Assumptions 1,2,3, let $C(K, \beta, \rho)$ be a function defined in Lemma 2, $\hat{\rho} = \max\{\rho, C(K, \beta, \rho)\}$, if ϵ satisfies that*

$$\epsilon \leq \min\left\{\frac{\hat{\rho}}{56\max\{c_2(K), c_3(K)\}\eta\beta}\log\left(\frac{d\beta}{\sqrt{\hat{\rho}\epsilon}\delta}\right), \left(\frac{\mathfrak{I}\hat{\rho}}{12\hat{c}\sqrt{\eta\beta}}\log\left(\frac{d\beta}{\sqrt{\hat{\rho}\epsilon}\delta}\right)\right)^2\right\} \tag{2.2}$$

where $c_2(K)$, $c_3(K)$ are defined in Lemma 4, then with probability $1-\delta$, perturbed Riemannian gradient descent with step size $c_{\max}/\beta$ converges to a $(\epsilon, -\sqrt{\hat{\rho}\epsilon})$-stationary point of f in

$$O\left(\frac{\beta(f(x_0) - f(x^*))}{\epsilon^2}\log^4\left(\frac{\beta d(f(x_0) - f(x^*))}{\epsilon^2\delta}\right)\right)$$

iterations.

Proof roadmap. For a function satisfying smoothness condition (Assumption 1 and 2), we use a local upper bound of the objective based on the third-order Taylor expansion

$$f(u) \leq f(x) + \langle \mathrm{grad} f(x), \mathrm{Exp}_x^{-1}(u)\rangle + \frac{1}{2}H(x)[\mathrm{Exp}_x^{-1}(u), \mathrm{Exp}_x^{-1}(u)] + \frac{\rho}{6}\|\mathrm{Exp}_x^{-1}(u)\|^3.$$

When the norm of the gradient is large (not near a saddle), the following lemma guarantees the decrease of the objective function in one iteration.

Lemma 1. *(Boumal et al., 2018) Under Assumption 1, by choosing $\bar{\eta} = \min\{\eta, \frac{\mathfrak{I}}{\|\mathrm{grad} f(u)\|}\} = O(1/\beta)$, the Riemannian gradient descent algorithm is monotonically descending, and $f(u^+) \leq$*

$f(u) - \frac{1}{2}\bar{\eta}\|\mathrm{grad} f(u)\|^2$.

Thus our main challenge in proving the main theorem is the Riemannian gradient behaviour at an approximate saddle point:

1. Similar to the Euclidean case studied by Jin et al. (2017a), we need to bound the probability where the perturbation fails, and we do it by bounding the "thickness" of the "stuck region" . We use a pair of hypothetical auxiliary sequences and study the "coupling" sequences. When two perturbations couple in the thinnest direction of the stuck region, their distance grows and one of them escapes from saddle point.

2. Our iterates are evolving on a manifold rather than a Euclidean space, so our strategy is to map the iterates back to an appropriate fixed tangent space where we can use the Euclidean analysis. This is done using the inverse of the exponential map and parallel transports.

3. Several key challenges arise in doing this. Unlike Jin et al. (2017a), the structure of the manifold interacts with the local approximation of the objective function in a complicated way. On the other hand, unlike recent work on Riemannian optimization by Boumal et al. (2016a), we do not have access to a second order oracle and we need to understand how the sectional curvature and the injectivity radius (which both capture intrinsic manifold properties) affect the behavior of the first order iterates.

4. Our main contribution is to carefully investigate how the various approximation errors arising from (a) the linearization of the iteration couplings and (b) their mappings to a common tangent space can be handled on manifolds with bounded sectional curvature. We address these challenges in a sequence of lemmas (Lemmas 3 through 6) we combine to linearize the coupling iterations in a common tangent space and precisely control the approximation error. This result is formally stated in the following lemma.

Lemma 2. *Define $\gamma = \sqrt{\hat{\rho}\epsilon}$, $\kappa = \frac{\beta}{\gamma}$, and $\mathscr{S} = \sqrt{\eta\beta}\frac{\gamma}{\hat{\rho}}\log^{-1}(\frac{d\kappa}{\delta})$. Let us consider x be a $(\epsilon, -\sqrt{\hat{\rho}\epsilon})$ saddle point, and define $u^+ = \mathrm{Exp}_u(-\eta \mathrm{grad} f(u))$ and $w^+ = \mathrm{Exp}_w(-\eta \mathrm{grad} f(w))$. Under Assumptions 1, 2, 3, if all pairwise distances between u, w, u^+, w^+, x are less than $12\mathscr{S}$,* *then for some explicit constant $C(K,\rho,\beta)$ depending only on K,ρ,β,*

$$\begin{aligned}&\|\mathrm{Exp}_x^{-1}(w^+)-\mathrm{Exp}_x^{-1}(u^+)-(I-\eta H(x))(\mathrm{Exp}_x^{-1}(w)-\mathrm{Exp}_x^{-1}(u))\| \\ &\leq C(K,\rho,\beta)d(u,w)\left(d(u,w)+d(u,x)+d(w,x)\right).\end{aligned} \tag{2.3}$$

The proof of this lemma includes novel contributions by strengthen known result (Lemmas 3) and also combining known inequalities in novel ways (Lemmas 4 to 6) that allow us to control all the approximation errors and arrive at the tight rate of escape for the algorithm.

2.5 Proof of Lemma 2

Lemma 2 controls the error of the linear approximation of the iterates when mapped in $T_x\mathcal{M}$. In this section, we assume that all points are within a region of diameter $R := 12\mathscr{S} \leq \mathfrak{I}$ (inequality follows from (2.2)), i.e., the distance of any two points in the following lemmas are less than R.

The proof of Lemma 2 is based on the sequence of following lemmas.

Lemma 3. *Let $x \in \mathcal{M}$ and $y, a \in T_x\mathcal{M}$. Let us denote by $z = \mathrm{Exp}_x(a)$ then under Assumption 3*

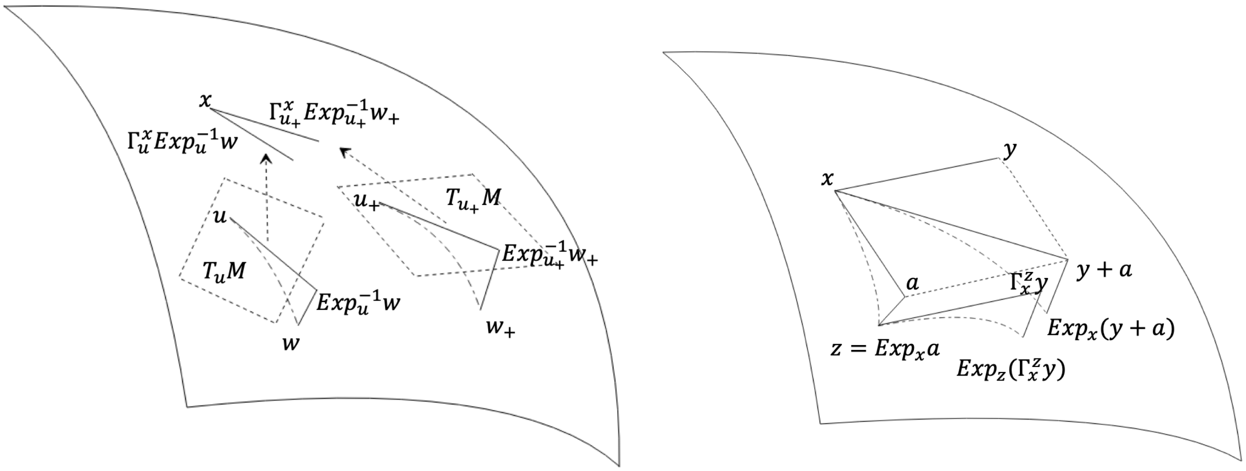
$$d(\mathrm{Exp}_x(y+a), \mathrm{Exp}_z(\Gamma_x^z y)) \leq c_1(K)\min\{\|a\|, \|y\|\}(\|a\|+\|y\|)^2. \tag{2.4}$$

This lemma tightens the result of Karcher (1977, C2.3), which only shows an upper-bound $O(\|a\|(\|a\|+\|y\|)^2)$. We prove the upper-bound $O(\|y\|(\|a\|+\|y\|)^2)$ in the Appendix A.3.

We also need the following lemma showing that both the exponential map and its inverse are Lipschitz.

Lemma 4. *Let $x, y, z \in M$, and the distance of each two points is no bigger than R. Then under Assumption 3,*

$$(1+c_2(K)R^2)^{-1}d(y,z) \leq \|\mathrm{Exp}_x^{-1}(y)-\mathrm{Exp}_x^{-1}(z)\| \leq (1+c_3(K)R^2)d(y,z).$$



(a)

(b)

Figure 2.2: (a) Eq(2.5), first map w and w_+ to $\mathcal{T}_u\mathcal{M}$ and $\mathcal{T}_{u_+}\mathcal{M}$, and transport the two vectors to $\mathcal{T}_x\mathcal{M}$, and get their relation. (b) Lemma 3 bounds the difference of two steps starting from x: (1) take $y + a$ step in $\mathcal{T}_x\mathcal{M}$ and map it to manifold, and (2) take a step in $\mathcal{T}_x\mathcal{M}$, map to manifold, call it z, and take $\Gamma_x^z y$ step in $\mathcal{T}_x\mathcal{M}$, and map to manifold. $\mathrm{Exp}_z(\Gamma_x^z y)$ is close to $\mathrm{Exp}_x(y + a)$.

Intuitively this lemma relates the norm of the difference of two vectors of $\mathcal{T}_x\mathcal{M}$ to the distance between the corresponding points on the manifold $\mathcal{M}$ and follows from bounds on the Hessian of the square-distance function (Sakai, 1996, Ex. 4 p. 154). The upper-bound is directly proven by Karcher (1977, Proof of Cor. 1.6), and we prove the lower-bound via Lemma 3.

The following contraction result is fairly classical and is proven using the Rauch comparison theorem from differential geometry (Cheeger & Ebin, 2008).

Lemma 5. *(Mangoubi et al., 2018, Lemma 1) Under Assumption 3, for $x, y \in \mathcal{M}$ and $w \in T_x\mathcal{M}$,*

$$d(\mathrm{Exp}_x(w), \mathrm{Exp}_y(\Gamma_x^y w)) \leq c_4(K) d(x, y).$$

Finally we need the following corollary of the Ambrose-Singer theorem (Ambrose & Singer, 1953).

Lemma 6. *(Karcher, 1977, Section 6) Under Assumption 3, for $x, y, z \in \mathcal{M}$ and $w \in T_x\mathcal{M}$,*

$$\|\Gamma_y^z \Gamma_x^y w - \Gamma_x^z w\| \leq c_5(K) d(x,y) d(y,z) \|w\|.$$

Lemma 3 through 6 are mainly proven in the literature, and we make up the missing part in Appendix A.3. Then we prove Lemma 2 in Appendix A.3.

The spirit of the proof is to linearize the manifold using the exponential map and its inverse, and to carefully bounds the various error terms caused by the approximation. Let us denote by $\theta = d(u,w) + d(u,x) + d(w,x)$.

1. We first show using twice Lemma 3 and Lemma 5 that

$$d(\mathrm{Exp}_u(\mathrm{Exp}_u^{-1}(w) - \eta \Gamma_w^u \mathrm{grad} f(w)), \mathrm{Exp}_u(-\eta \mathrm{grad} f(u) + \Gamma_{u_+}^u \mathrm{Exp}_{u_+}^{-1}(w_+))) = O(\theta d(u,w)).$$

2. We use Lemma 4 to linearize this iteration in $\mathcal{T}_u\mathcal{M}$ as

$$\|\Gamma_{u_+}^u \mathrm{Exp}_{u_+}^{-1}(w_+) - \mathrm{Exp}_u^{-1}(w) + \eta[\mathrm{grad} f(u) - \Gamma_w^u \mathrm{grad} f(w)]\| = O(\theta d(u,w)).$$

3. We use the Hessian Lipschitzness

$$\|\Gamma_{u_+}^u \mathrm{Exp}_{u_+}^{-1}(w_+)) - \mathrm{Exp}_u^{-1}(w) + \eta H(u) \mathrm{Exp}_u^{-1}(w)\| = O(\theta d(u,w)).$$

3. We use Lemma 6 to map to $T_x\mathcal{M}$ and the Hessian Lipschitzness to compare $H(u)$ to $H(x)$. This is an important intermediate result.

$$\|\Gamma_{u_+}^x \mathrm{Exp}_{u_+}^{-1}(w_+) - \Gamma_u^x \mathrm{Exp}_u^{-1}(w) + \eta H(x) \Gamma_u^x \mathrm{Exp}_u^{-1}(w)\| = O(\theta d(u,w)). \tag{2.5}$$

4. We use Lemma 3 and 4 to approximate two iteration updates in $\mathcal{T}_x\mathcal{M}$.

$$\|\mathrm{Exp}_x^{-1}(w) - (\mathrm{Exp}_x^{-1}(u) + \Gamma_u^x \mathrm{Exp}_u^{-1}(w))\| \leq O(\theta d(u,w)). \tag{2.6}$$

And same for the u_+, w_+ pair replacing u, w.

5. Combining (2.5) and (2.6) together, we obtain

$$\|\text{Exp}_x^{-1}(w^+) - \text{Exp}_x^{-1}(u^+) - (I - \eta H(x))(\text{Exp}_x^{-1}(w) - \text{Exp}_x^{-1}(u))\| \leq O(\theta d(u, w)).$$

Now note that, the iterations u, u_+, w, w_+ of the algorithm are both on the manifold. We use $\text{Exp}_x^{-1}(\cdot)$ to map them to the same tangent space at x.

Therefore we have linearized the two coupled trajectories $\text{Exp}_x^{-1}(u_t)$ and $\text{Exp}_x^{-1}(w_t)$ in a common tangent space, and we can modify the Euclidean escaping saddle analysis thanks to the error bound we proved in Lemma 2.

2.6 Proof of main theorem

In this section we suppose all assumptions in Section 2.4 hold. The proof strategy is to show with high probability that the function value decreases of $\mathscr{F}$ in $\mathscr{T}$ iterations at an approximate saddle point. Lemma 7 suggests that, if after a random perturbation and $\mathscr{T}$ steps, the iterate is $\Omega(\mathscr{S})$ far from the approximate saddle point, then the function value decreases. If the iterates do not move far, the perturbation falls in a stuck region. Lemma 8 uses a coupling strategy, and suggests that the width of the stuck region is small in the negative eigenvector direction of the Riemannian Hessian.

Define

$$\mathscr{F} = \eta\beta\frac{\gamma^3}{\hat{\rho}^2}\log^{-3}(\frac{d\kappa}{\delta}),\ \mathscr{G} = \sqrt{\eta\beta}\frac{\gamma^2}{\hat{\rho}}\log^{-2}(\frac{d\kappa}{\delta}),\ \mathscr{T} = \frac{\log(\frac{d\kappa}{\delta})}{\eta\gamma}.$$

At an approximate saddle point $\tilde{x}$, let y be in the neighborhood of $\tilde{x}$ where $d(y, \tilde{x}) \leq \mathfrak{I}$, denote

$$\tilde{f}_y(x) := f(y) + \langle \text{grad} f(y), \text{Exp}_y^{-1}(\tilde{x})\rangle + \frac{1}{2}H(\tilde{x})[\text{Exp}_y^{-1}(\tilde{x}), \text{Exp}_y^{-1}(\tilde{x})].$$

Let $\|\text{grad} f(\tilde{x})\| \leq \mathscr{G}$ and $\lambda_{\min}(H(\tilde{x})) \leq -\gamma$. We consider two iterate sequences, $u_0, u_1, ...$ and $w_0, w_1, ...$ where u_0, w_0 are two perturbations at $\tilde{x}$.

Lemma 7. *Assume Assumptions 1, 2, 3 and* (2.2) *hold. There exists a constant* $c_{\max}$*,* $\forall \hat{c} > 3, \delta \in (0, \frac{d\kappa}{e}]$*, for any* u_0 *with* $d(\tilde{x}, u_0) \leq 2\mathscr{S}/(\kappa \log(\frac{d\kappa}{\delta}))$*,* $\kappa = \beta/\gamma$*.*

$$T = \min \left\{ \inf_t \left\{ t | \tilde{f}_{u_0}(u_t) - f(u_0) \leq -3\mathscr{F} \right\}, \hat{c}\mathscr{T} \right\},$$

then $\forall \eta \leq c_{\max}/\beta$*, we have* $\forall 0 < t < T$*,* $d(u_0, u_t) \leq 3(\hat{c}\mathscr{S})$*.*

Lemma 8. *Assume Assumptions 1, 2, 3 and* (2.2) *hold. Take two points* u_0 *and* w_0 *which are perturbed from an approximate saddle point, where* $d(\tilde{x}, u_0) \leq 2\mathscr{S}/(\kappa \log(\frac{d\kappa}{\delta}))$*,* $\mathrm{Exp}_{\tilde{x}}^{-1}(w_0) - \mathrm{Exp}_{\tilde{x}}^{-1}(u_0) = \mu r e_1$*,* e_1 *is the smallest eigenvector*[5] *of* $H(\tilde{x})$*,* $\mu \in [\delta/(2\sqrt{d}), 1]$*, and the algorithm runs two sequences* $\{u_t\}$ *and* $\{w_t\}$ *starting from* u_0 *and* w_0*. Denote*

$$T = \min \left\{ \inf_t \left\{ t | \tilde{f}_{w_0}(w_t) - f(w_0) \leq -3\mathscr{F} \right\}, \hat{c}\mathscr{T} \right\},$$

then $\forall \eta \leq c_{\max}/l$*, if* $\forall 0 < t < T$*,* $d(\tilde{x}, u_t) \leq 3(\hat{c}\mathscr{S})$*, we have* $T < \hat{c}\mathscr{T}$*.*

We prove Lemma 7 and 8 in Appendix A.4. We also prove, in the same section, the main theorem using the coupling strategy of Jin et al. (2017a). but with the additional difficulty of taking into consideration the effect of the Riemannian geometry (Lemma 2) and the injectivity radius.

2.7 Numerical examples

kPCA. We consider the kPCA problem, where we want to find the $k \leq n$ principal eigenvectors of a symmetric matrix $H \in \mathbb{R}^{n \times n}$, as an example (Tripuraneni et al., 2018). This corresponds to

$$\min_{X \in \mathbb{R}^{n \times k}} -\frac{1}{2}\mathrm{tr}(X^T H X) \quad \text{subject to } X^T X = I,$$

[5] "smallest eigenvector" means the eigenvector corresponding to the smallest eigenvalue.

which is an optimization problem on the Grassmann manifold defined by the constraint $X^TX = I$. If the eigenvalues of H are distinct, we denote by v_1,...,v_n the eigenvectors of H, corresponding to eigenvalues with decreasing order. Let $V^* = [v_1, ..., v_k]$ be the matrix with columns composed of the top k eigenvectors of H, then the local minimizers of the objective function are V^*G for all unitary matrices $G \in \mathbb{R}^{k\times k}$. Denote also by $V = [v_{i_1}, ..., v_{i_k}]$ the matrix with columns composed of k distinct eigenvectors, then the first order stationary points of the objective function (with Riemannian gradient being 0) are VG for all unitary matrices $G \in \mathbb{R}^{k\times k}$. In our numerical experiment, we choose H to be a diagonal matrix $H = \text{diag}(0, 1, 2, 3, 4)$ and let $k = 3$. The Euclidean basis (e_i) are an eigenbasis of H and the first order stationary points of the objective function are $[e_{i_1}, e_{i_2}, e_{i_3}]G$ with distinct basis and G being unitary. The local minimizers are $[e_3, e_4, e_5]G$. We start the iteration at $X_0 = [e_2, e_3, e_4]$ and see in Fig. 2.3 the algorithm converges to a local minimum.

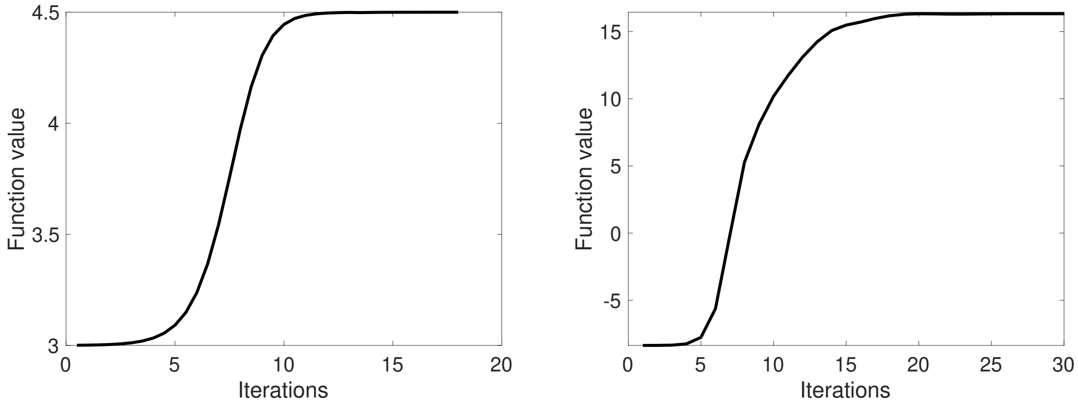
Burer-Monteiro approach for certain low rank problems. Following Boumal et al. (2016b), we consider, for $A \in \mathbb{S}^{d\times d}$ and $r(r + 1)/2 \leq d$, the problem

$$\min_{X\in\mathbb{S}^{d\times d}} \mathbf{tr}(AX), \ \ s.t. \ \text{diag}(X) = 1, X \succeq 0, \text{rank}(X) \leq r.$$

We factorize X by YY^T with an overparametrized $Y \in \mathbb{R}^{d\times p}$ and $p(p + 1)/2 \geq d$. Then any local minimum of

$$\min_{Y\in\mathbb{R}^{d\times p}} \mathbf{tr}(AYY^T), \ \ s.t. \ \text{diag}(YY^T) = 1,$$

is a global minimum where $YY^T = X^*$ (Boumal et al., 2016b). Let $f(Y) = \frac{1}{2}\mathbf{tr}(AYY^T)$. In the experiment, we take $A \in \mathbb{R}^{100\times 20}$ being a sparse matrix that only the upper left 5×5 block is random and other entries are 0. Let the initial point $Y_0 \in \mathbb{R}^{100\times 20}$, such that $(Y_0)_{i,j} = 1$ for $5j - 4 \leq i \leq 5j$ and $(Y_0)_{i,j} = 0$ otherwise. Then Y_0 is a saddle point. We see in Fig. 2.3 the algorithm converges to the global optimum.



(a)

(b)

Figure 2.3: (a) kPCA problem with $H = \mathrm{diag}(0, 1, 2, 3, 4)$, $X \in \mathbb{R}^{5\times 3}$, $\eta = 0.1$, $X_0 = [e_2, e_3, e_4]$. plot $f(X) = \frac{1}{2}\mathrm{trace}(X^T H X)$ versus iterations. We start from an approximate saddle point, and it converges to a local minimum (which is also global minimum). (b) Burer-Monteiro approach with $A \in \mathbb{R}^{100\times 100}$ such that the first 5×5 block is random and other entries are 0, $Y \in \mathbb{R}^{100\times 20}$, $\eta = 0.1$, $(Y_0)_{i,j} = 1$ if $5j - 4 \leq i \leq 5j$. Plot $f(Y) = \frac{1}{2}\mathrm{trace}(AYY^T)$ versus iterations. We start from the saddle point, and it converges to a local minimum (which is also global minimum).

2.8 Conclusion and future directions

Previous works have shown that in Euclidean space, although the gradient descent can converge to an approximate second order minimum in exponential time, by simply adding a random perturbation at the stationary points, the gradient descent iteration escapes from saddle points and converges to an approximate second order minimum with provable polynomial rate. However, they require the problem being unconstrained, which does not allow a smooth manifold constraint, or the optimization problem set up in Riemannian manifolds. No result was given about the second order convergence of perturbed first order optimization methods on Riemannian manifolds, and it is unknown how the curvature constant of the manifold contributes to the rate of escaping from saddle points. We have shown that for the constrained optimization problem of minimizing $f(x)$ subject to a manifold constraint, if the function and the manifold are appropriately smooth, a perturbed Riemannian gradient descent algorithm will escape saddle points with a rate of order $1/\epsilon^2$ in the accuracy ϵ, polylog in manifold dimension d, and depends polynomially on the curvature and smoothness parameters.

A natural extension of our result is to consider other variants of gradient descent, such as the heavy ball method, Nesterov's acceleration, and the stochastic setting. The question is whether these algorithms with appropriate modification (with manifold constraints) would have a fast convergence to second-order stationary point (not just first-order stationary as studied in recent literature), and whether it is possible to show the relationship between convergence rate and smoothness of manifold.

Chapter 3

ANALYSIS OF POLICY OPTIMIZATION FOR CONTROL: GLOBAL OPTIMALITY VIA CONVEX PARAMETERIZATION

This chapter proposes a framework that builds the mapping between a few control problems with their associated convex parameterized form. With the mapping, we show that all stationary points of the cost functions, as functions of the policy, are global minima despite their nonconvexity. The fact allows first order optimization methods (i.e., policy gradient method) to converge the globally optimal controller. We give a comprehensive theory covering many control problems, including continuous/discrete time/Markov jump LQR, distributed optimal control, minimizing the $\mathcal{L}_2$ gain that unifies the conclusion of each specific work.

This work is published as Sun & Fazel (2021).

3.1 Introduction

During the recent boom of reinforcement learning (RL), many optimal control problem are revisited as RL problems. However, we see a sharp difference between the training techniques in RL and in control theory. In RL, policy optimization is widely used, where one formulates a cost function as a function of the controller/policy, and runs zeroth or first order update in the policy space. This method is straightforward and empirically finds good policy, but in control they usually end up with a non-convex objective, where it is unknown whether gradient based algorithm converges to the global minimum. In control theory, one reparameterizes the cost function, and ends up with a convex objective in the reparameterized space instead of the policy space and solves the convex optimization problem. We are interested in explaining the success of the first order nonconvex policy optimization in RL, especially its convergence to global optimum, via our understanding of the convex parameterization technique in control theory.

We start by reviewing linear quadratic regulator (LQR), which is one of the most well studied optimal control problems (Kalman et al., 1960). Consider the continuous time linear time-invariant dynamical system,

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \tag{3.1}$$

where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^p$ is the input, and A, B are constant matrices describing the dynamics. The goal of optimal control is to determine the input series $u(t)$ that minimizes some cost function (that typically depends on the state and input). In the infinite horizon LQR problem, we define constant matrices $Q \in \mathbf{S}^n_{++}$, $R \in \mathbf{S}^p_{++}$, and minimize the cost as a function of input

$$\mathrm{cost}(u(t)) := \mathbf{E}_{x_0} \int_0^\infty (x(t)^\top Q x(t) + u(t)^\top R u(t)) \mathrm{d}t. \tag{3.2}$$

The optimal controller is linear in the state, called static state feedback controller, and can be described as $u(t) = Kx(t)$ for a constant $K \in \mathbb{R}^{p \times n}$ (Kalman et al., 1960).

We can define this cost function with variable K,

$$\begin{aligned}\mathcal{L}(K) &:= \mathbf{E}_{x_0} \int_0^\infty (x(t)^\top Q x(t) + u(t)^\top R u(t)) \mathrm{d}t, \text{ s.t. } u(t) = Kx(t) \\ &= \mathbf{E}_{x_0} \int_0^\infty x(t)^\top (Q + K^\top R K) x(t) \mathrm{d}t. \end{aligned} \tag{3.3}$$

Policy Optimization. The policy optimization (aka. policy gradient descent method if applying gradient descent or first order optimization method in policy space) is to minimize

$\mathcal{L}(K)$ by running first order method with respect to K. [1]We run[2]

$$K_{t+1} \leftarrow K_t - \eta_t \nabla \mathcal{L}(K_t).$$

It is shown that, the cost functions of control problems are typically nonconvex in K, e.g., continuous/discrete time LQR (Mohammadi et al., 2019b; Fazel et al., 2018). However, gradient descent for nonconvex optimization is widely used in machine learning, or control tasks with the context of reinforcement learning.

Convex parameterization. In classical control theory literature, due to the nonconvex nature, policy optimization is not commonly used. Instead, one can introduce another parameterization of the cost to make it convex, and apply convex optimization method with global convergence guarantee. This approach is in sharp contrast to how one would typically minimize a cost function through gradient descent on K.

Motivation: There are many papers that show convergence of first order policy optimization methods (which we will review below). They investigate different control problems and the proofs are given case by case. However, we observe that all the results are proven by the gradient dominance property, a special case of Lojasiewicz inequality[3], and all of them were solved by convex parameterization methods in classical control literature. Thus, we ask whether there is a proof that unifies the proofs of the gradient dominance property for different control problems, and bridges nonconvex methods with convex methods in classical control literature.

Contributions: We make a connection between nonconvex first order policy optimization

[1] We initialize K_0 as a stabilizing controller, so that $\mathcal{L}(K_0)$ is well defined. For LQR as an example, Perdomo et al. (2021) demonstrates an algorithm to find a stabilizing controller by policy optimization. They begin with an arbitrary controller, and define an alternative cost with a discount factor to make the cost finite, and run gradient based method on that cost and later anneal the discount factor.

[2] Zeroth order method is a specific implementation of stochastic gradient descent method and is used a lot in reinforcement learning. Duchi et al. (2015) proposes the two point estimation method for zeroth order optimization. Malik et al. (2019) is a survey of the zeroth order realization of policy optimization method on discrete time LQR with sample complexity analysis.

[3] $\|\nabla \mathcal{L}(K)\|_F \gtrsim (\mathcal{L}(K) - \mathcal{L}(K^*))^\alpha$ for a positive number α (Lojasiewicz, 1963).

and known convex parameterization methods with a map between the two parameters. This map maintains the Lojasiewicz inequality when mapping from the convex landscape to the nonconvex landscape.

Our result is quite general—we show that continuous-time LQR is a special case that the main theorems apply to, and we generalize the guarantees provided by this method to a range of other control problems. The instances cover LQR for continuous, discrete time system, and Markov jump system, maximizing $\mathcal{L}_2$ gain and system level synthesis. To judge whether a nonconvex landscape can be optimized globally using first order method in policy space, one can directly check if it is covered by the theorems, avoiding a case-by-case analysis. Also, as discussed in Fazel et al. (2018), theoretical guarantees for first-order methods naturally lead to guarantees for the more practical zeroth-order optimization or sampling-based methods, which do not need access to the gradient of the cost with respect to K.

Outline: The rest of this paper is structured as follows. Sec. 3.2 reviews the continuous-time LQR problem. Sec. 3.3 presents our main result on the the nonconvex cost, showing all stationary points are global minima. Sec. 3.4 lists more examples of control problems covered by the main theorem. Although Sec. 3.4 covered many problems, Sec. 3.5 further generalizes our main result using different parameterizations and Sec. 3.6 covers examples under the more generic result. Sec. 3.7 gives a proof sketch with intuitive connections between the nonconvex and convex formulations.

3.1.1 Prior art

LQR with unknown system matrices: model-based and model-free. There are two major types of algorithms when system matrices are not known. The first type is model-based methods, when we first estimate the system matrices and then a controller is constructed based on the identified system.

The second type of method is model free method, when the controller is directly trained by observing the cost function or its gradient, without characterizing the dynamics. Here one does not necessarily estimate the system matrices A, B and runs zeroth order update based on function value estimators that a simulator usually provides without explicitly giving the system matrices Fazel et al. (2018). [4]

Recent works on policy optimization. First order policy optimization method calls for an estimate of the cost and its gradient with respect to controller K. The goal is to show that gradient descent with respect to K converges to the optimal controller (we can call it K^*). The policy gradient descent is recently reviewed by Kakade (2002); Rajeswaran et al. (2017). Fazel et al. (2018) provides a counterexample showing that minimizing the quadratic LQ cost as a function of K is not convex, quasi-convex or star-convex.

There has been recent evidence of the *empirical* success of first order methods in solving nonconvex reinforcement learning problems. (Mårtensson, 2012, Ch. 3) proposes the gradient based method for optimal control and extends to decentralized control. Roberts et al. (2011) studies feedback control with dynamical controllers, and observes that gradient descent with Youla parameterization is robust within the set of stabilizing controllers while other parameterizations are not. On the *theoretical* side, despite the nonconvexity of $\mathcal{L}(K)$, for certain types of control problems, there are works showing the *gradient dominance* property, which enables first order methods to converge to the global optimum. Fazel et al. (2018) gives the first result by proving the coercivity and the gradient dominance properties of $\mathcal{L}(K)$ for the discrete time LQR. Based on this, Fazel et al. (2018) shows the linear convergence of gradient based method. Later Mohammadi et al. (2019b) shows a similar result for the continuous time case, papers Bu et al. (2019a, 2020) give a more detailed analysis for both discrete and continuous time LQR. Bu et al. (2019b); Zhang et al. (2019) show the convergence for two types of zero-sum LQ games. Zhang et al. (2020) studies the convergence of gradient descent on $\mathcal{H}_2$ control with $\mathcal{H}_\infty$ constraint, and shows that gradient descent implicitly makes the controller robust. Furieri et al. (2020) shows the convergence for finite-horizon distributed control under the quadratic invariance assumption.

[4] Lewis & Vrabie (2009) is a review of reinforcement learning area and optimal control, which studies a few fixed point type dynamic programming methods . Q-learning is a typical model free method for reinforcement learning, and it is applied to LQR as in Bradtke et al. (1994); Lee et al. (2012); Lee & Hu (2018).

Negative results. We note that, we cannot cover all optimal control problems with our theory. The feasible domain of structured LQR (where the state feedback controller K is in a low dimensional subspace) (Li et al., 2021; Feng & Lavaei, 2020) and static output feedback LQ problem (Feng & Lavaei, 2020) are not connected thus cannot be globally optimized using first order methods. The LQG problem, although has a convex parameterization, due to the non-smoothness of the parameterization, does not satisfy the setup of the main theorem in this chapter, and its cost is proven to have saddle points in policy space (Tang et al., 2021). We will discuss them in more detail in Sec. 3.8.

3.2 Review of convex parameterization for continuous-time LQR

Convex parameterization (e.g., solving optimal control by linear matrix inequalities (LMI) in Boyd et al. (1994)) is widely used in optimal control problems, and here we discuss its application for continuous time LQR (Mohammadi et al., 2019b). We will introduce new variables, construct an equivalent convex optimization problem with new variables, and the pair of variables are proven to be linked by a bijection. In the next section we use the critical properties of the nonconvex and convex problems as an intuition to generalize to a more general form.

Consider a continuous time linear time invariant system (3.1) where x is the state, u is the input, and x_0 is the initial state, which we assume is randomly picked from a zero-mean distribution with covariance $\Sigma := \mathbf{E}(x_0 x_0^\top) \succ 0$. This is a commonly used setup in e.g., the theoretical study (Bu et al., 2019a, §3.3) and the practical work (Mårtensson, 2012, Ch. 3). With $\Sigma \succ 0$, the optimal controller is not dependent on the initial state; when Σ is low rank, then a controller K that gives finite LQR cost does not stabilize the system for all initial state $x_0 \in \text{null}(\Sigma)$.

One can then consider minimizing the linear quadratic (LQ) cost (3.2) as a function of $u(t)$ where Q, R are positive definite matrices. Kalman et al. (1960) proves that, the input signal that minimizes the cost function $\text{cost}(u)$ is given by a static state feedback controller, denoted by $u(t) = K^* x(t)$. K^* can be obtained by solving linear equations, called Riccati equations.

Once we know that the optimal state feedback controller is static, we can write cost as $\mathcal{L}(K)$ as (3.3). It is a function of K, and we search only static state feedback controllers.

Solving the LQR problem, the classical way. The LQR problem is often solved using the algebraic Riccati equation (ARE) (Stengel, 1994; Dullerud & Paganini, 2013). The ARE has been widely studied in the literature, with solution methods including iterative algorithms (Hewer, 1971), algebraic solution methods (Lancaster & Rodman, 1995), and semidefinite programming (Balakrishnan & Vandenberghe, 2003).

An alternative approach is reparameterization, to obtain a convex optimization problem, as used in Mohammadi et al. (2019b). We will review it here, starting from the Lyapunov equation. Suppose the initial state satisfies $\mathbf{E}(x_0x_0^\top) = \Sigma \succ 0$, and $\dot{x}(t) = Ax(t)$. Then with a matrix $P \in \mathbf{S}_{++}^{n\times n}$ (P is a positive definite matrix) as the variable, the Lyapunov equation is written as

$$AP + PA^\top + \Sigma = 0.$$

In our setup (3.1), we use a state feedback controller $u = Kx$, thus we have $\dot{x} = (A + BK)x$. We denote the set of stabilizing controllers as $\mathcal{S}_{K,\text{sta}}$, which is defined as

$$\mathcal{S}_{K,\text{sta}} = \{K :\ \text{Re}(\lambda_i(A + BK)) < 0,\ i = 1, ..., n\}.$$

If a state feedback controller is applied, the cost is only bounded when $K \in \mathcal{S}_{K,\text{sta}}$ and is coercive in $\mathcal{S}_{K,\text{sta}}$ (Bu et al., 2020). Replace A by the closed loop system matrix $A + BK$ in the Lyapunov equation, and let $L = KP \in \mathbb{R}^{p\times n}$, we get

$$AP + PA^\top + BL + L^\top B^\top + \Sigma = 0.$$

Let $\mathcal{A}(P) = AP + PA^\top,\ \mathcal{B}(L) = BL + L^\top B^\top$, which are referred to as Lyapunov maps.

Assume $\mathcal{A}$ is invertible, then we have the relation

$$\mathcal{A}(P) + \mathcal{B}(L) + \Sigma = 0. \tag{3.4}$$

Indeed, once we fix the system and any stabilizing controller A, B, K, the matrices P as well as $L = KP$ are uniquely determined. P is the Grammian matrix

$$P = \int_0^\infty e^{t(A+BK)}\ \Sigma\ e^{t(A+BK)^\top}\ \mathrm{d}t. \tag{3.5}$$

The matrix P is positive definite if $\Sigma \succ 0$. We are interested in the cost function $\mathcal{L}(K)$ when $K \in \mathcal{S}_{K,\text{sta}}$, which corresponds to (3.2) by inserting $u(t) = Kx(t)$,

$$\mathcal{L}(K) = \begin{cases} \mathbf{tr}((Q + K^\top RK)P), & K \in \mathcal{S}_{K,\text{sta}}; \\ +\infty, & K \notin \mathcal{S}_{K,\text{sta}}. \end{cases} \tag{3.6}$$

One can construct a bijection from P, L to K, and prove that, if we minimize $f(L, P)$ subject to (3.4), the optimizer P^*, L^* will map to the optimal K^*, and this minimization problem is convex.

Convex parameterization for continuous time LQR: Suppose the dynamics and costs are (3.1) and (3.2), and let $\mathbf{E}(x_0 x_0^\top) = \Sigma \succ 0$. Denote the (static) state feedback controller by K, so that $u(t) = Kx(t)$. The optimal control problem is

$$\min_K\ \mathcal{L}(K), \quad \text{s.t.} \quad K \in \mathcal{S}_{K,\text{sta}} \tag{3.7}$$

where $\mathcal{L}(K)$ is the cost in (3.2) with $u = Kx$. This problem can be expressed as the following equivalent convex problem,

$$\min_{L,P,Z} \; f(L,P,Z) := \mathbf{tr}(QP) + \mathbf{tr}(ZR) \tag{3.8a}$$

$$\text{s.t. } \mathcal{A}(P) + \mathcal{B}(L) + \Sigma = 0, \;\; P \succ 0, \tag{3.8b}$$

$$\begin{bmatrix} Z & L \\ L^\top & P \end{bmatrix} \succeq 0. \tag{3.8c}$$

The connection between the two problems is distilled in Sec. 3.3. For all feasible (L, P, Z) triplets in (3.8), we can take the first two elements (L, P), and they form a bijection with all stabilizing controllers K in (3.7). The cost function values are equal under the bijection. So we can solve for L^*, P^*, and $K^* = L^*(P^*)^{-1}$.

3.3 Main result

In this section, we propose the main theorem. We first destill the property of the convex parameterization for continuous time LQR. The nonconvex cost function in the policy space has a convex counterpart, whose reparameterization is a smooth bijection and the cost value maps between the two parameterized forms. Based on that, we propose the main theorem: the norm of gradient of the nonconvex cost in policy space is lower bounded by the suboptimality gap from the global minimum.

Motivated by methods that use gradient descent in the policy space, we ask whether running a gradient-based algorithm and getting $\nabla_K \mathcal{L}(K) = 0$ for some K in fact gives the globally optimum K^*. Fazel et al. (2018); Mohammadi et al. (2019b) show the coercivity and gradient dominance property of $\mathcal{L}(K)$ for discrete- and continuous-time LQR respectively. In this chapter, we generalize these results from the special case of continuous-time LQR to a much broader set of control problems, showing the gradient dominance property of the nonconvex costs as functions of policy.

We present our main result in Theorem 2. We consider a pair of problems satisfying Assumptions 4, 6. In Sec. 3.4 we catalog a number of examples showing the generality of this result.

We begin by considering an abstract description of the pair of problems (3.7) and (3.8). These problem descriptions cover LQR as discussed in the last section, as well as more problems discussed in Sec. 3.4. Consider the problems

$$\min_{K} \quad \mathcal{L}(K), \quad \text{s.t. } K \in \mathcal{S}_K, \tag{3.9}$$

and

$$\min_{L,P,Z} \quad f(L,P,Z), \quad \text{s.t.} \quad (L,P,Z) \in \mathcal{S}, \tag{3.10}$$

where the sets $\mathcal{S}_K, \mathcal{S}$ capture the control constraints. They are defined differently for each specific example in Sec. 3.4. For example, for continuous time LQR, $\mathcal{S}_K$ is the set of all stabilizing controllers (3.7) and $\mathcal{S}$ is the intersection of (3.8b) & (3.8c). In infinite horizon problems, we need a stabilizing K so that $\mathcal{S}_K$ is equal to or a subset of the set of stabilizing controllers. We allow special cases when (3.10) depends only on L, P,

$$\min_{L,P} \quad f(L,P), \quad \text{s.t.} \quad (L,P) \in \mathcal{S}. \tag{3.11}$$

We distill the properties of the two problems (3.9) and (3.10) that will be critical for Theorem 2, and allow us to cover more problems as discussed in Sec. 3.4.

Assumption 4. *The feasible set $\mathcal{S}$ is convex in (L,P,Z). The cost function $f(L,P,Z)$ is convex, bounded, and differentiable over an open domain that contains the set $\mathcal{S}$.*

Assumption 4 indicates the second problem is convex. Next, we examine the connection between (3.7) and (3.8), formalized in the following assumption.

Assumption 5. *Let P be invertible*[5] *whenever $(L,P,Z) \in \mathcal{S}$. Assume we can express $\mathcal{L}(K)$*

[5] The invertibility of P holds for all instances in Sec. 3.4.

as follows,

$$\mathcal{L}(K) = \min_{L,P,Z} \; f(L,P,Z)$$
$$\text{s.t. } (L,P,Z) \in \mathcal{S},\; LP^{-1} = K.$$

Denote $\nabla\mathcal{L}(K)[V] := \mathbf{tr}(V^\top \nabla\mathcal{L}(K))$ as the directional derivative of $\mathcal{L}(K)$ is the direction V. With the assumptions above, we will present the main theorem.

Theorem 2. *Suppose Assumptions 4,5 hold, and consider the two problems* (3.9) *and* (3.10)*. Let K^* denote the global minimizer of $\mathcal{L}(K)$ in S_K. Then there exist constants $C_1, C_2 > 0$ and a direction V with $\|V\|_F = 1$, in the descent cone of $\mathcal{S}_K$ at K such that,*

1. *if f is convex, the gradient of $\mathcal{L}$ satisfies*[6]

$$\nabla\mathcal{L}(K)[V] \le -C_1(\mathcal{L}(K) - \mathcal{L}(K^*)). \tag{3.12}$$

2. *(a) if f is μ-strongly convex, or*

   (b) let $\mathcal{P}_\mathcal{S}(-\nabla f(L,P,Z))$ be the projection of $-\nabla f(L,P,Z)$ in the descent cone of $\mathcal{S}$ at (L,P,Z), if for any

$$(L,P,Z) = \arg\min_{L',P',Z'} \; f(L',P',Z'),\; \text{s.t. } (L',P',Z') \in \mathcal{S},\; L'(P')^{-1} = K,$$

   we have $\|\mathcal{P}_\mathcal{S}(-\nabla f(L,P,Z))\|_F^2 \ge \mu(f(L,P,Z) - f(L^*,P^*,Z^*))$,

   the gradient of $\mathcal{L}$ satisfies

$$\nabla\mathcal{L}(K)[V] \le -C_2(\mu(\mathcal{L}(K) - \mathcal{L}(K^*)))^{1/2}. \tag{3.13}$$

Remark 1. *The constants in the above theorem can be computed or bounded in a case by case manner, as discussed further in Appendix B.2. They typically depend on the norm of*

[6] We always consider the directional derivative of a feasible direction within descent cone.

system parameters and the radius of the feasible domain[7]*. We study continuous time LQR as an example. Let the sublevel set be where $\mathcal{L}(K) \leq a$. Define*

$$\nu = \frac{\lambda_{\min}^2(\Sigma)}{4} \left(\sigma_{\max}(A)\lambda_{\min}^{-1/2}(Q) + \sigma_{\max}(B)\lambda_{\min}^{-1/2}(R)\right)^{-2},$$

then

$$C_1 = \frac{\nu \lambda_{\min}^{1/2}(Q)\lambda_{\min}^{1/2}(R)}{4a^4} \cdot \min\left\{a^2,\ \nu\lambda_{\min}(Q)\right\}.$$

Mohammadi et al. (2019b) gives another convex formulation with strong convexity and we can get C_2 for that form,

$$C_2 = \frac{\nu}{2a^3} \min\left\{a^2,\ \nu\lambda_{\min}^{1/2}(Q)\lambda_{\min}^{1/2}(R)\right\}.$$

See Appendix B.2 for more details.

We know that $\|\nabla\mathcal{L}(K)\|_F \geq \left|\nabla\mathcal{L}(K)[\frac{V}{\|V\|_F}]\right|$ for any direction V. The lower bound $\|\nabla\mathcal{L}(K)\|_F \gtrsim (\mathcal{L}(K) - \mathcal{L}(K^*))^\alpha$ on the norm of the gradient is known as Lojasiewicz inequality (Lojasiewicz, 1963). The case when $\alpha = 1/2$ is also called the *gradient dominance* property. If this inequality holds for all K, all stationary points of the objective function are global minima, and an iterative method in which the norm of the gradient decreases to zero will have to converge to a global minimum. Nonconvex functions that satisfy Lojasiewicz inequality are easily optimized, compared to those with spurious local minima. In practice, Lojasiewicz inequality often holds in a neighborhood of a local minimum, and Lojasiewicz inequality is typically used as a tool for local convergence analysis (it is rare that Lojasiewicz inequality holds for $\mathcal{L}(K)$ globally, but it holds for example problems in this chapter).

Next, we consider a stronger assumption covered by Assumption 5, where we assume

[7] Although the set can be unbounded, when we run gradient descent with respect to $\mathcal{L}(K)$, the cost is typically bounded by the initial value $\mathcal{L}(K_0)$ so the iterates are in a sublevel set, therefore boundedness of this sublevel set suffices for our purpose.

that there is a bijection of a specific form between K and (L, P) (Assumption 6). This is true for many control problems including continuous time LQR. Theorem 2 also holds with Assumptions 4, 6. We emphasize the special case with the bijection for illustration. (In this section, the map between the variables is $K = LP^{-1}$, and in Sec. 3.5, we will present the result for a general $K = \Phi(P)$) In Sec. 3.7, we will illustrate the key proof steps: we use the fact that the convex function $f(L, P)$ is gradient dominant, and apply the bijection between K and (L, P) to calculate $\nabla\mathcal{L}(K)$.

Assumption 6. *1. (Bijection between the two feasible sets) Let P be invertible, and let $K = LP^{-1}$ define a bijection[8] $K \leftrightarrow (L, P)$, where there exists an auxiliary variable Z such that $(L, P, Z) \in \mathcal{S}$.*

2. (Equivalence of functions) Choose a controller $K \in \mathcal{S}_K$ with corresponding $(L, P) \in \mathcal{S}$. Then $\mathcal{L}(K) = \min_Z f(L, P, Z)$ subject to $(L, P, Z) \in \mathcal{S}$.

Theorem 2 suggests that when the original nonconvex optimization problem can be mapped to a convex optimization problem that satisfies Assumptions 4, 5 or 4, 6, all stationary points of the nonconvex objective are global minima. So if we can evaluate the gradient of nonconvex objective and run gradient descent algorithm, the iterates converge to the optimal controller.

3.4 Control problems covered by main theorem

Theorem 2 requires an optimal control problem (3.9), and its convex form (3.10) that satisfies a few assumptions. This is an abstract and general description that does not need the exact continuous time LQR formulation in Sec. 3.2. Sec. 3.2 implies that the continuous time LQR satisfies Assumptions 4,6, thus we can directly apply Theorem 2 to argue that the continuous time LQR cost $\mathcal{L}(K)$ satisfies (3.12).

In this section, we discuss more examples, showing that Theorem 2 covers a wide range of control design problems. This illustrates the **generality** of Theorem 2. If a new control

[8] Note that generally $K = LP^{-1}$ cannot guarantee a bijection. However bijection is possible with the extra constraint $(L, P) \in \mathcal{S}$.

problem is encountered, the assumptions for Theorem 2 can be checked, in order to directly conclude that the stationary points of the original cost function are all global minima, and further, the nonconvex function can be globally optimized by policy optimization.

3.4.1 Discrete time infinite horizon LQR

We will show that minimizing the LQ cost as a function of the state feedback controller K, and the convex form, satisfy the assumptions for Theorem 2. So that all stationary points of the LQ cost as a function of K are global minima, same as the result in Fazel et al. (2018).

We consider a discrete time linear system

$$x(t+1) = Ax(t) + Bu(t),\ x(0) = x_0,$$

The goal is to find a state feedback controller K such that the cost function

$$\mathcal{L}(K) = \mathbf{E}_{x_0} \sum_{i=0}^{\infty} x(t)^\top Q x(t) + u(t)^\top R u(t),\ u(t) = Kx(t)$$

is minimized. In other words, we will solve

$$\min_{K}\ \mathcal{L}(K),\ \text{s.t. } K \text{ stabilizes.} \tag{3.14}$$

Here we assume that $\mathbf{E}(x_0 x_0^\top) = \Sigma$. Similar to the continuous time system, one can choose the same parameterization P, L, Z and another PSD matrix $G \in \mathbb{R}^{n\times n} \succeq 0$ and solve the following problem

$$\min_{L,P,Z,G}\ f(L,P,Z,G) := \mathbf{tr}(QP) + \mathbf{tr}(ZR), \tag{3.15a}$$

$$\text{s.t. } P \succ 0,\ G - P + \Sigma = 0, \tag{3.15b}$$

$$\begin{bmatrix} Z & L \\ L^\top & P \end{bmatrix} \succeq 0,\ \begin{bmatrix} G & AP + BL \\ (AP+BL)^\top & P \end{bmatrix} \succeq 0. \tag{3.15c}$$

The goal is to argue that $\mathcal{L}(K)$ and (3.15) has the connection such that Theorem 2 applies, so that the stationary point of $\mathcal{L}(K)$ has to be the global optimum.

Lemma 9. *The LQR problems* (3.14) *and* (3.15) *satisfy Assumptions 4, 5.*

Proof. (3.15) is a convex optimization problem. Now we prove Assumption 5, i.e., we prove that $L(K)$ equals the minimum of the problem (3.15) with an extra constraint $K = LP^{-1}$.

- We first minimize over Z, the minimizer is $Z = LP^{-1}L^\top$. Now we plug $Z = LP^{-1}L^\top$ into cost, replace L by KP and the cost becomes $\mathbf{tr}((Q + K^\top RK)P)$.
- We will eliminate G by

$$G - P + \Sigma = 0, \quad \begin{bmatrix} G & AP + BL \\ (AP + BL)^\top & P \end{bmatrix} \succeq 0.$$

  Using Schur complement, it is equivalent to

$$(AP + BL)P^{-1}(AP + BL)^\top - P + \Sigma \preceq 0.$$

  Plug in $L = KP$, we have

$$(A + BK)P(A + BK)^\top - P + \Sigma \preceq 0.$$

  The cost does not involve G so it does not change.

- Now, we need to prove that $\mathcal{L}(K)$ is equal to

$$\begin{aligned} &\min_P \ \mathbf{tr}((Q + K^\top RK)P), \\ &\text{s.t. } (A + BK)P(A + BK)^\top - P + \Sigma \preceq 0. \end{aligned} \tag{3.16}$$

  The constraint (3.16) can be written as

$$(A + BK)P(A + BK)^\top - P + \Theta = 0, \ \Theta \succeq \Sigma.$$

- Denote the solution to $(A+BK)P(A+BK)^\top - P + \Theta = 0$ as $P(\Theta)$. $P(\Theta)$ for all $\Theta \succeq \Sigma$ covers the feasible points of (3.16). $P(\Theta)$ is expressed as:

$$P(\Theta) = \sum_{t=0}^{\infty}(A+BK)^t\Theta((A+BK)^\top)^t.$$

  So $P(\Theta) \succeq P(\Sigma)$, for all $\Theta \succeq \Sigma$. Since Q and $K^\top RK$ are positive semidefinite, the cost $\mathbf{tr}((Q+K^\top RK)P)$ achieves the minimum at $P = P(\Sigma)$.

- At the end, $P(\Sigma)$ is the Grammian $\boldsymbol{E}\sum_{t=0}^{\infty} x(t)x(t)^\top$ when $\boldsymbol{E}x(0)x(0)^\top = \Sigma$. We studied the connection between continous time Grammian (3.5) and the cost (3.6), and a similar result holds for discrete time LQR:

$$\mathbf{tr}((Q+K^\top RK)P(\Sigma)) = \mathcal{L}(K).$$

□

We built the connection between minimizing $\mathcal{L}(K)$, and the convex optimization (3.15). We argued this pair of problems satisfies the assumptions of Theorem 2. Theorem 2 suggests that $\mathcal{L}(K)$ is gradient dominant, so we can approach K^* by gradient descent on K. This is essentially the conclusion of Fazel et al. (2018); Bu et al. (2019a). Note that the proof of discrete time LQR (Fazel et al., 2018; Bu et al., 2019a) and continuous time LQR (Mohammadi et al., 2019b; Bu et al., 2020) cannot trivially extend to each other, but our result can cover both continuous and discrete time cases.

3.4.2 LQR with Markov jump linear system

We generalize the discrete time linear system to multiple linear systems with transitions, called Markov jump linear system in this part. We show that, the LQR with Markov jump linear system can be covered by the conclusion of Theorem 2. It means all stationary points of the linear quadratic cost as a function of policy/controllers are global minima.

Markov jump linear system. Suppose there are N linear systems, the i-th one being

$$x(t+1) = A_i x(t) + B_i u(t).$$

Now we study the LQR of Markov jump linear system (Jansch-Porto et al., 2020). At each time t, the dynamics linking $x(t+1)$ and the past state and input $x(t), u(t)$ is given by

$$x(t+1) = A_{w(t)} x(t) + B_{w(t)} u(t),\ w(t) \in [N] := \{1, ..., N\}.$$

At time t, a system $w(t)$ from number 1 to N is randomly chosen by some probabilistic model. The transition of the linear systems, or the transition of $w(t)$, follows the following probabilistic model

$$\mathbf{Pr}(w(t+1) = j | w(t) = i) = \rho_{ij} \in [0,1],\ \forall t \geq 0.$$

Suppose $\mathbf{Pr}(w(0) = i) = p_i$. For the i-th system, we will use a state feedback controller K_i. Let $K = [K_1, ..., K_N]$. Define the cost as

$$\mathcal{L}(K) = \mathbf{E}_{w,x_0} \sum_{t=0}^{\infty} x(t)^\top Q x(t) + u(t)^\top R u(t),\ \text{s.t. } u(t) = K_{w(t)} x(t),\ \mathbf{Pr}(w(0) = i) = p_i.$$

The nonconvex problem we target to solve is

$$\min_K\ \mathcal{L}(K), \quad \text{s.t. } \mathcal{L}(K) \text{ is finite.} \tag{3.17}$$

Convex formulation. We propose the following convex formulation. Denote $\boldsymbol{X}_0, \boldsymbol{X}_1, ..., \boldsymbol{X}_N \in \mathbb{R}^{n\times n}$, $L_1, ..., L_N \in \mathbb{R}^{p\times n}$, $Z_0, Z_1, ..., Z_N \in \mathbb{R}^{p\times p}$, $U_{ji} \in \mathbb{R}^{n\times n}$ for $i, j \in [N]$. The following problem is convex:

$$\begin{aligned}
&\min\ \mathbf{tr}(Q\boldsymbol{X}_0) + \mathbf{tr}(Z_0 R),\\
&\text{s.t. } \boldsymbol{X}_0 = \sum_{i=1}^{N} \boldsymbol{X}_i,\ Z_0 = \sum_{i=1}^{N} Z_i,\ \begin{bmatrix} Z_i & L_i \\ L_i^\top & \boldsymbol{X}_i \end{bmatrix} \succeq 0,\\
&\boldsymbol{X}_i - p_i \Sigma = \sum_{j=1}^{N} U_{ji},\ \begin{bmatrix} \rho_{ji}^{-1} U_{ji} & A_j \boldsymbol{X}_j + B_j L_j \\ (A_j \boldsymbol{X}_j + B_j L_j)^\top & \boldsymbol{X}_j \end{bmatrix} \succeq 0,\ \forall i, j \in [N].
\end{aligned}$$

The mapping between the controller K_i and the new variables are $K_i = L_i(\boldsymbol{X}_i)^{-1}$. When the convex problem is minimized, X_i^* represents the Grammian matrix $\boldsymbol{X}_i^* = \sum_{t=0}^{\infty} \boldsymbol{E}(x(t)x(t)^\top \mathbf{1}_{w(t)=i})$.

We prove that (3.17) and the convex formulation satisfy Assumptions 4, 5 in Appendix B.3.1, so that we apply Theorem 2 to claim that all stationary points of $\mathcal{L}(K)$ are global minima.

3.4.3 Minimizing $\mathcal{L}_2$ gain

We quote from Boyd et al. (1994) the problem of minimizing the $\mathcal{L}_2$ gain with static state feedback controller K and the convex formulation. We can apply Theorem 2 to argue that all stationary points of $\mathcal{L}_2$ gain as a function of K are global minima. The $\mathcal{L}_2$ gain is also the $\mathcal{H}_\infty$ norm of transfer function (Boyd et al., 1994, §6.3.2). This problem has an associated convex optimization problem and we can show that they satisfy Assumptions 4,5.

We consider minimizing the $\mathcal{L}_2$ gain of a closed loop system. The continuous time linear dynamical system is

$$\dot{x} = Ax + Bu + B_w w,\ y = Cx + Du.$$

For any signal z, denote

$$\|z\|_2 := \left(\int_0^\infty \|z(t)\|_2^2 \mathrm{d}t\right)^{1/2}$$

Suppose we use a state feedback controller $u = Kx$, and aim to find the optimal controller K^* that minimizes the $\mathcal{L}_2$ gain. We minimize the squared $\mathcal{L}_2$ gain as

$$\min_K \ \mathcal{L}(K) := (\sup_{\|w\|_2=1} \|y\|_2)^2, \text{ s.t. } u = Kx.$$

This problem can be further reformulated as the formulation in (Boyd et al., 1994, Sec 7.5.1)

$$\begin{aligned} &\min_{L,P,\gamma} \ f(L, P, \gamma) := \gamma, \text{ s.t.} \\ &\begin{bmatrix} AP + PA^\top + BL + L^\top B^\top + B_w B_w^\top & (CP + DL)^\top \\ CP + DL & -\gamma I \end{bmatrix} \preceq 0. \end{aligned} \tag{3.19}$$

The minimum $\mathcal{L}_2$ gain is $\sqrt{\gamma^*}$ and $K^* = L^* P^{*-1}$. We will show in the Appendix B.3.2 that the above nonconvex and convex problems satisfy Assumptions 4,5. Thus we can claim that all stationary points of $\mathcal{L}(K)$ are global minima.

3.4.4 Dissipativity

We quote from Boyd et al. (1994) the problem of maximizing the dissipativity with static state feedback controller K and the convex formulation, and apply Theorem 2 to show that all stationary points of the dissipativity as a function of K are global minima.

We study the dynamical system

$$\dot{x} = Ax + Bu + B_w w, \ y = Cx + Du + D_w w \tag{3.20}$$

The notion of dissipativity can be found in (Boyd et al., 1994, §6.3.3, §7.5.2). Our goal is to maximize the dissipativity, which is defined and formulated as with a convex parameterization

(Boyd et al., 1994, §7.5.2).

The dissipativity is defined as all $\eta > 0$ (if it exists, we usually take the maximum one) that satisfy the following inequality for all w and all $T > 0$,

$$\int_0^T w^\top y - \eta w^\top w \mathrm{d}t \geq 0.$$

We use a state feedback controller K, and the goal is to find K^* that maximizes the dissipativity η. Same as before, let K be factorized as LP^{-1}. We can maximize the dissipativity η as a function of K. From the formulation in (Boyd et al., 1994, §7.5.2), we maximize η subject to the dissipativity constraint (3.21),

$$\begin{aligned} &\max_{\eta, L, P} \ \eta, \\ &\text{s.t.} \ \begin{bmatrix} AP + PA^\top + BL + L^\top B^\top & B_w - PC^\top - (DL)^\top \\ B_w^\top - CP - DL & 2\eta I - (D + D^\top) \end{bmatrix} \preceq 0. \end{aligned} \tag{3.21}$$

We can claim that all stationary points of $\mathcal{L}(K)$ are global minima.

3.4.5 System level synthesis (SLS) for finite horizon time varying discrete time LQR

In this part, we switch to the discrete time system in finite horizon. We study the finite horizon time varying LQR problem, and its solution using SLS, and show that it satisfies Assumptions 4,6. Hence we can apply Theorem 2 to conclude that all stationary points of the nonconvex objective functions are global minima.

This problem and its convex form are introduced in Anderson et al. (2019). We consider the following linear dynamical system

$$x(t+1) = A(t)x(t) + B(t)u(t) + w(t) \tag{3.22}$$

over a finite horizon $0, \dots T$. Let the state be x and the input be u. Define

$$
\begin{aligned}
X &= \begin{bmatrix} x(0) \\ ... \\ x(T) \end{bmatrix}, \; U = \begin{bmatrix} u(0) \\ ... \\ u(T) \end{bmatrix}, \\
W &= \begin{bmatrix} x(0) \\ w(0) \\ ... \\ w(T-1) \end{bmatrix}, Z = \begin{bmatrix} 0 & 0 & ... & 0 & 0 \\ I & 0 & ... & 0 & 0 \\ 0 & I & ... & 0 & 0 \\ ... & & & & \\ 0 & 0 & ... & I & 0 \end{bmatrix}, \\
\mathcal{A} &= \mathrm{diag}(A(0), ..., A(T-1), 0), \\
\mathcal{B} &= \mathrm{diag}(B(0), ..., B(T-1), 0).
\end{aligned}
$$

Now we consider the time varying controller K that links state and input as

$$u(t) = \sum_{i=0}^{t} K(t, t-i)x(i), \tag{3.23}$$

and let

$$
\mathcal{K} = \begin{bmatrix} K(0,0) & 0 & ... & 0 \\ K(1,1) & K(1,0) & ... & 0 \\ ... & & & \\ K(T,T) & K(T,T-1) & ... & K(T,0) \end{bmatrix}.
$$

We will minimize some cost function with the constraint. For example, in the discrete time LQR regime (more examples of nonquadratic cost in (Anderson et al., 2019, Sec 2.2)), let the input be (3.23) and define

$$\mathcal{L}(\mathcal{K}) = \sum_{t=0}^{T} x(t)^\top Q(t)x(t) + u(t)^\top R(t)u(t), \tag{3.24}$$

here $Q(t), R(t) \succeq 0$. We will minimize $\mathcal{L}(\mathcal{K})$ where $\mathcal{K}$ is the variable.

Parameterization: The dynamics (3.22) can be written as

$$X = Z\mathcal{A}X + Z\mathcal{B}U + W = Z(\mathcal{A} + \mathcal{B}\mathcal{K})X + W$$

We define the mapping from W to X, U by

$$\begin{bmatrix} X \\ U \end{bmatrix} = \begin{bmatrix} \Phi_X \\ \Phi_U \end{bmatrix} W.$$

where Φ_X, Φ_U are block lower triangular. There is a constraint on Φ_X, Φ_U:

$$\begin{bmatrix} I - Z\mathcal{A} & -Z\mathcal{B} \end{bmatrix} \begin{bmatrix} \Phi_X \\ \Phi_U \end{bmatrix} = I. \tag{3.25}$$

It is proven in (Anderson et al., 2019, Thm. 2.1) that $\mathcal{K} = \Phi_U \Phi_X^{-1}$, $\mathcal{K}$ and Φ_X, Φ_U is a bijection given Φ_X, Φ_U satisfying (3.25).

Let $\mathcal{Q} = \text{diag}(Q(0), ..., Q(T))$, $\mathcal{R} = \text{diag}(R(0), ..., R(T))$, the LQR cost with $x(0) \sim \mathcal{N}(0, \Sigma)$ and no noise is

$$f(\Phi_X, \Phi_U) = \left\| \text{diag}(\mathcal{Q}^{1/2}, \mathcal{R}^{1/2}) \begin{bmatrix} \Phi_X(:,0) \\ \Phi_U(:,0) \end{bmatrix} \Sigma^{1/2} \right\|_F^2,$$

$\Phi_X(:,0), \Phi_U(:,0)$ are the first n columns of Φ_X, Φ_U. The LQR cost with $x(0), w(t)$ being i.i.d from $\mathcal{N}(0,\Sigma)$ is

$$f(\Phi_X,\Phi_U) = \left\| \operatorname{diag}(\mathcal{Q}^{1/2},\mathcal{R}^{1/2}) \begin{bmatrix} \Phi_X \\ \Phi_U \end{bmatrix} (I_{T+1}\otimes\Sigma^{1/2}) \right\|_F^2 .$$

The symbol $\otimes$ means Kronecker product. If we solve

$$\min_{\mathcal{K}} \ \mathcal{L}(\mathcal{K}), \ \mathcal{K} \text{ is block lower left triangular}$$

with the above two costs of $w(t)$, both can be minimized with constraint (3.25):

$$\min_{\Phi_X,\Phi_U} \ f(\Phi_X,\Phi_U), \ \text{s.t.} \ \begin{bmatrix} I - Z\mathcal{A} & -Z\mathcal{B} \end{bmatrix} \begin{bmatrix} \Phi_X \\ \Phi_U \end{bmatrix} = I,$$
$$\Phi_X,\Phi_U \text{ are block lower left triangular.}$$

This problem is convex. The theorem (Anderson et al., 2019, Thm. 2.1) suggests the relation between $\mathcal{L}$ and f satisfying Assumption 6 for Theorem 2. With Theorem 2, we can argue that all stationary points of $\mathcal{L}(\mathcal{K})$ are global minimum.

The paper Alonso et al. (2021) proposes some generalization of SLS. It introduces a localization constraint, where the state is constrained in a convex set. For example, the constraint is (Alonso et al., 2021, Eq. (9))

$$\Phi_X(:,0)x_0 \in \mathcal{P}$$

for a convex set $\mathcal{P}$. We can add it to the problem in convex parameterized problem and map it as a constraint in the controller K space. The nonconvex problem is still gradient dominant.

3.5 A more general description of Assumption 5

In this section, we will give a more general theorem, based on replacing the map $K = LP^{-1}$ by arbitrary function Φ defined below. This allows the theorem to cover more examples in Sec. 3.6.

We chose $K = LP^{-1}$ because this is frequently used for the convex parameterization of the optimal control problem. For example, with the continuous time LQR problem motivated in Sec. 3.2, the mapping between K and L, P is almost the only widely used convex parameterization method. If we choose another change of variable, the resulting objective function is usually not convex in the new variables.

On the other hand, although the mapping $K = LP^{-1}$ is studied, we can generalize Theorem 2 with arbitrary mappings if the reformulated problem is convex – the new mappings still have to satisfy a few assumptions to preserve the Lojasiewicz inequality.

Here we will propose the following assumptions which replace the mapping $K = LP^{-1}$ by an abstract mapping Φ.

Suppose we consider the problems

$$\min_K \quad \mathcal{L}(K), \quad \text{s.t. } K \in \mathcal{S}_K, \tag{3.26}$$

and

$$\min_P \quad f(P), \quad \text{s.t.} \quad P \in \mathcal{S}. \tag{3.27}$$

The matrix P can be a concatenation of many variables, just as a shortlisted expression. For example, P represents (P, L, Z) of continuous LQR. We will study the original optimization problem (3.26), and map it to a convex optimization problem (3.27) where the mapping between K and the variable of the other problem P is abstractly denoted by $K = \Phi(P)$ in (3.28).

Assumption 7. *The feasible set $\mathcal{S}$ is convex in P. The cost function $f(P)$ is convex, finite* *and differentiable in $P \in \mathcal{S}$. $\mathcal{L}(K)$ is Lipschitz in K.*

Assumption 8. *Assume we can express $\mathcal{L}(K)$ as:*

$$\mathcal{L}(K) = \min_{P} \; f(P), \; \textit{s.t. } P \in \mathcal{S}, \; K = \Phi(P). \tag{3.28}$$

And we assume the first order Taylor expansion of the mapping Φ is well defined as

$$\Phi(P + \mathrm{d}P) = \Phi(P) + \Psi(P)[\mathrm{d}P] + o(\mathrm{d}P).$$

for any $P \in \mathcal{S}$ and any perturbation $\mathrm{d}P$ such that $\mathrm{d}P$ is in the descent cone of $\mathcal{S}$ at P.

We mentioned that, P represents (P, L, Z) in continuous LQR. And we can see that Assumption 5 is very similar to Assumption 8. We just apply $\Phi(P, L, Z) = LP^{-1}$ and get Assumption 5.

As a description of the connection between the controller and its parameterization, Assumption 8 is more general than Assumption 6. In Assumption 6, if Z does not exist, it means that the two parameterization and cost functions are diffeomorphic, so that the minimums of the two cost functions map to each other. Assumption 8 is more general with a surjective map. However, if assume the right hand side of the following equation is unique for any K and define

$$g(K) = \arg\min_{P} \; f(P), \; \text{s.t. } P \in \mathcal{S}, \; K = \Phi(P).$$

Then $g(\cdot)$, whose inverse exists, gives a bijection between K and P. And we have $\mathcal{L}(K) = f(P) = f(g(K))$ and $\mathcal{L}(g^{-1}(P)) = f(P)$. Suppose g is smooth, then this gives a diffeomorphism between $K \in \mathcal{S}_K$ and $P \in \{P' \mid \exists K \in \mathcal{S}_K, \; g(K) = P'\} := \mathcal{S}_P$. $\mathcal{S}_P$ is a subset of $\mathcal{S}$ and it is a manifold. Generally we cannot claim convexity of a function defined on a manifold. As long as $P \neq P^*$, the Riemannian gradient of $f(P)$ is non-zero. With a diffeomorphism, we can claim that $\nabla\mathcal{L}(K) \neq 0$ as long as $K \neq K^*$.

Remark 2. *Note that, because of* (3.28)*, the assumption does not trivially hold for any smooth mapping* Φ *in the very general context. For example, the paper Lasserre (2001) proposes the sum-of-squares method for solving polynomial optimizations, which has a convex parameterization of lifting the problem to a higher dimensional space. We explain the idea in a simple paradigm. let* $x \in \mathbb{R}^2$ *and the objective function is power* 2*. The objective function is*

$$\mathcal{L}(x) = a_1 x_1^2 + a_2 x_1 x_2 + a_3 x_2^2.$$

One can define a matrix $X \in \mathbf{S}^{2\times 2} \succeq 0$ *and a cost function that is linear in* X*,*

$$f(X) = \begin{bmatrix} a_1 & a_2/2 \\ a_2/2 & a_3 \end{bmatrix} X.$$

It can be proven that X^* *is rank-*1*, and it maps to* $\begin{bmatrix} x_1^2 & x_1 x_2 \\ x_1 x_2 & x_2^2 \end{bmatrix}$*. However, the map creates many meaningless points while lifting the dimension – extra points when* X *is rank-*2 *that are not mapped from the original problem* x_1, x_2*, and the extra points do not necessarily satisfy* (3.28)*.*

The following conclusion holds with the above Assumptions 7, 8. It generalizes beyond the specific mapping $\Phi(P, L) = LP^{-1}$ to a more general definition, and we propose some instances of convex formulations with different Φ in the next section. We propose the following theorem and the proof is in Appendix B.1.

Theorem 3. *Denote* $\Delta K = \Psi(P)[P^* - P]$*. Let* $\nabla\mathcal{L}(K)[\Delta K]$ *be the directional derivative of* $\mathcal{L}(K)$ *in direction* ΔK*. Then with Assumptions 7, 8 we have*

$$\nabla\mathcal{L}(K)[\Delta K] \leq \mathcal{L}(K^*) - \mathcal{L}(K).$$

If K is not optimal, then the right hand side is strictly less than 0, which means the directional derivative of $\mathcal{L}$ is not 0. Therefore $\nabla\mathcal{L}(K) = 0$ holds only at the global minima.

Remark 3. *Theorem 3 means that,*

$$\|\nabla\mathcal{L}(K)\|_F \geq -\nabla\mathcal{L}(K)[\frac{\Delta K}{\|\Delta K\|_F}] \geq C(K)(\mathcal{L}(K) - \mathcal{L}(K^*))$$

where

$$\begin{aligned} C(K) &= \|\Psi(P)[P^* - P]\|_F^{-1} = \|\Psi(\Phi^{-1}(K))[\Phi^{-1}(K^*) - \Phi^{-1}(K)]\|_F^{-1} \\ &\geq \|\Psi(P)\|_{\mathrm{op}}^{-1}\|P^* - P\|_F^{-1} \end{aligned}$$

For continuous time LQR, P represents the list of variables (P, L, Z) there. Remember

$$\nu = \frac{\lambda_{\min}^2(\Sigma)}{4}\left(\sigma_{\max}(A)\lambda_{\min}^{-1/2}(Q) + \sigma_{\max}(B)\lambda_{\min}^{-1/2}(R)\right)^{-2}.$$

In the sublevel set where $\mathcal{L}(K) \leq a$, we have that

$$\begin{aligned} \|\Psi(P)\|_{\mathrm{op}} &\leq \frac{2a}{\nu}\max\left\{1, \ \frac{a^2}{\nu(\lambda_{\min}(Q)\lambda_{\min}(R))^{1/2}}\right\}, \\ \|P^* - P\|_F &\leq \frac{a}{\lambda_{\min}^{1/2}(Q)}\max\left\{\lambda_{\min}^{-1/2}(Q), \lambda_{\min}^{-1/2}(R)\right\}. \end{aligned}$$

3.6 Control problems with generalized map

This section will cover examples where the parameterization is based on the general map Φ, not necessarily $\Phi(P, L) = LP^{-1}$. We can apply Theorem 3 to these problems.

3.6.1 Distributed finite horizon LQR

(Mårtensson, 2012, Ch. 3) is an *empirical* study (i.e., proposing an algorithm without a proof of convergence) of the gradient descent method for distributed control synthesis. For such a problem, the controller is distributed with a graph structure, showing the accessibility of the distributed controllers to the states: if controller i has no access to state j, then $K_{ij} = 0$, otherwise $K_{ij} \in \mathbb{R}$. Thus there is an extra subspace constraint regarding the graph structure of K, and (Mårtensson, 2012, Ch. 3) applies projected gradient descent on (3.2) with respect to K. It allows a fixed or random of initial state as in (3.2). Generally it is NP-hard to find a global optimum with the subspace constraint, so the paper only proposes an algorithm without a proof.

With an extra condition called quadratic invariance, the problem is not NP-hard. We review the solutions in Furieri et al. (2020) with the connection to our framework.

We consider the time varying linear system

$$\begin{aligned} x(t+1) &= A(t)x(t) + B(t)u(t) + w(t), \\ y(t) &= C(t)x(t). \end{aligned}$$

This is in finite time horizon $t = 0, ..., T$. The state evolution is same as the setup in our SLS example (Sec. 3.4.5), and we can use the same notations $X, U, W, Z, \mathcal{A}, \mathcal{B}$. We further define

$$Y = \begin{bmatrix} y(0) \\ ... \\ y(T) \end{bmatrix}, \ V = \begin{bmatrix} v(0) \\ ... \\ v(T) \end{bmatrix}, \ \mathcal{C} = \operatorname{diag}(C(0), ..., C(T)).$$

Now we will consider the control policy

$$u(t) = \sum_{i=0}^{t} K(t, t-i)y(i).$$

The search space of policy is same as SLS, and we define $\mathcal{K}$ matrix in the same way. Furieri et al. (2020) studies the problem under the context of distributed control. One searches for the controller $K \in \mathcal{S}_K$ where $\mathcal{S}_K$ a subset of controllers. In distributed control, there is a graph model for controllers such that the i-th controller might not be able to access the state j for (i, j) in a set of indices $\mathcal{S}_{\text{idx}}$. In this case, $K_{i,j} = 0$ is an extra constraint for the control problem. Therefore, if one searches for the optimal controller in $\mathcal{S}_K$, we can define the subspace

$$\mathcal{S}_K := \{K \mid K_{i,j} = 0,\ \forall (i,j) \in \mathcal{S}_{\text{idx}}\}.$$

The extra constraint is not always easily handled, but (Furieri et al., 2020, §3) proposes an extra assumption, called quadratic invariance (QI), and introduces the equivalent convex optimization.

Remember we defined

$$\mathcal{K} = \begin{bmatrix} K(0,0) & 0 & ... & 0 \\ K(1,1) & K(1,0) & ... & 0 \\ ... & & & \\ K(T,T) & K(T,T-1) & ... & K(T,0) \end{bmatrix},\ \mathcal{C} = \text{diag}(C(0), ..., C(T)).$$

And we define

$$P_{11} = (I - Z\mathcal{A})^{-1},\ P_{12} = (I - Z\mathcal{A})^{-1} Z\mathcal{B}.$$

QI means that, for all $\mathcal{K} \in \mathcal{S}_K$, $\mathcal{K}\mathcal{C}P_{12}\mathcal{K} \in \mathcal{S}_K$.

The cost function is:

$$\mathcal{L}(\mathcal{K}) = \sum_{t=0}^{T} y(t)^\top Q(t) y(t) + u(t)^\top R(t) u(t).$$

Define

$$\Phi(\mathcal{G}) = (I + \mathcal{G}\mathcal{C}P_{12})^{-1}\mathcal{G}.$$

Then we can get a new variable $\mathcal{G}$ and a function Φ. With $\mathcal{K} = \Phi(\mathcal{G})$, the cost can be proven to be convex in $\mathcal{G}$. The variable $\mathcal{G}$ is in the same subspace as $\mathcal{K}$ determined by $\mathcal{S}_K$. Indeed, the mapping satisfies Assumptions 7, 8, and the exact formulation of the two optimization problems are described in (Furieri et al., 2020, Append. A, Lem. 5). Define $\mathcal{Q} = \text{diag}(Q(0), ..., Q(T))$, $\mathcal{R} = \text{diag}(R(0), ..., R(T))$. Let $w(t)$ be Gaussian random vectors with stationary covariance, $w(t_1)$ and $w(t_2)$ are independent $\forall t_1 \neq t_2$. $\Sigma_w = I_T \otimes \text{Cov}(w)$ ($\otimes$ means Kronecker product), $\Sigma_x = \text{diag}(\boldsymbol{E}(x_0 x_0^\top), 0, ..., 0)$. The convex cost function takes the form

$$f(\mathcal{G}) = \left\| \mathcal{Q}^{1/2}\mathcal{C}(I + P_{12}\mathcal{G}\mathcal{C})P_{11} \begin{bmatrix} \Sigma_w^{1/2} & \Sigma_x^{1/2} \end{bmatrix} \right\|_F^2 + \left\| \mathcal{R}^{1/2}\mathcal{G}\mathcal{C}P_{11} \begin{bmatrix} \Sigma_w^{1/2} & \Sigma_x^{1/2} \end{bmatrix} \right\|_F^2 .$$

In summary, we have a pair of problems: 1) minimize $\mathcal{L}(\mathcal{K})$ over $\mathcal{K}$ and 2) minimize $f(\mathcal{G})$ over $\mathcal{G}$. They are related under the Assumptions 4, 6 of Theorem 2. Thus we can claim via Theorem 2 that, all stationary points of $\mathcal{L}(\mathcal{K})$ are global minima.

3.7 Proof sketch

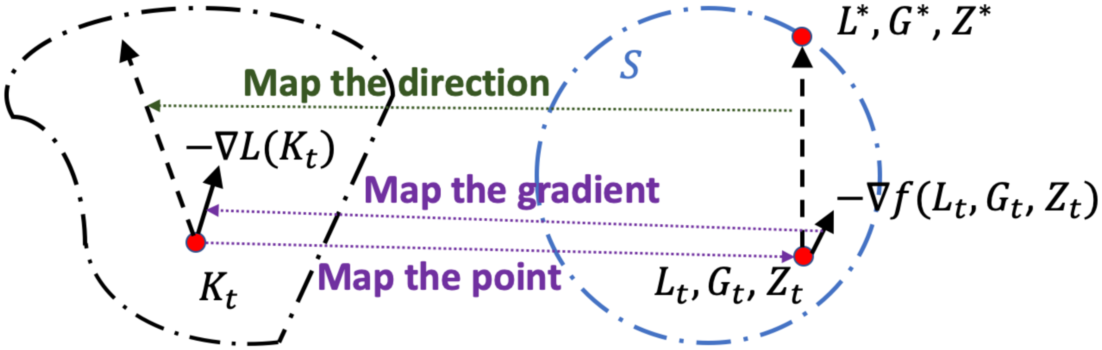


Figure 3.1: Mapping between nonconvex and convex landscapes. Suppose we run gradient descent at iteration t, for any controller K, we can map it to L, P, Z in the other parameterized space. and then we map the direction $(L^*, P^*, Z^*) - (L, P, Z)$ and the gradient $\nabla f(L, P, Z)$ back to the original K space. Since in (L, P, Z) space the objective function is convex, then $\langle \nabla f(L, P, Z), (L^*, P^*, Z^*) - (L, P, Z) \rangle < 0$. We prove that similar correlation holds for the nonconvex objective.

We put the full proof of Theorem 2 in Appendix B.1, and give a sketch of the proof in this section. We illustrate the idea in Figure 3.1, which, on the high level, maps the original space of controller K where the cost is nonconvex, and the parameterized space with L, P, Z where the cost is convex.

For simplicity, we sketch the proof using Assumptions 4,6. For any point K, we can find a point (L, P, Z) in the parameterized space. If it is not the optimizer, we can find the line segment linking (L, P, Z) and the optimizer (L^*, P^*, Z^*). Note that the optimization problem is convex in this space so that $\langle \nabla f(L, P, Z), (L^*, P^*, Z^*) - (L, P, Z) \rangle$ is upper bounded by $f(L^*, P^*, Z^*) - f(L, P, Z)$. Then with the assumptions, we can map the directional derivative back to the original K space, and show that the directional derivative in $\mathcal{L}(K)$ is not 0.

3.8 Conclusion and future directions

The future work is to refine the analysis to obtain the best case-specific convergence rates, and to provide an interpretation of the associated constants in terms of control theoretic notions.

We also note that, not all control problems are easy to solve by first order methods in policy space. The distributed control problem is an example. The controller K has a sparsity pattern, i.e., K is in a subspace. Ref. (Li et al., 2021) shows that, generally the set of stabilizing controllers is highly disconnected and the problem is NP-hard without the extra assumption of quadratic invariance in Furieri et al. (2020). Similarly, the static output feedback controller design is NP-hard. The goal is to minimize the LQ cost, but we can only observe an output $y = Cx$ but cannot observe the full state, and we are only allowed to use a static output feedback controller $u = Ky$. If C is not full row rank, the set of stabilizing controllers is also highly disconnected (Feng & Lavaei, 2020). If C is full row rank, the problem almost reduces to state feedback control since one can recover the state x from y, and Ref. Duan et al. (2021) shows that first order policy optimization finds the optimal controller. Ref. Tang et al. (2021) shows that, the cost of the LQR problem with an output-feedback dynamical controller, i.e., the LQG cost, has saddle points (the problem setup in Tang et al. (2021) expresses the dynamic controller in state-space and the optimization variables are the state-space matrices A_K,B_K, C_K for the controller). Although a parameterization can construct an equivalent convex optimization problem, the map for such parameterization is generally not smooth, and the nonsmoothness breaks the gradient dominance and generates saddle points; thus this negative example is not covered by the results in this paper.

We are also interested in understanding the cost landscape of LQG problem (Tang et al., 2021) and output estimation problem (Umenberger et al., 2022), specifically their connection with the convex parameterization, and investigating the second order landscape analysis via the mapping to the convex problem. For such problems, one possible approach is in (Umenberger et al., 2022, Sec. I.2). They propose a regularizer which is a barrier function in the reparameterized space, with a closed form expression in the policy space. The boundary of the barrier consists of singular matrices that breaks the smoothness of the mapping. Hence with the barrier function, when the regularized objective function is finite, we might be able to get a Lipschitz Φ function and can apply Theorem 3.

Chapter 4

LEARNING LINEAR DYNAMICAL SYSTEMS VIA NUCLEAR NORM REGULARIZATION

In this chapter, we investigate the regularzation method for learning low-order linear dynamical systems from input-output data, named as system identification problem. It is known that with appropriate regularizers, the prior information of the structure learning model can be exploited. We show that, with a designed random matrix as input and the Hankel nuclear norm regularizer, one can recover the system using optimal number of observations and achieve strong statistical estimation rates. We propose a training-validation procedure for tuning the regularization weight and accurately selecting the recovered model. Our synthetic experiment shows the strict advantage of regularized algorithm over the unregularized counterpart, and our experiments on real dataset shows that regularized algorithm has lower sample complexity and returns a Hankel matrix with a clear singular value gap.

This work is published as Sun et al. (2020).

4.1 Introduction

System identification is an important topic in control theory. Accurate estimation of system dynamics is the basis of control or policy decision problems in tasks varying from linear-quadratic control to deep reinforcement learning. Consider a linear time-invariant system of order R with the *minimal* state-space representation

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t, \\ y_t &= Cx_t + Du_t + z_t, \end{aligned} \tag{4.1}$$

where $x_t \in \mathbb{R}^R$ is the state, $u_t \in \mathbb{R}^p$ is the input, $y_t \in \mathbb{R}^m$ is the output, $z_t \in \mathbb{R}^m$ is the output noise, $A \in \mathbb{R}^{R\times R}$, $B \in \mathbb{R}^{R\times p}$, $C \in \mathbb{R}^{m\times R}$, $D \in \mathbb{R}^{m\times p}$ are the system parameters, and x_0 is the initial state (in this chapter, we assume $x_0 = 0$). Generally with the same input and output, the dimension of the hidden state x can be any number no less than R, and we are interested in the minimum dimensional representation (i.e., minimal realization) in this chapter.

The goal of system identification is to find the system parameters, such as A, B, C, D matrices or impulse response, given input and output observations. If $(C, D) = (I, 0)$, we directly observe the state. A notable line of work derives statistical bounds for system identification with limited *state* observations from a single output trajectory (defined in Fig. 4.2) with a random input Abbasi-Yadkori & Szepesvári (2011); Simchowitz et al. (2018); Sarkar & Rakhlin (2019).

The state evolves as $x_{t+1} = Ax_t + \eta_t$ where η_t is the white noise that provides excitation to states Simchowitz et al. (2018); Sarkar & Rakhlin (2019). They recover A by solving a least-squares problem. The main proof approach comes from an analysis of martingales (Abbasi-Yadkori et al., 2011, Thm 2,3). Simchowitz et al. (2018) assumes that the system is stable whereas Sarkar & Rakhlin (2019) removes the assumptions on the spectral radius of A.

When we do not directly observe the state x (also known as hidden-state), one has only access to u_t and y_t and lack the full information on x_t. We recover the impulse response (also known as the Markov parameters) sequence $h_0 = D$, $h_t = CA^{t-1}B \in \mathbb{R}^{m\times p}$ for $t = 1, 2, \dots$ that uniquely identifies the end-to-end behavior of the system. The impulse response can have infinite length, and we let $h = [D, CB, CAB, CA^2B, \dots, CA^{2n-3}B]^\top$ denote its first $2n - 1$ entries, which can be later placed into an $n \times n$ Hankel matrix. Without knowing the system order, we consider recovering the first n terms of h where n is larger than system order R. To this end, let us also define the Hankel map $\mathcal{H} : \mathbb{R}^{m\times(2n-1)p} \to \mathbb{R}^{mn\times pn}$ as

$$H := \mathcal{H}(h) = \begin{bmatrix} h_1 & h_2 & ... & h_n \\ h_2 & h_3 & ... & h_{n+1} \\ ... & & & \\ h_n & h_{n+1} & ... & h_{2n-1} \end{bmatrix}. \tag{4.2}$$

If $n \geq R$, the Hankel matrix H is of rank R regardless of n (Sontag, 2013, Sec. 5.5). Specifically, we will assume that R is small, so the Hankel matrix is low rank. Our goal is to recover a low rank Hankel matrix. It is known that nuclear norm regularization is used to find a low rank matrix Recht et al. (2010); Fazel et al. (2001), and Fazel (2002) uses it for recovering a low rank Hankel matrix.

Low-rank Hankel matrices arise in a range of applications, from dynamical systems – where the rank corresponds to a low order or MacMillan degree for the system Sontag (2013); Fazel (2002) – to signal processing problems. The latter includes recovering sum of complex exponentials Cai et al. (2016); Xu et al. (2018) (where the rank of the Hankel matrix is the number of summands), shape-from-moments estimation in tomography and geophysical inversion Elad et al. (2004) (where the vertices of an object are probed and the output is a sum of exponentials), and video in-painting Ding et al. (2007) (where the video is regarded as a low order system).

Performance criteria for system identification: To explain our contributions, we introduce common performance metrics. Refs. Oymak & Ozay (2018) and Sarkar et al. (2019) recover the system from single rollout/trajectory ("rollout" is defined in Sec. 4.3) of the input signal, whereas our work, Tu et al. (2017) and Cai et al. (2016) require multiple rollouts. To ensure a standardized comparison, we define *sample complexity* to be the number of equations (equality constraints in variables h_t) used in the problem formulation, which is same as the number of observed outputs (see Fig. 4.2 and Sec. 4.3). With this, we explore the following performance metrics for learning the system from T output measurements.

- **Sample complexity:** The minimum sample size T for recovering system parameters with zero error when the noise is set to $z = 0$. This quantity is lower bounded by the system order. System order can be seen as the “degrees of freedom" of the system.
- **Impulse Response (IR) Estimation Error:** The Frobenius norm error $\|\hat{h} - h\|_F$ for the IR. A good estimate of IR enables the accurate prediction of the system output.
- **Hankel Estimation Error:** The spectral norm error $\|\mathcal{H}(\hat{h} - h)\|$ of the Hankel matrix. This metric is particularly important for system identification as described below.

The Hankel spectral norm error is a critical quantity for several reasons. First, the Hankel spectral norm error connects to the $\mathcal{H}_\infty$ estimation of the system Sanchez-Pena & Sznaier (1998). Secondly, bounding this error allows for robustly finding balanced realizations of the system; for example, the error in reconstructing state-space matrices (A, B, C, D) via the Ho-Kalman procedure is bounded by the Hankel spectral error. Finally, it is beneficial in model selection, as a small spectral error helps distinguish the true singular values of the system from the spurious ones caused by estimation error. Indeed, as illustrated in the experiments, the Hankel singular value gap of the solution of the regularized algorithm is more visible compared to least-squares, which aids in identifying the true order of the system as explored in Sec. 4.8.

Algorithms: Hankel-regularization & OLS. In our analysis, we consider a multiple rollout setup where we measure the system dynamics with T separate rollouts. For each rollout, the input sequence is $u^{(i)} = [u^{(i)}_{2n-1}, ..., u^{(i)}_1] \in \mathbb{R}^{(2n-1)p}$ and we measure the system output at time $2n - 1$. Note that the i^{th} output at time $2n - 1$ is simply $h^\top u^{(i)}$. Define $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$ where the i^{th} row is $u^{(i)}$. Let $y \in \mathbb{R}^{T\times m}$ denote the corresponding observed outputs. Hankel-regularization refers to the nuclear norm regularized problem (HNN).

$$\hat{h} = \arg\min_{h'} \quad \frac{1}{2}\|\bar{\boldsymbol{U}}h' - y\|_F^2 + \lambda\|\mathcal{H}(h')\|_*, \tag{HNN}$$

Finally, setting $\lambda = 0$, we obtain the special case of ordinary least-squares (OLS).

4.2 Contributions

Our main contribution is establishing data-driven guarantees for Hankel nuclear norm regularization and shedding light on the benefit of regularization through a comparison to the ordinary least-squares (OLS) estimator. Specifically, a summary of our findings are as follows.

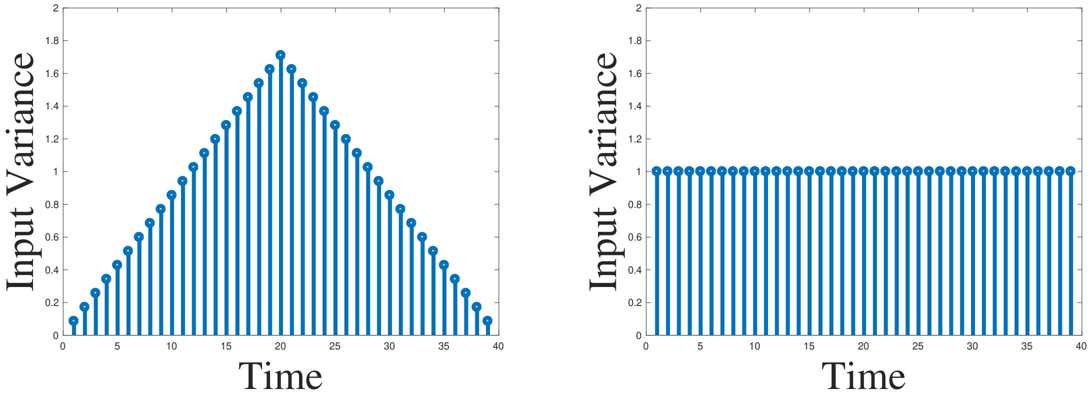
• **Hankel nuclear norm** (Sec. 4.4 & 4.5): For multi-input/single-output (MISO) systems (p input channels), we establish *near-optimal sample complexity* bounds for the Hankel-regularized system identification, showing the required sample size grows as $\mathcal{O}(pR\log^2 n)$ where R is the system order and n is the Hankel size. This result utilizes an *input-shaping* strategy (rather than i.i.d. excitation, see Fig. 4.1a) and builds on Cai et al. (2016) who studied the recovery of a sum-of-exponentials signal. Our bound significantly improves over naive bounds. For instance, without Hankel structure, enforcing low-rank would require $\mathcal{O}(nR)$ samples and enforcing Hankel structure without low-rank would require $\mathcal{O}(n)$ samples.

We also establish finite sample bounds on the IR and Hankel spectral errors. Our rates are on par with the OLS rates; however, unlike OLS, they also apply in the small sample size regime $pn \gtrsim T \gtrsim pR\log^2 n$.

Surprisingly, Sec. 4.5 shows that the *input-shaping* is necessary for the logarithmic sample complexity in n. Specifically, we prove that if the inputs are i.i.d. standard normal (Fig. 4.1b), the minimum number of observations to exactly recover the impulse response in the noiseless case grows as $T \gtrsim n^{1/6}$.

• **Sharpening OLS bounds** (Sec. 4.6): For multi-input/multi-output (MIMO) systems, we establish a *near-optimal spectral error rate* for the Hankel matrix when $T \gtrsim np$. Our error rate improves over that of Oymak & Ozay (2018) and our sample complexity improves over Sarkar et al. (2019) and Tu et al. (2017) which require $\mathcal{O}(n^2)$ samples rather than $\mathcal{O}(n)$. This refinement is accomplished by relating the IR and Hankel errors. Specifically, using the fact that rows of the Hankel matrix are subsets of the IR sequence, we always have the inequality

$$\|\hat{h} - h\|_F/\sqrt{2} \leq \|\mathcal{H}(\hat{h} - h)\| \leq \sqrt{n}\|\hat{h} - h\|_F. \tag{4.3}$$



(a)

(b)

Figure 4.1: (a) Shaped input (where variance of $u(t)$ changes over time): recovery is guaranteed when $T \approx R$; (b) i.i.d input (fixed variance): recovery fails with high probability when $T \lesssim n^{1/6}$. See Sec. 4.5.

Observe that there is a factor of $\sqrt{n}$ gap between the left-hand and right-hand side inequalities. We show that the left-hand side is typically the tighter one, thus $\|\hat{h} - h\|_F \sim \|\mathcal{H}(\hat{h} - h)\|$.

• **Guarantees on accurate model-selection (Sec. 4.7):** The Hankel-regularized algorithm requires a proper choice of the regularization parameter λ. In practice, the optimal choice is data dependent and one usually estimates λ via trial and error based on the validation error. We provide a complete procedure for model selection (training & validation phases), and establish statistical guarantees for it.

• **Contrasting Hankel regularization and OLS (Sec. 4.8):** Finally, we assess the benefits of regularization via numerical experiments on system identification focusing on data collected from a single-trajectory.

We first consider synthetic data and focus on low-order systems with slow impulse-response decay. The slow-decay is intended to exacerbate the FIR approximation error arising from truncating the impulse-response at $2n - 1$ terms. In this setting, OLS as well as Sarkar et al. (2019) are shown to perform poorly. In constrast, Hankel-regularization better avoids the truncation error as it allows for fitting a long impulse-response with few data (due to logarithmic dependence on n).

Our real-data experiments (on a low-order example from the DaISy datasets De Moor et al. (1997)) suggest that the regularized algorithm has empirical benefits in sample complexity, estimation error, and Hankel spectral gap, and demonstrate that the regularized algorithm is less sensitive to the choice of the tuning parameter, compared to OLS whose tuning parameter is the Hankel size n. Finally, comparison of least-squares approaches in Oymak & Ozay (2018) (OLS) and Sarkar et al. (2019) reveals that OLS (which directly estimates the impulse response) performs substantially better than the latter (which estimates the Hankel matrix). This highlights the role of proper parameterization in system identification.

4.2.1 Prior Art

The traditional unregularized methods include Cadzow approach (Cadzow, 1988; Gillard, 2010), matrix pencil method (Sarkar & Pereira, 1995), Ho-Kalman approach (Ho & Kálmán, 1966) and the subspace method raised in Ljung (1999); Van Overschee & De Moor (1995, 2012), further modified as frequency domain subspace method in McKelvey et al. (1996) when the inputs are single frequency (sine/cosine) signals.

The algorithms reduce the rank of the estimated Hankel matrix or the order of the system impulse response in the following ways: Cadzow (1988) uses alternative projections to get a low rank Hankel; Sarkar & Pereira (1995) recovers the subspace of the Hankel matrix by columns of Vandermonde decomposition matrix and the system order is the column space dimension; Ho & Kálmán (1966) recovers system parameters A, B, C, D from a low rank approximation of Hankel matrix estimation, with the size of A, B, C, D corresponding to the system order; Van Overschee & De Moor (2012) rewrites the system dynamics as a relation of input, output and state, leverages the subspaces spanned by them and does system identification. Recent works show that least-squares can be used to recover the Markov parameters. To identify a stable system from a single trajectory, Oymak & Ozay (2018) estimates the impulse response and Sarkar et al. (2019) estimates the Hankel matrix via least-squares. The latter provides optimal Hankel spectral norm error rates, however has suboptimal sample complexity (see the table in Section 4.3). While Oymak & Ozay (2018);

Sarkar et al. (2019) use random input, (Tu et al., 2017, Thm 1.1, 1.2) use impulse and single frequency signal respectively as input. They both recover impulse response. These works assume known system order, or traverse the Hankel size n to fit the system order. Zheng & Li (2020) proves that least-squares can identify any (including unstable) linear systems with multiple rollout data. Reyhanian & Haupt (2021) studies online system identification. It applies online gradient descent on least-squares loss and shows the identification error. Fattahi (2020) shows that, when the system is strictly stable ($\rho(A) < 1$), the sample complexity is only polynomial in $(1 - \rho(A))^{-1}$ and logarithmic in dimension.

There are several interesting generalizations of least squares with non-asymptotic guarantees for different goals. Hazan et al. (2018) and Simchowitz et al. (2019) introduced filtering strategies on top of least squares. The filters in Hazan et al. (2018) is the top eigenvectors of a special deterministic matrix, used for output prediction in stable systems. Simchowitz et al. (2019) uses filters in frequency domain to recover the system parameters of a stable system, Tsiamis & Pappas (2019) gives a non-asymptotic analysis for learning a Kalman filter system, which can also be applied to an auto-regressive setting. As an extension, Dean et al. (2019) and Mania et al. (2019) apply system identification guarantee for robust control, where the system is identified and controlled in an episodic way. Lu & Mo (2021) extended the online LQR to a non-episodic way. Agarwal et al. (2019) studies online control and regret analysis in adversarial setting, whose algorithm directly learns the policy in an end-to-end way. Talebi et al. (2020) controls an unknown unstable system with no initial stabilizing controller. Another area is system identification with non-linearity. Mhammedi et al. (2020) learns a linear system using nonlinear output observations. Oymak (2019); Khosravi & Smith (2020); Foster et al. (2020); Bahmani & Romberg (2019); Sattar & Oymak (2020) consider guarantees for certain nonlinear systems with state observations and Mania et al. (2020); Wagenmaker & Jamieson (2020) study active learning where the new input adapts with respect to previous observations. Rutledge et al. (2020) studies the estimation and proposes the subsequent model-based control algorithm with missing data. Du et al. (2019); Sattar et al. (2021) study clustering and identification for Markov jump system and Du et al. (2021)

further analyzes the optimal control strategy based on the estimated system parameters. Chen & Poor (2022) studies learning mixtures of linear systems from multiple trajectories. Each trajectory is generated by one of a collection of systems while we don't know which one it comes from. The algorithm first identifies the clusters and then identifies each system. The technique about learning mixture data and subspace-based meta-learning is covered in Chapter 5.

Nuclear norm regularization has been shown to recover an unstructured low-rank matrix in a sample-efficient way in many settings (e.g., Recht et al. (2010); Candes & Plan (2010)). The regularized subspace method are introduced in Hansson et al. (2012); Verhaegen & Hansson (2016). Liu et al. (2013); Fazel et al. (2013) propose slightly different algorithms which regress low rank output Hankel matrix. Grossmann et al. (2009) specifies the application of Hankel nuclear norm regularization when some output data are missing. Ayazoglu & Sznaier (2012) proposes a fast algorithm on solving the regularization algorithm. All above regularization works emphasize on optimization algorithm implementation and have no statistical bounds. More recently Cai et al. (2016) theoretically proves that a low order SISO system from multi-trajectory input-outputs can be recovered by this approach. Blomberg (2016) gives a thorough analysis on Hankel nuclear norm regularization applied in system identification, including discussion on proper error metrics, role of rank/system order in formulating the problem, implementable algorithm and selection of tuning parameters.

The rest of the chapter is organized as follows. Next section introduces the technical setup. Sections 4.4 proposes our results on nuclear norm regularization. Section 4.5 discusses the role of the input distribution and establishes lower bounds. Section 4.6 provides our results on least-squares estimator. Section 4.7 discusses model selection algorithms. Finally Section 4.8 presents the numerical experiments.

4.3 Problem Setup and Algorithms

Let $\|\cdot\|, \|\cdot\|_*, \|\cdot\|_F$ denote the spectral norm, nuclear norm and Frobenius norm respectively. Throughout, we estimate the first $2n-1$ terms of the impulse response denoted by h. The

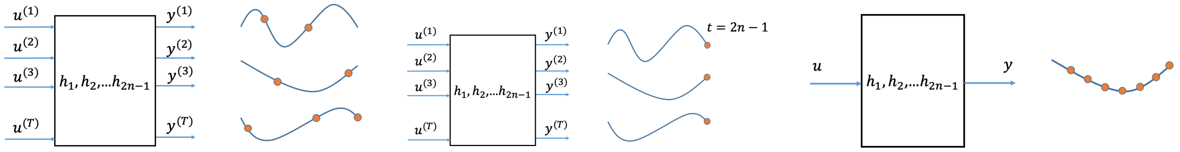


Figure 4.2: (a) Arbitrary sampling on output data, and two specific data aqcuisition models: (b) multi-rollout, and (c) single rollout.

system is excited by an input u over the time interval $[0, t]$ and the output y is measured at time t, i.e.,

$$y_t = \sum_{i=1}^{t} h_{t+1-i} u_i + z_t. \tag{4.4}$$

We start by describing data acquisition models. Generally there are several rounds (ith round is denoted with super script (i) in Fig. 4.2) of inputs sent into the system, and the output can be collected or neglected at arbitrary time. In the setting that we refer to as “multi-rollout" (Fig. 4.2(b)), for each input signal $u^{(i)}$ we take only one output measurement y_t at time $t = 2n - 1$ and then the system is restarted with a new input. Here the *sample complexity* is T, the number of output measurements as well as the round of inputs. Recent papers (e.g., Oymak & Ozay (2018) and Sarkar et al. (2019)) use the “single rollout" model (Fig. 4.2(c)) where we apply an input signal from time 1 to $T + 2n - 2$ without restart, and collect all output from time $2n - 1$ to $T + 2n - 2$, in total T output measurements; we use this model in the numerical experiments in Sec. 4.8.

We consider two estimators in this chapter: the *nuclear norm regularized estimator* and the *least squares estimator* defined later.

We will bound the various error metrics mentioned earlier in terms of the sample complexity T, the true system order R, the dimension of impulse response $n \gg R$, and signal to noise ratio (SNR) defined as $\mathbf{snr} = \mathbb{E}[\|u\|_F^2/n]/\,\mathbb{E}[\|z\|_F^2]$. Table 4.1 provides a summary and comparison of these bounds. All bounds are order-wise and hide constants and log factors. We can see that, with nuclear norm regularization, our result matches the least squares impulse response

Table 4.1: Comparison of recovery error of impulse response. The Hankel matrix is $n \times n$, the system order is R, and the number of samples is T, and $\sigma = 1/\sqrt{\mathbf{snr}}$ denotes the noise level. LS-IR and LS-Hankel stands for least squares regression on the impulse response and on the Hankel matrix.

Paper	This work	This work	Oymak & Ozay (2018)	Sarkar et al. (2019)
Sample complexity	R	n	n	n^2
Method	Nuc-norm	LS-IR	LS-IR	LS-Hankel
Impulse response error	see (4.7)	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$(1+\sigma)\sqrt{n/T}$
Hankel spectral error	see (4.7)	$\sigma\sqrt{n/T}$	$\sigma n/\sqrt{T}$	$(1+\sigma)\sqrt{n/T}$

and Hankel spectral error bound while sample complexity can be as small as $\mathcal{O}(R^2)$, and we can recover the impulse response with guaranteed suboptimal error when sample complexity is $\mathcal{O}(R)$. Our least square error bound matches the best error bounds among Oymak & Ozay (2018) and Sarkar et al. (2019), which is proven optimal for least squares.

Next, we discuss the design of the input signal and introduce input shaping matrix.

Input shaping: Note $\mathcal{H}$ operator does not preserve the Euclidean norm, so Cai et al. (2016) proposes using a normalized operator $\mathcal{G}$, where they first define the weights

$$K_j = \begin{cases} \sqrt{j}, & 1 \le j \le n, \\ \sqrt{2n-j}, & n < j \le 2n-1. \end{cases} \tag{4.5}$$

and let $K \in \mathbb{R}^{(2n-1)p\times(2n-1)p}$ be a block diagonal matrix where the jth diagonal block of size $p \times p$ is equal to $K_j I_{p\times p}$. In other words,

$$K = \begin{bmatrix} K_1 I & 0 & 0 & ... & 0 \\ 0 & K_2 I & 0 & ... & 0 \\ ... & & & & \\ 0 & 0 & 0 & ... & K_{2n-1} I \end{bmatrix}$$

Define the mapping $\mathcal{G}(h) = \mathcal{H}(K^{-1}h)$. In other words, if $\beta = Kh$ then $\mathcal{G}(\beta) = \mathcal{H}(h)$. Define $\mathcal{G}^* : \mathbb{R}^{mn\times np} \to \mathbb{R}^{m\times(2n-1)p}$ as the adjoint of $\mathcal{G}$, where $[\mathcal{G}^*(M)]_i = \sum_{j+k-1=i} M_{(j)(k)}/K_i$ if we denote the j,k-th block of M (defined in (4.2)) by $M_{(j)(k)}$. Using this change of variable and letting $\boldsymbol{U} = \bar{\boldsymbol{U}}K^{-1}$, problem (HNN) can be written as

$$\hat{\beta} = \arg\min_{\beta'} \quad \frac{1}{2}\|\boldsymbol{U}\beta' - y\|_F^2 + \lambda\|\mathcal{G}(\beta')\|_*. \tag{4.6}$$

4.4 Hankel nuclear norm regularization

To promote a low-rank Hankel matrix, we add nuclear norm regularization in our objective and solve the regularized regression problem. Here we give a finite sample analysis for the recovery of the Hankel matrix and the impulse response found via this approach. We consider a random input matrix $\bar{\boldsymbol{U}}$ and observe the corresponding noisy output vector y as in (4.4). We then regress y and $\bar{\boldsymbol{U}}$ such that $y = \bar{\boldsymbol{U}}h + z$ where z is the noise vector.

Theorem 4. *Consider the problem* (HNN) *in the MISO (multi-input single-output) setting (m=1, p inputs). Suppose the system is order R, $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$, each row consists of an input rollout $u^{(i)} \in \mathbb{R}^{(2n-1)p}$, and the scaled $\boldsymbol{U} = \bar{\boldsymbol{U}}K^{-1}$ has i.i.d Gaussian entries. Let $\boldsymbol{snr} = \mathbb{E}[\|u\|^2/n]/\,\mathbb{E}[\|z\|^2]$ and $\sigma = 1/\sqrt{\boldsymbol{snr}}$. Let $\lambda = \sigma\sqrt{\frac{pn}{T}}\log(n)$. Then, the problem* (HNN) *returns $\hat{h}$ such that*

$$\frac{\|\hat{h} - h\|_2}{\sqrt{2}} \leq \|\mathcal{H}(\hat{h} - h)\| \lesssim \begin{cases} \sqrt{\frac{np}{\boldsymbol{snr}\times T}}\log(n) & \textit{if} \quad T \gtrsim \min(R^2, n) \\ \sqrt{\frac{Rnp}{\boldsymbol{snr}\times T}}\log(n) & \textit{if} \quad R \lesssim T \lesssim \min(R^2, n). \end{cases} \tag{4.7}$$

Thm. 4 jointly bounds the impulse response and Hankel spectral errors of the system under mild conditions. We highlight the improvements that our bounds provide: (1) When the system is low order, the sample complexity T is logarithmic in n and improves upon the $O(n)$ bound of the least-squares algorithm. (2) The error rate with respect to the system parameters n, R, T is same as Oymak & Ozay (2018), Sarkar et al. (2019) and Tu et al. (2017) (e.g. compare to Thm. 7).

The regularized method also has the intrinsic advantage that it does not require knowledge of the rank or the singular values of the Hankel matrix beforehand. Numerical experiments on real data in Section 4.8 demonstrate the performance and robustness of the regularized method.

The theorem above follows by combining statistical analysis with a more general deterministic result (Thm. 5). We will state this result in terms of a restricted singular value (RSV) condition. While RSV is a common condition in sparse estimation literature, our analysis requires introducing a spectral norm variation of RSV. Given a matrix M spectral RSV over a set S is defined as follows:

$$\|M\|_S = \max_{v\in S, v\neq 0} \|\mathcal{G}(Mv)\|/\|\mathcal{G}(v)\|.$$

Theorem 5. *Consider the problem* (4.6) *in the MISO setting, where* $\boldsymbol{U} \in \mathbb{R}^{T\times(2n-1)p}$. *Let* β *denote the (weighted) impulse response of the true system which has order* R*, i.e.,* $\operatorname{rank}(\mathcal{G}(\beta)) = R$*, and let* $y = \boldsymbol{U}\beta + \xi$ *be the measured output, where* ξ *is the measurement noise. Finally, denote the minimizer of* (4.6) *by* $\hat{\beta}$. *Define*

$$\mathcal{J}(\beta) := \left\{ v \mid \langle v, \partial(\frac{1}{2}\|\boldsymbol{U}\beta - y\|_2^2 + \lambda\|\mathcal{G}(\beta)\|_*)\rangle \leq 0 \right\}, \quad \Gamma := \|I - \boldsymbol{U}^\top\boldsymbol{U}\|_{\mathcal{J}(\beta)},$$

where $\mathcal{J}(\beta)$ *is the normal cone at* β*, and* Γ *is the spectral RSV. If* $\Gamma < 1$*,* $\hat{\beta}$ *satisfies*

$$\|\mathcal{G}(\hat{\beta} - \beta)\| \leq \frac{\|\mathcal{G}(\boldsymbol{U}^\top\xi)\| + \lambda}{1-\Gamma}.$$

This theorem determines the generic conditions on the measurements $\boldsymbol{U}$ to ensure successful system identification. As future work, it would be desirable to extend our results to a wider range of measurement models.

4.5 IID inputs and the importance of input shape

I.i.d. input (without shaping matrix K) is typically used for the recovery of impulse response, and Oymak & Ozay (2018) proves an optimal bound in terms of Frobenius norm error of least squares algorithm. However, we used a scaling matrix K in our algorithm. We ask that, when a system is low order, is i.i.d. input without shaping optimal regarding the sample complexity?

In the following theorem, we prove that, for a special case, the Gaussian width with unweighted input is polynomial in n, compared to $O(\log n)$ in weighted setting. Since the Gaussian width bound is tight with respect to sample complexity for high probability recovery (McCoy & Tropp, 2013, Thm. 1), Thm. 6 indicates that the sample complexity in i.i.d. input regime is larger than weighted input regime.

Theorem 6. *Suppose the system impulse response is h such that $h_t = 1$, $\forall t \geq 1$, which is order 1. The Gaussian width of the set $\{x \mid \|\mathcal{H}(h+x)\|_* \leq \|\mathcal{H}(h)\|_*\} \cap \mathbb{S}$ is lower bounded by $Cn^{1/6}$ for some constant C.*

Thus in the noiseless setting, the sample complexity is $T \gtrsim n^{1/6}$, which is bigger than $\log n$ dependence with input shape. This result is rather counter-intuitive since i.i.d. inputs are often optimal for structured parameter estimation tasks (e.g. compressed sensing). Our result shows the provable benefit of input shaping.

4.6 Refining the bounds on least-squares estimator

In this section, we revisit the least-squares estimator given measurements $y = \bar{\boldsymbol{U}}h + z$. We consider the MIMO setup where $y \in \mathbb{R}^{T\times m}$ and $h \in \mathbb{R}^{(2n-1)p\times m}$. This is obtained by setting $\lambda = 0$ in (HNN) hence the estimator is given via the pseudo-inverse

$$\hat{h} := h + \bar{\boldsymbol{U}}^\dagger z = \min_{h'} \frac{1}{2}\|\bar{\boldsymbol{U}}h' - y\|_F^2. \tag{4.8}$$

The next theorem bounds the error when inputs and noise are randomly generated.

Theorem 7. *Denote the solution to (4.8) as $\hat{h}$. Let $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$ be input matrix obtained from multiple rollouts, with i.i.d. standard normal entries, $y \in \mathbb{R}^{T\times m}$ be the corresponding* *outputs and $z \in \mathbb{R}^{T\times m}$ be the noise matrix with i.i.d. $\mathcal{N}(0,\sigma_z^2)$ entries. Then the spectral norm error obeys $\|\mathcal{H}(\hat{h}-h)\| \lesssim \sigma_z\sqrt{\frac{mnp}{T}}\log(np)$.*

This theorem improves the spectral norm bound compared to Oymak & Ozay (2018) which naively bounds the spectral norm in terms of IR error using the right-hand side of (4.3). Instead, we show that spectral error is same as the IR error up to a log factor (when there is only output noise). We remark that $O(\sigma_z\sqrt{np/T})$ is a tight lower bound for $\|\mathcal{H}(h-\hat{h})\|$ as well as $\|h-\hat{h}\|$ (Oymak & Ozay, 2018; Arias-Castro et al., 2012). The proof of the theorem above is in Sec. C.5.

4.7 Model selection for regularized system identification

In Thm. 4, we established the recovery error of system impulse response for a particular parameter choice λ, which depends on the noise level. In practice, we do not know the noise level, thus, given the candidates of regularization parameter, which is denoted by a set Λ containing positive numbers, we try a list of $\lambda \in \Lambda$ and check the validation error to perform a model selection. Denote the cardinality of Λ by N_λ. In Algorithm 2, we state our training and validation procedure. The theorem below states our performance guarantee for this algorithm.

Theorem 8. *Consider the setting of Thm. 4. Sample T i.i.d. training rollouts $(\boldsymbol{U}, y)$ and T_{val} i.i.d. validation rollouts $(\boldsymbol{U}_{val}, y_{val})$. Set $\lambda^* = C\sigma\sqrt{\frac{pn}{T}}\log(n)$ which is the choice in Thm. 4. Fix failure probability $P \in (0,1)$. Suppose that:*

(a) *There is a candidate $\hat{\lambda} \in \Lambda$ obeying $\lambda^*/2 \le \hat{\lambda} \le 2\lambda^*$.*

(b) *Validation set obeys $T_{val} \gtrsim \left(\frac{T\log^2(|\Lambda|/P)}{R\log^2(n)}\right)^{1/3}$.*

Set $\bar{R} = \min(R^2, n)$. With probability at least $1-P$, Algorithm 2 achieves an estimation error equivalent to (4.7):

$$\|\mathcal{H}(\hat{h}-h)\| \lesssim \begin{cases} \sqrt{\frac{np}{\boldsymbol{snr}\times T}}\log(n), & \text{if } T \gtrsim \bar{R}, \\ \sqrt{\frac{Rnp}{\boldsymbol{snr}\times T}}\log(n), & \text{if } R \lesssim T \lesssim \bar{R}. \end{cases} \tag{4.9}$$

Algorithm 2 System identification and model selection

```
Require: Training data:  Input feature matrix $\bar{\boldsymbol{U}}$, output vector $y$, with $T$ measurements
  Validation data:  Input feature matrix $\bar{\boldsymbol{U}}_{\text{val}}$, output vector $y_{\text{val}}$, with $T_{\text{val}}$ measurements
  Hyperparameters: Hankel dimension $n$, candidate set $\Lambda$.
  for $\lambda_i \in \Lambda$ do
      Training: Solve $\hat{h}_\lambda \leftarrow \arg\min_{h'} \ \frac{1}{2}\|\bar{\boldsymbol{U}}h' - y\|_2^2 + \lambda_i\|\mathcal{H}(h')\|_*$. Record $\hat{h}_\lambda$.
  Model selection: Choose $\hat{h}_\lambda$ that corresponds to the smallest validation error $\|\bar{\boldsymbol{U}}_{\text{val}}\hat{h}_\lambda - y_{\text{val}}\|_2^2$,
  and call it $\hat{h}$.
  return $\hat{h}$
```

In Algorithm 2, we fix the size of Hankel matrix (which is usually large/overparameterized) and tune λ. In contract to this, Sarkar et al. (2019) introduces a model selection method for unregularized least squares, which is accomplished by changing the size of the Hankel matrix. In the next section, we will run experiments and contrast these two methods, and provide insights on how regularization can improve over least-squares for certain class of dynamical systems.

Sample complexity analysis: model selection with data being requested online. Algorithm 2 uses static data for training and validation, which means that, the total $T + T_{\text{val}}$ samples are given and fixed, and we split the data and run Algorithm 2. We denote the total sample complexity $T_{\text{tot}} = T + T_{\text{val}}$. To be fully efficient in sample complexity, we can start from $T_{\text{tot}} = 0$, keep requesting new samples, which means increasing T_{tot}, and run Algorithm 2 for each T_{tot}. When the validation error is small enough (which happens when $T_{\text{tot}} \gtrsim R$), we know the algorithm recovers a meaningful impulse response estimation and we can **terminate** the algorithm.

We compare it with the model selection algorithm in Sarkar et al. (2019) for least squares estimator, and we find that it does **not terminate** until $T_{\text{tot}} \gtrsim n$. For least squares, the parameter to be tuned is the dimension of the variable, i.e., we vary the length of estimated impulse response. We call the tuning variable n_{t} and it is upper bounded by n. We keep increasing T_{tot} and train by varying $n_{\text{t}} \in [1, T_{\text{tot}}/2]$ (so the least squares problems are overdetermined). The output y is collected at time $2n_{\text{t}} - 1$. We consider two impulse responses with horizon n: $h^1 = \mathbf{1}_n$ (order = 1) and $h^{n_1} = [\mathbf{1}_{n_1}; \mathbf{0}_{n-n_1}]$ (order = n_1). As long as $T_{\text{tot}} \ll n_1$, one cannot differentiate h^1 and h^{n_1} since y is collected at time T_{tot} and the $T_{\text{tot}} + 1$-th to the n-th terms of h^1, h^{n_1} do not contribute to y. Even if the system is order 1, one does not know it and cannot terminate the algorithm. Thus the tuning algorithm in Sarkar et al. (2019) requires $T_{\text{tot}} \gtrsim n$. This does not happen with regularization, because we collect y at time n in Algorithm 2, but not time $2n_{\text{t}} - 1$, thus the algorithm always detects the difference between h^1, h^{n_1} after time n_1.

4.8 Experiments and insights

4.8.1 Experiments with synthetic data

We use an experiment with synthetic data to answer the following question.

When does regularization beat least-squares: Low-order slow-decay systems (Fig. 4.3). So far, we showed that for fixed Hankel size, nuclear norm regularization requires less data than unregularized least-squares especially when the Hankel size is set to be large. However, for least squares, one can choose to use a *smaller Hankel size* that $n \approx R$, so that we solve a problem of small dimension compared to $n \gg R$. We ask if there is a scenario in which fine-tuned nuclear norm regularization strictly outperforms fine-tuned least-squares.

In what follows, we discuss a single trajectory scenario.An advantage of the regularized algorithm is that, we can set up the problem with **large** n, when the sample complexity T and the system order R are both small. Least squares suffers an error of order $(1 - \rho(A)^n)^{-1}$ Oymak & Ozay (2018). The error comes from FIR truncation of impulse response so that happens for both regularized and unregularized algorithms. Thus, if system decays slowly, i.e., $\rho(A) \approx 1$, we will suffer from significant truncation error. As an example, when sample size is 40 and $\rho(A) = 0.98$, if the problem is kept overdetermined (i.e. $n < 40$), then n will not be large enough to make the truncation error $(1 - \rho(A)^n)^{-1}$ small. In regularized method, as long as n is large, we can recover a system with slowly-decaying impulse response even if the number of parameters (i.e., Hankel size) is larger than sample size. This motivates us to compare the performance on recovering systems with *low-order slow-decay*. In Fig. 4.3, we set up an order-1 system with a pole at 0.98 and generate single rollout data with size 40. We tune λ when applying regularized algorithm with $n = 45$ (as it is safe to choose a large Hankel dimension), whereas in unregularized method, Hankel size cannot be larger than 20 ($n \times n$ Hankel has $2n - 1$ parameters and we need least squares to remain overdetermined). With these in mind, in the first two figures we can see that the best validation error of regularization algorithm is 0.44, which is drastically smaller than the unregularized validation error 0.73. In the third figure, we use regularized least squares with $n = 20$, which also causes large truncation error (due to large $\rho(A)$) compared to the initial choice of $n = 45$ (the first figure). In this case, the best validation error is 0.56 which is again noticeably worse than the error 0.44 in the first figure.

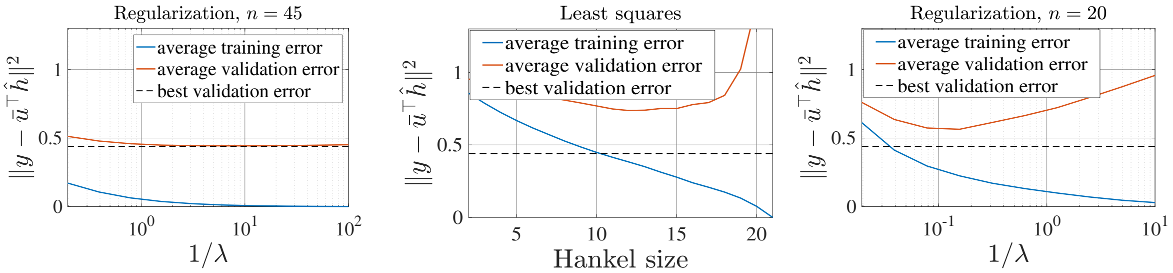
When the number of variables $2n - 1$ is larger than T, the problem is overparameterized and there can be infinitely many impulse responses that achieves zero squared loss on training dataset. This happens in the first figure when $\lambda \to 0$, and in the second figure when n is large. In this case, regularized algorithm chooses the solution with the smallest Hankel nuclear norm and the least squares chooses the one with smallest ℓ_2 norm. We can see that, the first figure has smaller validation error when $1/\lambda$ tends to infinity. So among the solutions that overfits the training dataset, the one with small Hankel nuclear norm has better generalization performance when the true system is low order[1].

4.8.2 Experiments with DaISy Dataset

Our experiment uses the DaISy dataset De Moor et al. (1997), where a known input signal (not random) is applied and the resulting noisy output trajectory is measured. Using the

[1] Codes for generating figures are available at `https://github.com/sunyue93/sunyue93.github.io/blob/main/sysIdFiles.zip`.

Figure 4.3: Synthetic data, single rollout. Order-1 system with pole= 0.98. Recovery by *regularized* algorithm varying λ and *unregularized* algorithms varying Hankel size n. Training sample size is 40 and validation sample size is 800. The figures are the training/validation error with (1) regularized algorithm, different λ and fixed $n = 45$; and (2) least squares with varying n; (3) regularized algorithm, different λ and fixed $n = 20$ (small size).



input and output matrices

$$\boldsymbol{U} = \begin{bmatrix} u_{2n-1}^T & u_{2n-2}^T & \ldots & u_1^T \\ u_{2n}^T & u_{2n-1}^T & \ldots & u_2^T \\ \ldots & & & \\ u_{2n+T-2}^T & u_{2n+T-3}^T & \ldots & u_T^T \end{bmatrix}, \tag{4.10}$$

$$y = [y_{2n-1}, \ldots, y_{2n+T-2}], \tag{4.11}$$

we solve the optimization problem (HNN) using single trajectory data.

While the input model is single instead of multiple rollout, experiments will demonstrate the advantage of Hankel-regularization over least-squares in terms of sample complexity, singular value gap and ease of tuning.

Large data regime: Both Hankel and least-squares algorithms work well (Fig. 4.4). The first two figures in Fig. 4.4 show the training and validation errors of Hankel-regularized and unregularized methods with hyperparameters λ and n respectively. We then choose the best system by tuning the hyperparameters to achieve the smallest validation error. The third figure in Fig. 4.4 plots the training and validation output sequence of the dataset for these algorithms. We see that with sufficient sample size, the system is recovered well. However, the validation error is more flat as a function of $1/\lambda$ (first figure) whereas it is sensitive to the choice of n (second figure), thus λ is easier to tune compared to n.

Small data regime: Hankel-regularization succeeds while least-squares may fail due to overfitting (Fig. 4.5). The first two figures in Fig. 4.5 show that the Hankel spectrums of the two algorithms have a notable difference: The system recovered by Hankel-regularization is low-order and has larger singular value gap. The last two figures in Fig. 4.5 show the advantage of regularization with much better validation performance. As expected from our theory, the difference is most visible in small sample size (this experiment uses 50 training samples). When the number of observations T is small, Hankel-regularization still returns a solution close to the true system while least-squares cannot recover the system properly.

Learning a linear approximation of a nonlinear system with few data (Fig. 4.6). Finally, we show that Hankel-regularization can identify a stable nonlinear system via its linearized approximation as well. We consider the inverted pendulum as the experimental environment. First we use a linearized controller to stabilize the system around the equilibrium, and apply single rollout input to the closed-loop system, which is i.i.d. random input of dimension 1. The dimension of the state is 4, and we observe the output of dimension 1, which is the displacement of the system. We then use the Hankel-regularization and least-squares to estimate the closed-loop system with a linear system model and predict the trajectory using the estimated impulse response. We use $T = 16$ observations for training, and set the dimension to $n = 45$. Fig. 4.6 shows the singular values and estimated trajectory of these two methods. Despite the nonlinearity of the ground-truth system, the regularized algorithm finds a linear model with order 6 and the predicted output has small error, while the correct order is not visible in the singular value spectrum of the unregularized least-squares.

4.9 Conclusion and future directions

This work established new sample complexity and estimation error bounds for system identification. We showed that nuclear norm penalization works well with small sample size regardless of the mis-specification in the problem (i.e. fitting impulse response with a much larger length rather than the true order). For least-squares we provide the first guarantee that is optimal in sample complexity and the Hankel spectral norm error. These results can be refined in several directions. In the proof of Thm. 5, we use a weighted version of the Hankel operator. We expect that directly computing the Gaussian width of the original Hankel operator will also lead to improvements from least square. It would also be interesting to extend the results to account for single trajectory analysis or process noise. In both cases, an accurate analysis of the regularized problem would lead to new algorithmic insights.

Figure 4.4: System identification for CD player arm data. Training data size $= 200$ and validation data size $= 600$. The first two figures are the training/validation errors of varying λ in regularized algorithm ($n = 10$), and training/validation errors of varying Hankel size n in unregularized algorithm. The last figure is the output trajectory of the true system and the recovered systems (best validation chosen for each).

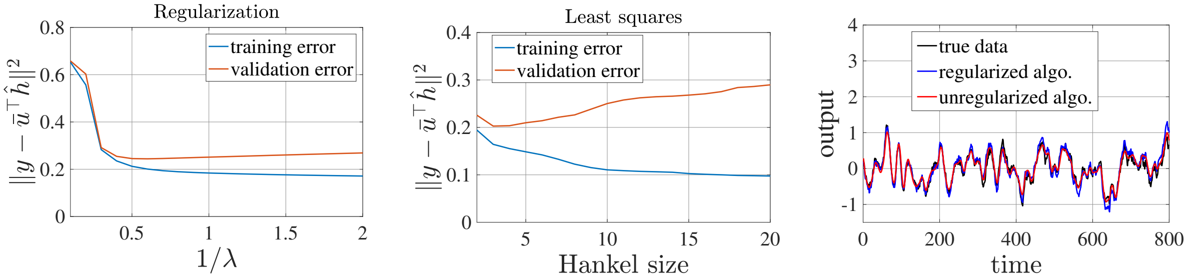


Figure 4.5: The first two figures: CD player arm data, singular values of the *regularized* and *unregularized* Hankel. The last two figures: Recovery by *regularized* and *unregularized* algorithms when Hankel matrix is 10 × 10. Training size is 50 and validation size is 400.

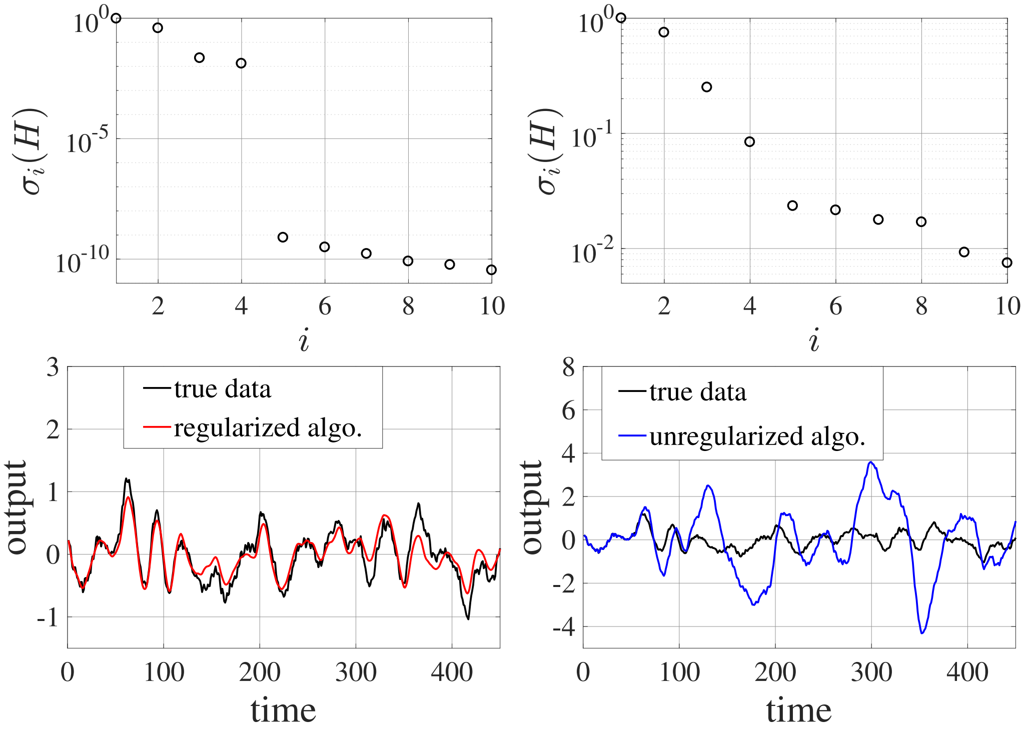
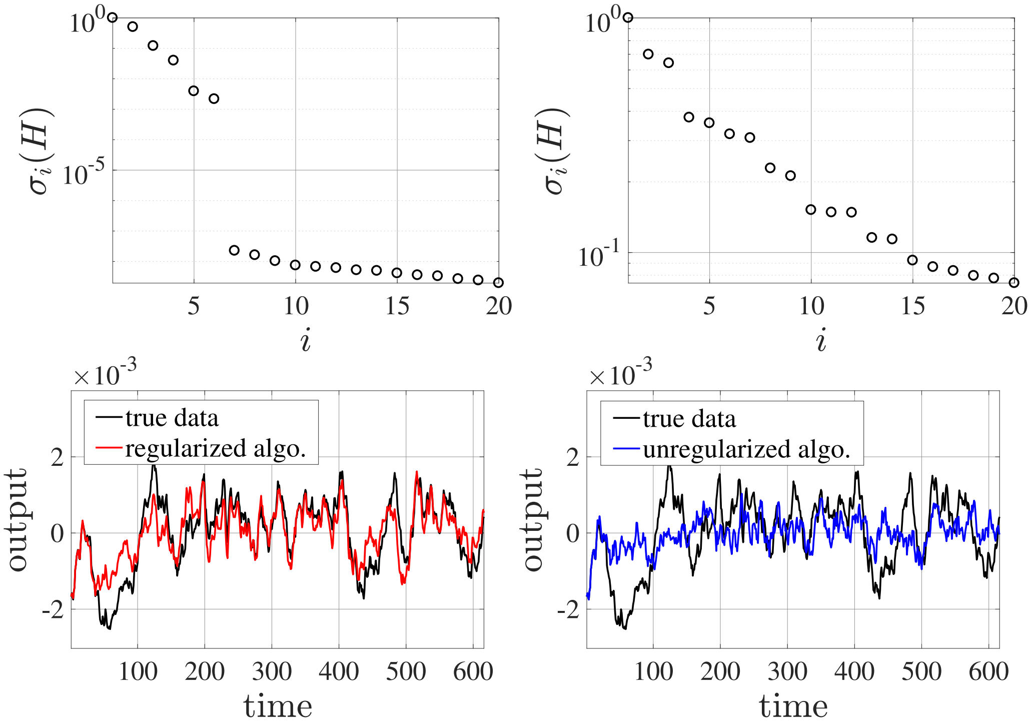


Figure 4.6: The first two figures: Stabilized inverted pendulum data, singular values of the *regularized* and *unregularized* Hankel. The last two figures: Recovery by *regularized* and *unregularized* algorithms when Hankel matrix is 40 × 40. Training size is 16 and validation size is 600.



Chapter 5

ANALYSIS OF OVERPARAMETERIZED LINEAR META-LEARNING

In this chapter, we study the overparameterized linear meta-learning. Here we have a sequence of linear-regression tasks and ask: (1) Given earlier tasks, what is the optimal linear representation of features for a new downstream task? and (2) How many samples do we need to build this representation? Specifically, for (1), we first show that learning the optimal representation coincides with the problem of designing a task-aware regularization to promote inductive bias. This inductive bias explains how the downstream task actually benefits from overparameterization, in contrast to prior works on few-shot learning. For (2), we develop a theory to explain how feature covariance can implicitly help reduce the sample complexity well below the degrees of freedom and lead to small estimation error. We then integrate these findings to obtain an overall performance guarantee for our meta-learning algorithm. Numerical experiments on real and synthetic data verify our insights on overparameterized meta-learning.

This work is published [1] as Sun et al. (2021).

5.1 Introduction

In a multitude of machine learning (ML) tasks with limited data, it is crucial to build accurate models in a sample-efficient way. Constructing a simple yet informative representation of features is a critical component of learning a model that generalizes well to an unseen test set. The field of meta-learning dates back to Caruana (1997); Baxter (2000) and addresses this challenge by transferring insights across distinct but related tasks. Usually, the meta-learner

[1] Codes for generating figures are available at `https://github.com/sunyue93/Rep-Learning/tree/main/nips_supp`.

first (1) learns a feature-representation from previously seen tasks and then (2) uses this representation to succeed at an unseen task. The first phase is called representation learning and the second is called few-shot learning. Such information transfer between tasks is the backbone of modern transfer and multitask learning and finds ubiquitous applications in image classification (Deng et al., 2009), machine translation (Bojar et al., 2014) and reinforcement learning (Finn et al., 2017).

Recent literature in ML theory has posited that overparameterization can be beneficial to generalization in traditional single-task setups for both regression (Mei & Montanari, 2019; Wu & Xu, 2020; Bartlett et al., 2020; Muthukumar et al., 2019; Montanari et al., 2019) and classification (Muthukumar et al., 2020; Montanari et al., 2020) problems. Empirical literature in deep learning suggests that overparameterization is of interest for both phases of meta-learning as well. Deep networks are stellar representation learners despite containing many more parameters than the sample size. Additionally, overparameterization is observed to be beneficial in the few-shot phase for transfer-learning in Figure 5.1(a). A ResNet-50 network pretrained on Imagenet was utilized to obtain a representation of R features for classification on CIFAR-10. All layers except the final (softmax) layer are frozen and are treated as a fixed feature-map. We then train the final layer of the network for the downstream task which yields a linear classifier on pretrained features. The figure plots the effect of increasing R on the test error on CIFAR-10, for different choices of training size n_2. For each choice of n_2, increasing representation dimension R beyond downstream samples n_2 is seen to reduce the test-error. These findings are corroborated by Finn et al. (2017) (MAML) and Vinyals et al. (2016), who successfully use a transfer learning method that adapts a pre-trained model, with 112980 parameters, to downstream tasks with only 1-5 new training samples.

In Figure 5.1(b), we consider a sequence of *linear* regression tasks and plot the few-shot error of our proposed projection and eigen-weighting based meta-learning algorithm for a fixed few-shot training size, but varying dimensionality of features. The resulting curve looks similar to Figure 5.1(a) and suggests that the observations regarding overparameterization

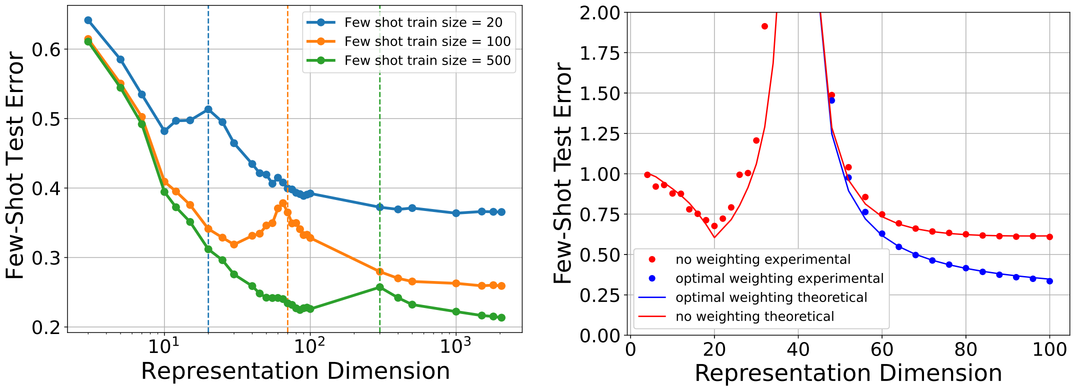


Figure 5.1: **Illustration of the benefit of overparameterization in the few-shot phase.** (a) Double-descent in transfer learning: dashed lines indicate the location where the number of features R exceed the number of training points; i.e., the transition from under to over-parameterization. The experimental details are contained in the supplement. (b) Illustration of the benefit of using Weighted minL2-interpolation in Definition 3 (blue). See Remark 4 for details and discussion.

for meta-learning in neural networks can, to a good extent, be captured by linear models, thus motivating their detailed study. This aligns with trends in recent literature: while deep nets are nonlinear, recent advances show that linearized problems such as kernel regression (e.g., via neural tangent kernel (Jacot et al., 2018; Du et al., 2018; Lee et al., 2019; Oymak et al., 2019; Chizat et al., 2018)) provide a good proxy to understand some of the theoretical properties of practical overparameterized deep nets.

However, existing analysis of subspace-based meta-learning algorithms for both the representation learning and few-shot phases of linear models have typically focused on the classical *underparameterized regime*. These works (see Paragraphs 2-3 of Section 5.1.2) consider the case where representation learning involves projection onto a lower-dimensional subspace. On the other hand, recent works on double descent show that an *overparameterized* interpolator beats PCA-based method. We aim to build upon these results to develop a theoretical understanding of overparameterized meta-learning.

5.1.1 Our contributions

We study meta-learning when each task is a linear regression problem, similar in spirit to Tripuraneni et al. (2020); Kong et al. (2020b). In the representation learning phase, the learner is provided with training data from T distinct tasks, with n_1 training samples per task: using this data, it selects a matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d\times R}$ with arbitrary R to obtain a linear *representation* of features via the map $\boldsymbol{x} \to \boldsymbol{\Lambda}^\top \boldsymbol{x}$. In the few-shot learning phase, the learner faces a new task with n_2 training samples and aims to use the representation $\boldsymbol{\Lambda}^\top \boldsymbol{x}$ to aid prediction performance.

We highlight that obtaining the representation consists of two steps: first the learner projects $\boldsymbol{x}$ onto R basis directions, and then performs *eigen-weighting* of each of these directions, as shown in Figure 5.2b. The overarching goal of this chapter is to propose a scheme to use the knowledge gained from earlier tasks to choose $\boldsymbol{\Lambda}$ that minimizes few-shot risk. This goal enables us to engage with important questions regarding overparameterization:

Q1: What should the size R and the representation $\boldsymbol{\Lambda}$ be to minimize risk at the few-shot phase?

Q2: Can we learn the Rd dimensional representation $\boldsymbol{\Lambda}$ with $N \ll Rd$ samples?

The answers to the questions above will shed light on whether overparameterization is beneficial in few-shot learning and representation learning respectively. Towards this goal, we make several contributions to the finite-sample understanding of *linear* meta-learning, under assumptions discussed in Section 5.2. Our results are obtained for a general data/task model with *arbitrary task covariance* $\boldsymbol{\Sigma}_\beta$ *and feature covariance* $\boldsymbol{\Sigma}_F$ which allows for a rich set of observations.

Optimal representation for few-shot learning. As a stepping stone towards the goal of characterizing few-shot risk for different $\boldsymbol{\Lambda}$, in Section 5.3 we first consider learning with **known covariances** $\boldsymbol{\Sigma}_T$ and $\boldsymbol{\Sigma}_F$ respectively (Algorithm 3). Compared to projection-only representations in previous works (see Paragraphs 2-3 of Section 5.1.2), our scheme applies

$\boldsymbol{\Sigma}_F$	Feature covariance
$\boldsymbol{\Sigma}_T$	Task covariance
$\tilde{\boldsymbol{\Sigma}}_T$	Canonical task covariance
n_1	Samples per each earlier task
T	Number of earlier tasks
N	Total sample size $T \times n_1$
n_2	Samples for new task
$\boldsymbol{\Lambda}$	Eigen-weighting matrix

Table 5.1: Main notation

eigen-weighting matrix $\boldsymbol{\Lambda}^*$ to incentivize the optimizer to place higher weight on promising eigen-directions. This eigen-weighting procedure has been shown in the single-task case to be extremely crucial to avail the benefit of overparameterization (Belkin et al., 2019; Montanari et al., 2019; Muthukumar et al., 2019): it captures an inductive bias that promotes certain features and demotes others. We show that the importance of eigen-weighting extends to the multi-task case as well.

Canonical task covariance. Our analysis in Section 5.3 also reveals that, the optimal subspace and representation matrix are closed-form functions of the *canonical task covariance* $\tilde{\boldsymbol{\Sigma}}_T = \boldsymbol{\Sigma}_F^{1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F^{1/2}$, which captures the feature saliency by summarizing the feature and task distributions.

Representation learning. In practice, task and feature covariances (and hence the canonical covariance) are rarely known apriori. However, we can estimate the principal subspace of the canonical task covariance $\tilde{\boldsymbol{\Sigma}}_T$ (which has a degree of freedom (DoF) of $\Omega(Rd)$) from data. In Section 5.4 we first present empirical evidence that feature covariance $\boldsymbol{\Sigma}_F$ is "positively correlated" with $\tilde{\boldsymbol{\Sigma}}_T$. Then we propose an efficient algorithm based on Method-of-Moments (MoM), and show that the sample complexity of representation learning is well below $\mathcal{O}(Rd)$ due to the inductive bias. Our sample complexity bound depends on interpretable quantities such as *effective ranks* $\boldsymbol{\Sigma}_F, \tilde{\boldsymbol{\Sigma}}_T$ and improves over prior art (e.g., Kong et al. (2020b); Tripuraneni et al. (2020)), even though the prior works were specialized to low-rank $\tilde{\boldsymbol{\Sigma}}_T$ and identity $\boldsymbol{\Sigma}_F$ (see Table 5.2).

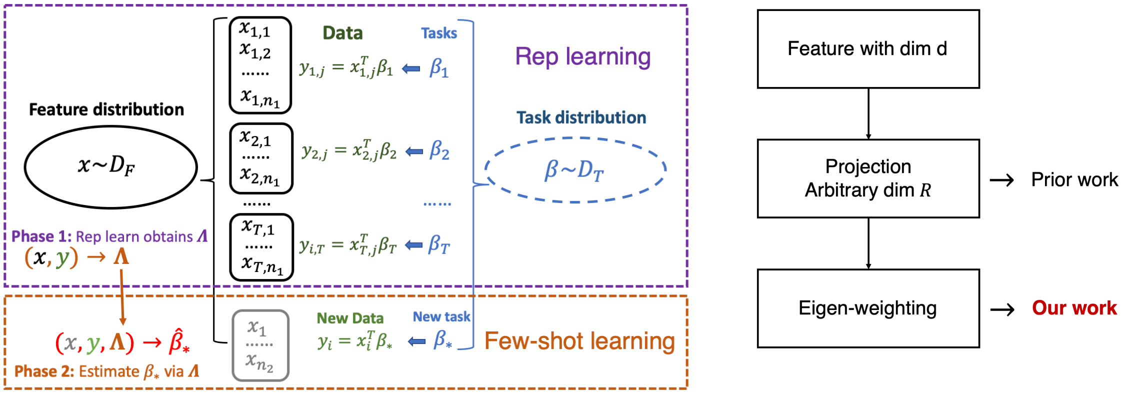

Data
Tasks
Rep learning
Feature distribution
$x \sim D_F$
Task distribution
$\beta \sim D_T$
$x_{1,1}$ $x_{1,2}$ …… x_{1,n_1}
$y_{1,j} = x_{1,j}^T \beta_1 \Leftarrow \beta_1$
$x_{2,1}$ …… x_{2,n_1}
$y_{2,j} = x_{2,j}^T \beta_2 \Leftarrow \beta_2$
$x_{T,1}$ …… x_{T,n_1}
$y_{i,T} = x_{T,j}^T \beta_T \Leftarrow \beta_T$
Phase 1: Rep learn obtains Λ
$(x, y) \rightarrow \Lambda$
$(x, y, \Lambda) \rightarrow \hat{\beta}_*$
Phase 2: Estimate β_* via Λ
x_1 …… x_{n_2}
New Data
New task
$y_i = x_i^T \beta_* \Leftarrow \beta_*$
Few-shot learning
Feature with dim d
Projection
Arbitrary dim R
→ Prior work
Eigen-weighting
→ Our work

Figure 5.2: (a) Steps of the meta-learning algorithm. (b) Our representation-learning algorithm has two steps: projection and eigen-weighting. We focus on the use of overparameterization+weighting matrix (Def. 3), and compare this with overparameterization with simple projection (no eigen-weighting), and underparameterization (for which eigen-weighting has no impact and is equivalent to projection). Tripuraneni et al. (2020); Kong et al. (2020b,a); Du et al. (2020) study underparameterized projections only. To distinguish from eigen-weighting, we will refer to simple projections as subspace-based representations.

End to end meta-learning guarantee. In Section 5.5, we consider the generalization of Section 5.3, where we have only estimates of the covariances instead of perfect knowledge. This leads to an overall meta-learning guarantee in terms of $\mathbf{\Lambda}^*$, N and n_2 and uncovers a bias-variance tradeoff: As N decreases, it becomes more preferable to use a smaller R (more bias, less variance) due to inaccurate estimate of the weak eigen-directions of $\tilde{\mathbf{\Sigma}}_T$. In other words, we find that overparameterization is only beneficial for few-shot learning if the quality of representation learning is sufficiently good. This explains why, in practice, increasing the representation dimension may not help reduce few-shot risk beyond a certain point (see Fig. A.2).

5.1.2 Related work

Overparameterized ML and double-descent. The phenomenon of double-descent was first discovered by Belkin et al. (2019). This paper and subsequent works on this topic Bartlett et al. (2020); Muthukumar et al. (2019, 2020); Montanari et al. (2019); Chang et al.

(2020) emphasize the importance of the right prior (sometimes referred to as inductive bias or regularization) to avail the benefits of overparameterization. However, an important question that arises is: where does this prior come from? Our work shows that the prior can come from the insights learned from related previously-seen tasks. Section 5.3 extends the ideas in Nakkiran et al. (2020); Wu & Xu (2020) to depict how the optimal representation described can be learned from imperfect covariance estimates as well.

Theory for representation learning. Recent papers (Kong et al., 2020b,a; Tripuraneni et al., 2020; Du et al., 2020) propose the theoretical bounds of representation learning when the tasks lie in an exactly r dimensional subspace. (Kong et al., 2020b,a; Tripuraneni et al., 2020) discuss method of moment estimators and (Tripuraneni et al., 2020; Du et al., 2020) discuss matrix factorized formulations. Tripuraneni et al. (2020) shows that the number of samples that enable meaningful representation learning is $\mathcal{O}(dr^2)$. Kong et al. (2020b,a); Tripuraneni et al. (2020) assume the features follow a standard normal distribution. Thekumparampil et al. (2021) proposes the alternating minimization algorithm for matrix factorization method, and shows that the algorithm converges to the estimator that achieves $\mathcal{O}((dr)^{-1/2})$ error with $\mathcal{O}(dr^2)$ samples. We define a canonical covariance which handles arbitrary feature and task covariances. We also show that our estimator succeeds with $\mathcal{O}(dr)$ samples when $n_1 \sim r$, and extend the bound to general covariances with effective rank defined.

Subspace-based meta learning. With tasks being low rank, Kong et al. (2020b,a); Tripuraneni et al. (2020); Gulluk et al. (2021); Du et al. (2020) do few-shot learning in a low dimensional space. Collins et al. (2022) applies MAML for matrix factorization in subspace-based linear model and shows that MAML learns the features. Yang et al. (2020, 2021) study meta-learning for linear bandits. Lucas et al. (2020) gives information theoretic lower and upper bounds. Bouniot et al. (2020) proposes subspace-based methods for nonlinear problems such as classification. As mentioned in Chapter 4, Chen & Poor (2022) studies learning mixtures of linear systems from multiple trajectories, which is a combination of system identification and learning mixed clusters. The clustering method is similar to the philosophy of clustering method, and then it applies individual system identification algorithm as downstream tasks. We investigate a representation with arbitrary dimension, specifically interested in overparameterized case and show it yields a smaller error with general task/feature covariances. Du et al. (2020) provides results on overparameterized representation learning, but Du et al. (2020) requires number of samples per pre-training task to obey $n_1 \gtrsim d$, whereas our results apply as soon as $n_1 \gtrsim 1$.

Mixed Linear Regression (MLR). In MLR (Zhong et al., 2016; Li & Liang, 2018; Chen et al., 2020), multiple linear regression are executed, similar to representation learning. The difference is that, the tasks are drawn from a finite set, and number of tasks can be larger than d and not necessarily low rank. Lounici et al. (2011); Cavallanti et al. (2010); Maurer et al. (2016) propose sample complexity bounds of representation learning for mixed linear regression. They can be combined with other structures such as binary task vectors (Balcan et al., 2015) and sparse task vectors (Argyriou et al., 2008).

5.2 Problem setup

The problem we consider consists of two phases:

1. Representation learning: Prior tasks are used to learn a suitable representation to process features.
2. Few-shot learning: A new task is learned with a few samples by using the suitable representation.

This section defines the key notations and describes the data generation procedure for the two phases. In summary, we study linear regression tasks, where the features and tasks are generated randomly i.i.d. from their associated distributions $\mathcal{D}_T$ and $\mathcal{D}_F$, and the two phases share the same feature and task distributions.The setup is summarized in Figure 5.2(a).

5.2.1 Data generation

Definition 1 (Task and feature distributions)**.** *Throughout, $\mathcal{D}_T$ and $\mathcal{D}_F$ denote the distributions of tasks β_i and features $\boldsymbol{x}_{ij}$ respectively. These distributions are subGaussian, zero-mean with corresponding covariance matrices $\boldsymbol{\Sigma}_T$ and $\boldsymbol{\Sigma}_F$.*

Definition 2 (Data distribution for a single task)**.** *Given a specific realization of task vector $\beta \sim \mathcal{D}_T$, the corresponding label/input distribution $(y, \boldsymbol{x}) \sim \mathcal{D}_\beta$ is obtained via $y = \boldsymbol{x}^\top \beta + \varepsilon$ where $\boldsymbol{x} \sim \mathcal{D}_F$ and ε is zero-mean subgaussian noise with variance σ^2.*

Data for Representation Learning (Phase 1). We have T tasks, each with n_1 training examples. The task vectors $(\beta_i)_{i=1}^T \subset \mathbb{R}^d$ are drawn i.i.d. from the distribution $\mathcal{D}_T$. The data for ith task is given by $(y_{ij}, \boldsymbol{x}_{ij})_{j=1}^{n_1} \overset{\text{i.i.d.}}{\sim} \mathcal{D}_{\beta_i}$. In total, there are $N = T \times n_1$ examples.

Data for Few-Shot Learning (Phase 2). Sample task $\boldsymbol{\beta}_\star \sim \mathcal{D}_T$. Few-shot dataset has n_2 examples $(y_i, \boldsymbol{x}_i)_{j=1}^{n_2} \overset{\text{i.i.d.}}{\sim} \mathcal{D}_{\boldsymbol{\beta}_\star}$.

We use representation learning data to learn a representation of feature-task distribution, called eigen-weighting matrix $\boldsymbol{\Lambda}$ in Defenition 3 below. The matrix $\boldsymbol{\Lambda}$ is passed to few-shot learning stage, helping learn $\boldsymbol{\beta}_\star$ with few data.

5.2.2 Training in Phase 2

We will define a weighted representation, called eigen-weighting matrix, and show how it is applied for few-shot learning. The matrix is learned during representation learning using the data from the T tasks. Denote $\boldsymbol{X} \in \mathbb{R}^{n_2 \times d}$ whose i^{th} row is $\boldsymbol{x}_i$, and $\boldsymbol{y} = [y_1, ..., y_m]^\top$. We are interested in studying the weighted 2-norm interpolator defined below for overparameterization regime $R \geq n_2$.

Definition 3 (Eigen-weighting matrix and Weighted ℓ_2-norm interpolator)**.** *Let the representation dimension be R, where R is any integer between 1 and d. We define an eigen-weighting matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times R}$ and the associated weighted ℓ_2-norm interpolator*

$$\hat{\beta}_{\boldsymbol{\Lambda}} = \arg\min_{\beta} \|\boldsymbol{\Lambda}^\dagger \beta\|_2 \quad \textit{s.t.} \quad \boldsymbol{y} = \boldsymbol{X}\beta \quad \textit{and} \quad \beta \in \text{range\_space}(\boldsymbol{\Lambda}).$$

The solution is equivalent to defining $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}^\dagger \hat{\beta}_{\boldsymbol{\Lambda}}$ and solving an unweighted minimum 2-norm regression with features $\boldsymbol{X\Lambda}$. This corresponds to our few-shot learning problem

$$\hat{\boldsymbol{\alpha}}_{\boldsymbol{\Lambda}} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_2 \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{X\Lambda\alpha}$$

from which we obtain $\hat{\beta}_{\mathbf{\Lambda}} = \mathbf{\Lambda}\hat{\boldsymbol{\alpha}}_{\mathbf{\Lambda}}$. When there is no confusion, we can replace $\hat{\beta}_{\mathbf{\Lambda}}$ with $\hat{\beta}$. One can easily see that $\hat{\beta} = \mathbf{\Lambda}(\boldsymbol{X}\mathbf{\Lambda})^{\dagger}\boldsymbol{y}$. We note that Definition 3 is a special case of the weighted ridge regression discussed in Wu & Xu (2020), as stated in Observation 1. An alternative equivalence between min-norm interpolation and ridge regression can be found in Muthukumar et al. (2019).

Observation 1. Let $\boldsymbol{X} \in \mathbb{R}^{n_2 \times d}$ and $\boldsymbol{y} \in \mathbb{R}^{n_2}$, define

$$\hat{\beta}_1 = \lim_{t \to 0} \text{argmin}_{\beta} \|\boldsymbol{X}\beta - \boldsymbol{y}\|_2^2 + t\beta^{\top}(\mathbf{\Lambda}\mathbf{\Lambda}^{\top})^{\dagger}\beta, \ \beta \in \text{column space of } \mathbf{\Lambda}. \tag{5.1}$$

We have that $\hat{\beta}_1 = \hat{\beta}$.

5.3 Canonical covariance and optimal representation

In this section, we ask the simpler question: if the covariances $\mathbf{\Sigma}_T$ and $\mathbf{\Sigma}_F$ are known, what is the best choice of $\mathbf{\Lambda}$ to minimize the risk of the interpolator from Definition 3? In general, the covariances are not known; however, the insights from this section help us study the more general case in Section 5.5. Define the risk as the expected error of inferring the label on the few-shot dataset,

$$\text{risk}(\mathbf{\Lambda}, \mathbf{\Sigma}_T, \mathbf{\Sigma}_F) = \boldsymbol{E}_{\boldsymbol{x},y,\beta}(y - \boldsymbol{x}^{\top}\hat{\beta}_{\mathbf{\Lambda}})^2 = \boldsymbol{E}_{\beta}(\hat{\beta}_{\mathbf{\Lambda}} - \beta)^{\top}\mathbf{\Sigma}_F(\hat{\beta}_{\mathbf{\Lambda}} - \beta) + \sigma^2. \tag{5.2}$$

The natural choice of optimization for choosing $\mathbf{\Lambda}$ would be to choose the weighting that minimizes the eventual risk of the learned interpolator.

$$\mathbf{\Lambda}^* = \arg \min_{\mathbf{\Lambda}' \in \mathbb{R}^{d \times R}} \text{risk}(\mathbf{\Lambda}', \mathbf{\Sigma}_T, \mathbf{\Sigma}_F) \tag{5.3}$$

Since the label y is bilinear in x and β, we introduce whitened features $\tilde{\boldsymbol{x}} = \mathbf{\Sigma}_F^{-1/2}\boldsymbol{x}$ and associated task vector $\tilde{\beta} = \mathbf{\Sigma}_F^{1/2}\beta$. This change of variables ensures $\boldsymbol{x}^T\beta = \tilde{\boldsymbol{x}}^T\tilde{\beta}$; now, the task covariance in the transformed coordinates takes the form

$$\tilde{\mathbf{\Sigma}}_T = \mathbf{\Sigma}_F^{1/2}\mathbf{\Sigma}_T\mathbf{\Sigma}_F^{1/2},$$

which we call the **canonical task covariance**; it captures the joint behavior of feature and task covariances $\mathbf{\Sigma}_F, \mathbf{\Sigma}_T$. Below, we observe that the risk in Equation (5.2) is invariant to the change of co-ordinates that we have described above i.e it does not change when $\mathbf{\Sigma}_F^{1/2}\mathbf{\Sigma}_T\mathbf{\Sigma}_F^{1/2}$ is fixed and we vary $\mathbf{\Sigma}_F$ and $\mathbf{\Sigma}_T$.

Observation 2 (Equivalence to problem with whitened features)**.** Let data be generated as in Phase 1. Denote $\tilde{\mathbf{\Sigma}}_T = \mathbf{\Sigma}_F^{1/2}\mathbf{\Sigma}_T\mathbf{\Sigma}_F^{1/2}$. Then $\mathrm{risk}(\mathbf{\Sigma}_F^{-1/2}\mathbf{\Lambda}, \mathbf{\Sigma}_T, \mathbf{\Sigma}_F) = \mathrm{risk}(\mathbf{\Lambda}, \tilde{\mathbf{\Sigma}}_T, \boldsymbol{I})$.

This observation can be easily verified by substituting the change-of-coordinates into (5.2) and evaluating the risk.

The risk in (5.2) quantifies the quality of representation $\mathbf{\Lambda}$; however it is not a manageable function of $\mathbf{\Lambda}$ that can be straightforwardly optimized. In this subsection, we show that it is asymptotically equivalent to a different optimization problem, which can be easily solved by analyzing KKT optimality conditions. Theorem 9 characterizes this equivalence; the computeReduction subroutine of Algorithm 3 calculates key quantities that are used in specifying the reduction, and the computeOptimalRep subroutine of Algorithm 3 uses the solution of the simpler problem to obtain a solution for the original.

Assumption 9 (Bounded feature covariance)**.** *There exist positive constants* $\Sigma_{\min}, \Sigma_{\max}$ *such that* $\mathbf{\Sigma}_F$ *is lower/upper bounded as follows:* $\mathbf{0} \prec \Sigma_{\min}\boldsymbol{I} \preceq \mathbf{\Sigma}_F \preceq \Sigma_{\max}\boldsymbol{I}$.

Assumption 10 (Joint diagonalizability)**.** $\mathbf{\Sigma}_F$ *and* $\mathbf{\Sigma}_T$ *are diagonal matrices.*[2]

Assumption 11 (Double asymptotic regime)**.** *We let the dimensions and the sample size grow as* $d, R, n_2 \to \infty$ *at fixed ratios* $\bar{\kappa} := d/n_2$ *and* $\kappa := R/n_2$.

Assumption 12. *The joint empirical distribution of the eigenvalues of* $\mathbf{\Lambda}_R$ *and* $\tilde{\mathbf{\Sigma}}_T^R$ *is given by the average of Dirac* δ*'s:* $\frac{1}{R}\sum_{i=1}^{R}\delta_{\mathbf{\Lambda}_{R,i},\sqrt{R}\tilde{\mathbf{\Sigma}}_{T,i}^R}$. *It converges to a fixed distribution as* $d \to \infty$.

[2] This is equivalent to the more general scenario where $\mathbf{\Sigma}_F$ and $\mathbf{\Sigma}_T$ are jointly diagonalizable.

Algorithm 3 Constructing the optimal representation

```
Require: Projection dimension $R$, noise level $\sigma$, canonical covariance $\tilde{\mathbf{\Sigma}}_T$, task covariance $\mathbf{\Sigma}_F$.
  function ComputeOptimalRep($R, \mathbf{\Sigma}_F, \tilde{\mathbf{\Sigma}}_T, \sigma, n_2$)
      $\boldsymbol{U}_1, \mathbf{\Sigma}_F^R, \tilde{\mathbf{\Sigma}}_T^R, \sigma_R$ = ComputeReduction$R, \mathbf{\Sigma}_F, \tilde{\mathbf{\Sigma}}_T, \sigma$
      Optimization: Get $\boldsymbol{\theta}^*$ from (OPT-REP).
      Map to eigenvalues: Set diagonal $\mathbf{\Lambda}_R^* \in \mathbb{R}^{R\times R}$ with entries $\mathbf{\Lambda}_{R,i}^* = (1/\boldsymbol{\theta}_i^* - 1)^{-2}$.
      Lifting and feature whitening: $\mathbf{\Lambda}^* \leftarrow \boldsymbol{U}_1(\mathbf{\Sigma}_F^R)^{-1/2}\mathbf{\Lambda}_R^*$.
      Return $\mathbf{\Lambda}^*$
  function ComputeReduction($R, \mathbf{\Sigma}_F, \tilde{\mathbf{\Sigma}}_T, \sigma$)
      Get eigen-decomposition $\tilde{\mathbf{\Sigma}}_T = \boldsymbol{U}\mathbf{\Sigma}\boldsymbol{U}^\top$.
      Principal eigenspace $\boldsymbol{U}_1 \in \mathbb{R}^{d\times R}$ = the first $R$ columns of $\boldsymbol{U}$.
      Top eigenvalues: Set $\tilde{\mathbf{\Sigma}}_T^R = \boldsymbol{U}_1^\top \tilde{\mathbf{\Sigma}}_T \boldsymbol{U}_1, \mathbf{\Sigma}_F^R = \boldsymbol{U}_1^\top \mathbf{\Sigma}_F \boldsymbol{U}_1$
      Equivalent noise level: $\sigma_R^2 \leftarrow \sigma^2 + \mathbf{tr}(\tilde{\mathbf{\Sigma}}_T) - \mathbf{tr}(\tilde{\mathbf{\Sigma}}_T^R)$.
      Return $\boldsymbol{U}_1, \mathbf{\Sigma}_F^R, \tilde{\mathbf{\Sigma}}_T^R, \sigma_R$
```

With these assumptions, we can derive an analytical expression to quantify the risk of a representation $\mathbf{\Lambda}$. We will then optimize this analytic expression to obtain a formula for the optimal representation.

Theorem 9 (Asymptotic risk equivalence)**.** *Suppose Assumptions 9, 10, 11, 12 hold. Let $\xi > 0$ be the unique number obeying $n_2 = \sum_{i=1}^R \left(1 + (\xi\mathbf{\Lambda}_i^2)^{-1}\right)^{-1}$. Define $\boldsymbol{\theta} \in \mathbb{R}^R$ with entries $\boldsymbol{\theta}_i = \frac{\xi\mathbf{\Lambda}_i^2}{1+\xi\mathbf{\Lambda}_i^2}$ and calculate $\tilde{\mathbf{\Sigma}}_T^R, \sigma_R$ using the* ComputeReduction *procedure of Algorithm 3. Then, define the analytic risk formula*

$$f(\boldsymbol{\theta}, \tilde{\mathbf{\Sigma}}_T^R, n_2) = \frac{1}{n_2 - \|\boldsymbol{\theta}\|_2^2}\left(n_2 \sum_{i=1}^R (1-\boldsymbol{\theta}_i)^2 \tilde{\mathbf{\Sigma}}_{T,i}^R + (\|\boldsymbol{\theta}\|_2^2 + 1)\sigma_R^2\right). \tag{5.4}$$

We have that

$$\lim_{n_2\to\infty} f(\boldsymbol{\theta}, \tilde{\mathbf{\Sigma}}_T^R, n_2) = \lim_{n_2\to\infty} risk(\mathbf{\Sigma}_F^{-1/2}\mathbf{\Lambda}, \mathbf{\Sigma}_T, \mathbf{\Sigma}_F) \tag{5.5}$$

The proof of Theorem 9 applies the convex Gaussian Min-max Theorem (CGMT) in Thrampoulidis et al. (2015) and can be found in the Appendix D.2. We show that as dimension grows, the distribution of the estimator $\hat{\beta}$ converges to a Gaussian distribution and we can calculate the expectation of risk.

Theorem 9 provides us with a closed-form risk for any linear representation. Now, one can solve for the optimal representation by computing (OPT-REP) below. In order to do this, we propose an algorithm for the optimization problem in Appendix D.2.5 via a study of the KKT conditions for the problem[3].

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \; f(\boldsymbol{\theta}, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F), \text{ s.t. } 0 \le \boldsymbol{\theta} < 1, \sum_{i=1}^{R} \boldsymbol{\theta}_i = n_2 \qquad \text{(OPT-REP)}$$

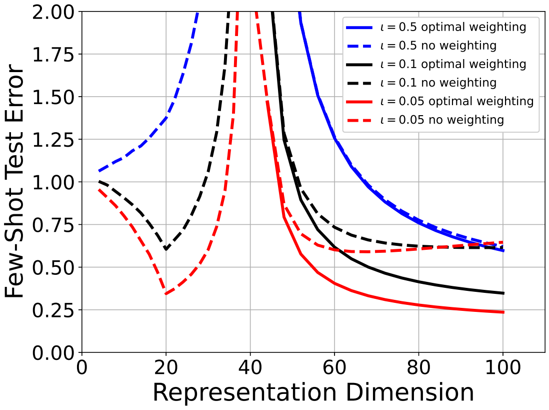


Figure 5.3: Theoretical risk of optimal representation. $\boldsymbol{\Sigma}_F = \boldsymbol{I}_{100}$, $\boldsymbol{\Sigma}_T = \text{diag}(\boldsymbol{I}_{20}, \iota\boldsymbol{I}_{80})$, $n_2 = 40$.

The optimal representation is[4] $\boldsymbol{\Lambda}^*_{R,i} = ((1/\boldsymbol{\theta}^*_i - 1)\xi)^{-2}$. The subroutine computeOptimalRep in Algorithm 3 summarizes this procedure.

Remark 4. *Thm. 9 states that risk*$(\boldsymbol{\Sigma}_F^{-1/2}\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F)$ *can be arbitrarily well-approximated by* $f(\boldsymbol{\theta}, \tilde{\boldsymbol{\Sigma}}_T^R, n_2)$ *if* n_2 *is sufficiently large. In Fig. 5.1(b), we set* $\boldsymbol{\Sigma}_F = \boldsymbol{I}_{100}$, $\boldsymbol{\Sigma}_T =$ *diag*$(\boldsymbol{I}_{20}, 0.1\boldsymbol{I}_{80})$, $n_2 = 40$. *The curves in Fig5.1(b) are the finite dimensional approximation of* f *(LHS of* (5.5)*); the dots are empirical approximations of the risk (RHS of* (5.5)*). We*

[3] In Sec. 5.5 the constraint is $\underline{\theta} \le \boldsymbol{\theta} \le 1 - \frac{d-n_2}{n_2}\underline{\theta}$ for robustness concerns.

[4] In the algorithm, $\xi = 1$ and $\boldsymbol{\Lambda}_{R,i} = (1/\boldsymbol{\theta}^*_i - 1)^{-2}$, because $c\boldsymbol{\Lambda}^*$ for any constant c gives the same $\hat{\beta}$.

tested two cases when $\mathbf{\Lambda}$ is the optimal eigen-weighting or projection matrix with no weighting. Our theorem is corroborated by the observation that the dots and curves are visibly very close. The approximation is already accurate for the finite dimensional problem with just $n_2 = 40$.

The benefit of overparameterization. Theorem 1 leads to an optimal eigen-weighting strategy via asymptotic analysis. In Figure 5.3, we plot the effect on the risk of increasing R for different shapes of task covariance; the parameter ι controls how spiked $\mathbf{\Sigma}_T$ is, with a smaller value for ι indicating increased spiked-ness. For the underparameterized problem, the weighting does not have any impact on the risk. In the overparameterized regime, the eigen-weighted learner achieves lower few-shot error than its unweighted ($\mathbf{\Lambda} = \boldsymbol{I}$) counterpart, showing that eigen-weighting becomes critical.

The eigen-weighting procedure can introduce inductive bias during few-shot learning, and helps explain how optimal representation minimizing the few-shot risk can be overparameterized with $R \gg n_2$. We note that, an R dimensional representation can be recovered by a d dimensional representation matrix of rank R, thus the underparameterized case can never beat d dimensional case in theory. The error with optimal eigen-weighting in overparameterized regime is smaller than the respective underparameterized counterpart. The error is lower with smaller ι. It implies that, while $\tilde{\mathbf{\Sigma}}_T$ gets closer to low-rank, the excess error caused by choosing small dimension R (equal to the gap $\sigma_R^2 - \sigma^2$ in Algo 3) is not as significant.

Low dimensional representations zero out features and cause bias. By contrast, when $\tilde{\mathbf{\Sigma}}_T \in \mathbb{R}^{d\times d}$ is not low rank, every feature contributes to learning with the importance of the features reflected by the weights. This viewpoint is in similar spirit to that of Hastie et al. (2019) where the authors devise a misspecified linear regression to demonstrate the benefits of overparameterization. Our algorithm allows arbitrary representation dimension R and eigen-weighting.

5.4 Representation learning

In this section, we will show how to estimate the useful distribution in representation learning phase that enables us to calculate eigen-weighting matrix $\boldsymbol{\Lambda}^*$. Note that $\boldsymbol{\Lambda}^*$ depends on the canonical covariance $\tilde{\boldsymbol{\Sigma}}_T = \boldsymbol{\Sigma}_F^{1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F^{1/2}$. Learning the R-dimensional principal subspace of $\tilde{\boldsymbol{\Sigma}}_T$ enables us[5] to calculate $\boldsymbol{\Lambda}^*$. Denote this subspace by $\tilde{\boldsymbol{S}}_T$.

Subspace estimation vs. inductive bias. The subspace-based representation $\tilde{\boldsymbol{S}}_T$ has degrees of freedom$=$ Rd. When $\tilde{\boldsymbol{\Sigma}}_T$ is exactly rank R and features are whitened, Tripuraneni et al. (2020) provides a sample-complexity lower bound of $\Omega(Rd)$ examples and gives an algorithm achieving $\mathcal{O}(R^2 d)$ samples. However, in practice, deep nets learn good representations despite overparameterization. In this section, recalling our **Q2**, we argue that the inductive bias of the feature distribution can implicitly accelerate learning the canonical covariance. This differentiates our results from most prior works such as Kong et al. (2020b,a); Tripuraneni et al. (2020) in two aspects:

1. Rather than focusing on a *low dimensional* subspace and assuming $N \gtrsim Rd$, we can estimate $\tilde{\boldsymbol{\Sigma}}_T$ or $\tilde{\boldsymbol{S}}_T$ in the overparameterized regime $N \lesssim Rd$.
2. Rather than assuming whitened features $\boldsymbol{\Sigma}_F = \boldsymbol{I}$ and achieving a sample complexity of $R^2 d$, our learning guarantee holds for arbitrary covariance matrices $\boldsymbol{\Sigma}_F, \boldsymbol{\Sigma}_T$. The sample complexity depends on *effective rank* and can be arbitrarily smaller than DoF. We showcase our bounds via a spiked covariance setting in Example 1 below.

For learning $\tilde{\boldsymbol{\Sigma}}_T$ or its subspace $\tilde{\boldsymbol{S}}_T$, we investigate the method-of-moments (MoM) estimator.

Definition 4 (MoM Estimator)**.** *For* $1 \le i \le T$*, define* $\hat{\boldsymbol{b}}_{i,1} = 2n_1^{-1}\sum_{j=1}^{n_1/2} y_{i,j}\boldsymbol{x}_{i,j}$*,* $\hat{\boldsymbol{b}}_{i,2} = 2n_1^{-1}\sum_{j=n_1/2+1}^{n_1} y_{i,j}\boldsymbol{x}_{i,j}$*. Set*

$$\hat{\boldsymbol{M}} = n_1^{-1}\sum_{i=1}^{T}(\boldsymbol{b}_{i,1}\boldsymbol{b}_{i,2}^{\top} + \boldsymbol{b}_{i,2}\boldsymbol{b}_{i,1}^{\top}),$$

[5] We also need to estimate $\boldsymbol{\Sigma}_F$ for whitening. Estimating $\boldsymbol{\Sigma}_F$ is rather easy and incurs smaller error compared to $\tilde{\boldsymbol{\Sigma}}_T$. The analysis is provided in Appendix D.3.1.

The expectation of $\hat{\boldsymbol{M}}$ is equal to $\boldsymbol{M} = \boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F$.

Inductive bias in representation learning: Recall that canonical covariance $\tilde{\boldsymbol{\Sigma}}_T = \boldsymbol{\Sigma}_F^{1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F^{1/2}$ is the attribute of interest. However, feature covariance $\boldsymbol{\Sigma}_F^{1/2}$ term implicitly modulates the estimation procedure because the population MoM is not $\tilde{\boldsymbol{\Sigma}}_T$ but $\boldsymbol{M} = \boldsymbol{\Sigma}_F^{1/2}\tilde{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_F^{1/2}$. For instance, when estimating the principle canonical subspace $\tilde{\boldsymbol{S}}_T$, the degree of alignment between $\boldsymbol{\Sigma}_F$ and $\tilde{\boldsymbol{\Sigma}}_T$ can make or break the estimation procedure: If $\boldsymbol{\Sigma}_F$ and $\tilde{\boldsymbol{\Sigma}}_T$ have *well-aligned* principal subspaces, $\tilde{\boldsymbol{S}}_T$ will be easier to estimate since $\boldsymbol{\Sigma}_F$ will amplify the $\tilde{\boldsymbol{S}}_T$ direction within $\boldsymbol{M}$.

We verify the inductive bias on practical image dataset, reported in Appendix D.1. We assessed correlation coefficient between covariances $\tilde{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F$ via the canonical-feature alignment score defined as the correlation coefficient

$$\rho(\boldsymbol{\Sigma}_F, \tilde{\boldsymbol{\Sigma}}_T) := \frac{\left\langle \boldsymbol{\Sigma}_F, \tilde{\boldsymbol{\Sigma}}_T \right\rangle}{\|\boldsymbol{\Sigma}_F\|_F\|\tilde{\boldsymbol{\Sigma}}_T\|_F} = \frac{\text{trace}(\boldsymbol{M})}{\|\boldsymbol{\Sigma}_F\|_F\|\tilde{\boldsymbol{\Sigma}}_T\|_F}.$$

Observe that, the MoM estimator $\boldsymbol{M}$ naturally shows up in the alignment definition because the inner product of $\tilde{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F$ is equal to trace($\boldsymbol{M}$). This further supports our inductive bias intuition. As reference, we compared it to canonical-identity alignment defined as $\frac{\text{trace}(\tilde{\boldsymbol{\Sigma}}_T)}{\sqrt{d}\|\tilde{\boldsymbol{\Sigma}}_T\|_F}$ (replacing $\boldsymbol{\Sigma}_F$ with $\boldsymbol{I}$). The canonical-feature alignment score is higher than the canonical-identity alignment score. This significant score difference exemplifies how $\boldsymbol{\Sigma}_F$ and $\tilde{\boldsymbol{\Sigma}}_T$ can synergistically align with each other (inductive bias). This alignment helps our MoM estimator defined below, illustrated by Example 1 (spiked covariance).

In the following subsections, let $N = n_1T$ refer to the total tasks in representation-learning phase. Let $r_F = \mathbf{tr}(\boldsymbol{\Sigma}_F)$, $r_T = \mathbf{tr}(\boldsymbol{\Sigma}_T)$, and $r = \mathbf{tr}(\tilde{\boldsymbol{\Sigma}}_T)$. Define the approximate low-rankness measure of feature covariance by[6]

$$s_F = \min\ s'_F,\ \text{s.t. } s'_F \in \{1, ..., d\},\ s'_F/d \geq \lambda_{s'_F+1}(\boldsymbol{\Sigma}_F)$$

[6] The $(s_F + 1)$-th eigenvalue is smaller than s_F/d. Note the top eigenvalue is 1.

feature cov	$\boldsymbol{\Sigma}_F = \boldsymbol{I}$, $\boldsymbol{\Sigma}_T = \text{diag}(\boldsymbol{I}_{s_T}, \boldsymbol{0})$			$\boldsymbol{\Sigma}_F = \text{diag}(\boldsymbol{I}_{s_F}, \iota_F \boldsymbol{I}_{d-s_F})$, $\boldsymbol{\Sigma}_T = \text{diag}(\boldsymbol{I}_{s_T}, \iota_T \boldsymbol{I}_{d-s_T})$		
estimator	sample N	sample n_1	error	sample N	sample n_1	error
MoM	ds_T^2	1	$(ds_T^2/N)^{1/2}$	$r_F r_T^2$	1	$(r_F r_T^2/N)^{1/2}$
MoM	ds_T	s_T	$(s_T/n_1)^{1/2}$	$r_F r_T$	r_T	$(r_T/n_1)^{1/2}$

Table 5.2: **Right side:** Sample complexity and error of MoM estimators. s_F (s_T) is the dimension of the principal eigenspace of the feature (task) covariance. $r_F = s_F + \iota_F(d - s_F)$, $r_T = s_T + \iota_T(d - s_T)$ are the effective ranks. **Left side:** This is the well-studied setting of identity feature covariance and low-rank task covariance. Our bound in the second row is the first result to achieve optimal sample complexity of $\mathcal{O}(ds_T)$ (cf. Tripuraneni et al. (2020); Kong et al. (2020b)).

We have two results for this estimator.

1. Generally, we can estimate $\boldsymbol{M}$ with $\mathcal{O}(r_F r^2)$ samples.
2. Let $n_1 \geq s_T$, we can estimate $\boldsymbol{M}$ with $\mathcal{O}(s_F r)$ samples.

Tripuraneni et al. (2020) has sample complexity $\mathcal{O}(dr^2)$ (r is exact rank). Our sample complexity is $\mathcal{O}(r_F r^2)$. r_F, r can be seen as effective ranks and our bounds are always smaller than Tripuraneni et al. (2020). We will discuss later in Example 1. Our second result says when $n_1 \geq s_T$, our sample complexity achieves the $\mathcal{O}(dr)$ which is proven a lower bound in Tripuraneni et al. (2020).

Theorem 10. *Let data be generated as in Phase 1. Assume* $\|\boldsymbol{\Sigma}_F\|, \|\boldsymbol{\Sigma}_T\| = 1$ *for normalization*[7]*.*

1. *Let* n_1 *be a even number. Then with probability at least* $1 - N^{-100}$*,*

$$\|\hat{\boldsymbol{M}} - \boldsymbol{M}\| \lesssim (r + \sigma^2)\sqrt{\frac{r_F}{N}} + \sqrt{\frac{r_T}{T}}.$$

2. *Assume* $T \geq s_F$*. If* $n_1 \gtrsim r + \sigma^2$*, then with probability at least* $1 - CT^{-100}$ *for some constant*

[7] This is simply equivalent to scaling $y_{i,j}$, which does not affect the normalized error $\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|/\|\boldsymbol{M}\|$. In the appendix we define $\mathcal{S} = \max\{\|\boldsymbol{\Sigma}_F\|, \|\boldsymbol{\Sigma}_T\|\}$ and prove the theorem for general $\mathcal{S}$.

$C > 0$,

$$\|\hat{\boldsymbol{M}} - \boldsymbol{M}\| \lesssim \left((r+\sigma^2)/n_1\right)^{1/2}.$$

Denote the top-R principal subspaces of $\boldsymbol{M}, \hat{\boldsymbol{M}}$ by $\boldsymbol{M}_{top}, \hat{\boldsymbol{M}}_{top}$ and assume the eigen-gap condition $\lambda_R(\boldsymbol{M}) - \lambda_{R+1}(\boldsymbol{M}) > 2\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|$. Then a direct application of Davis-Kahan Theorem (Davis & Kahan, 1970) bounds the subspace angle as follows

$$\textit{angle}(\boldsymbol{M}_{top}, \hat{\boldsymbol{M}}_{top}) \lesssim \|\hat{\boldsymbol{M}} - \boldsymbol{M}\|/(\lambda_R(\boldsymbol{M}) - \lambda_{R+1}(\boldsymbol{M})).$$

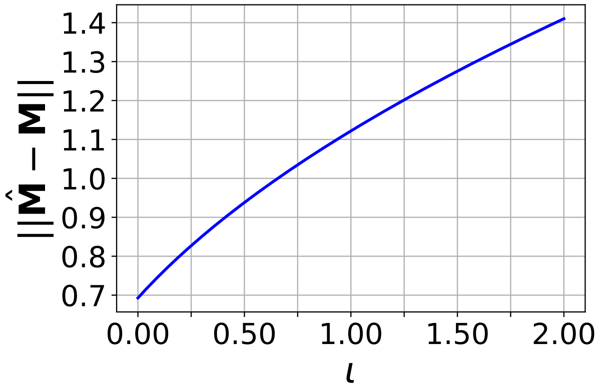


Figure 5.4: Error of MoM estimator

Estimating eigenspace of canonical covariance. Note that if $\boldsymbol{\Sigma}_F$ and $\boldsymbol{\Sigma}_T$ are aligned, (e.g. Example 1 below with $s_F = s_T = R$), then $\boldsymbol{M}_{\text{top}} = \tilde{\boldsymbol{S}}_T$ is exactly the principal subspace of $\tilde{\boldsymbol{\Sigma}}_T$. Theorem 10 indeed gives estimation error for the principal subspace of $\tilde{\boldsymbol{\Sigma}}_T$. Note that, such alignment is and more general requirement compared to related works which require whitened features (Tripuraneni et al., 2020; Kong et al., 2020b).

Example 1 (Spiked $\tilde{\boldsymbol{\Sigma}}_T$, Aligned principal subspaces)**.** *Suppose the spectra of $\boldsymbol{\Sigma}_F$ and $\tilde{\boldsymbol{\Sigma}}_T$ are bimodal as follows $\boldsymbol{\Sigma}_F = diag(\boldsymbol{I}_{s_F}, \iota_F \boldsymbol{I}_{d-s_F})$, $\boldsymbol{\Sigma}_T = diag(\boldsymbol{I}_{s_T}, \iota_T \boldsymbol{I}_{d-s_T})$. Set statistical error $Err_{T,N} := \sqrt{r_T^2 r_F/N} + \sqrt{r_T/T}$. When $\iota_T, \iota_F < 1$, $s_F \geq s_T$, the recovery error of $\tilde{\boldsymbol{\Sigma}}_T$ and its* *principal subspace* $\tilde{\boldsymbol{S}}_T$ *are bounded as*

$$angle(\hat{\boldsymbol{M}}_{top}, \tilde{\boldsymbol{S}}_T) \lesssim Err_{T,N} + \iota_F^2 \iota_T \quad and \quad \|\hat{\boldsymbol{M}} - \tilde{\boldsymbol{\Sigma}}_T\| \lesssim Err_{T,N} + \iota_F \iota_T.$$

The estimation errors for $\tilde{\boldsymbol{\Sigma}}_T, \tilde{\boldsymbol{S}}_T$ are controlled in terms of the effective ranks and the spectrum tails ι_F, ι_T. Typically $s_F s_T \gtrsim n_1$ so $\sqrt{r_T^2 r_F / N}$ term dominates the statistical error in practice. In Figure 5.4 we plot the error of estimating $\boldsymbol{M}$ (whose principal subspace coincides with $\tilde{\boldsymbol{\Sigma}}_T$). $\boldsymbol{\Sigma}_F = \mathrm{diag}(\boldsymbol{I}_{30}, \iota \boldsymbol{I}_{70})$, $\boldsymbol{\Sigma}_T = \mathrm{diag}(\boldsymbol{I}_{30}, \boldsymbol{0}_{70})$. $T = N = 100$. We can see that the error increase with ι .

5.5 Robustness of optimal representation and overall meta-learning bound

In Section 5.3, we described the algorithm for computing the optimal representation with *known* distributions of features and tasks. In Section 5.4, we proposed the MoM estimator in representation learning phase to estimate the unknown covariance matrices. In this section, we study the algorithm's behaviors when we calculate $\boldsymbol{\Lambda}$ using the *estimated* canonical covariance, rather than the full-information setting of Section 5.3.

Armed with the provably reliable estimators of Section 5.4, we can replace $\tilde{\boldsymbol{\Sigma}}_T$ and $\boldsymbol{\Sigma}_F$ in Algorithm 3 with our estimators. In this section, we inquire: how does the estimation error in covariance-estimation in representation learning stage affect the downstream few-shot learning risk? That says, we are interested in[8] $\mathrm{risk}(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F) - \mathrm{risk}(\boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F)$.

We replace the constraint in (OPT-REP) by $\underline{\theta} \leq \boldsymbol{\theta} \leq 1 - \frac{d-n_2}{n_2}\underline{\theta}$. This changes the "optimization" step in Algorithm 3. Theorem 11 does not require an explicit computation of the optimal representation by enforcing $\underline{\theta}$. Instead, we use the robustness of such a representation (due to its well-conditioned nature) to deduce its stability. Therefore, for practical computation of optimal representation, we simply use Algorithm 3. We can then evaluate $\underline{\theta}$ after-the-fact as the minimum singular value of this representation to apply

[8] Note that Sec.6 of Wu & Xu (2020) gives the exact value of $\mathrm{risk}(\boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F)$ so we have an end to end error guarantee.

Theorem 11 without assuming an explicit $\underline{\theta}$.

Let $\mathbf{\Lambda}_{\underline{\theta}}(R) = \textsc{ComputeOptimalRep}(R, \mathbf{\Sigma}_F, \hat{\boldsymbol{M}}, \sigma, n_2)$ denote the estimated optimal representation and $\mathbf{\Lambda}^*_{\underline{\theta}}(R) = \textsc{ComputeOptimalRep}(R, \mathbf{\Sigma}_F, \tilde{\mathbf{\Sigma}}_T, \sigma, n_2)$ denote the true optimal representation, which cannot be accessed in practice. Below we present the bound of the whole meta-learning algorithm. It shows that a bounded error in representation learning leads to a bounded increase on the downstream few-shot learning risk, thus quantifying the robustness of few-shot learning to errors in covariance estimates.

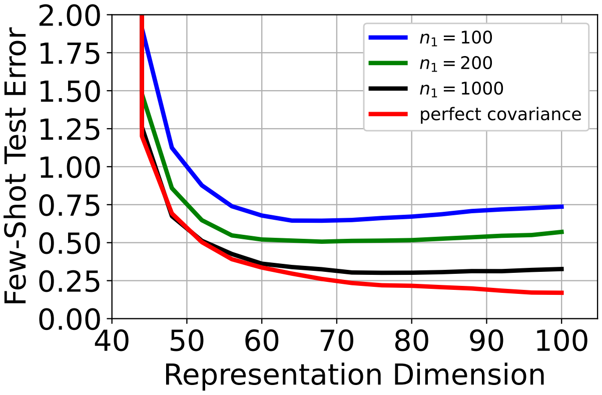


Figure 5.5: End to end learning guarantees. $d = 100, n_2 = 40, T = 200$, $\mathbf{\Sigma}_T = (\boldsymbol{I}_{20}, 0.05 \cdot \boldsymbol{I}_{80})$, $\mathbf{\Sigma}_F = \boldsymbol{I}_{100}$.

Theorem 11. *Let $\mathbf{\Lambda}_{\underline{\theta}}(R)$, $\mathbf{\Lambda}^*_{\underline{\theta}}(R)$ be as defined above, and $r_F = \mathbf{tr}(\mathbf{\Sigma}_F)$, $r_T = \mathbf{tr}(\mathbf{\Sigma}_T), r = \mathbf{tr}(\tilde{\mathbf{\Sigma}}_T)$. The risk of meta-learning algorithm satisfies*[9]

$$\textit{risk}(\mathbf{\Lambda}_{\underline{\theta}}(R), \mathbf{\Sigma}_T, \mathbf{\Sigma}_F) - \textit{risk}(\mathbf{\Lambda}^*_{\underline{\theta}}(R), \mathbf{\Sigma}_T, \mathbf{\Sigma}_F) \lesssim \frac{n_2^2}{d(R-n_2)(2n_2 - R\underline{\theta})\underline{\theta}} \left[(r+\sigma^2)\sqrt{\frac{r_F}{N}} + \sqrt{\frac{r_T}{T}} \right].$$

Notice that, as the number of previous tasks T and total representation learning samples N observed increase, the risk of the estimated $\mathbf{\Lambda}_{\underline{\theta}}(R)$ approaches the risk of the optimal $\mathbf{\Lambda}^*_{\underline{\theta}}(R)$ as we expect. The result only applies to the overparameterized regime of interest

[9] The bracketed expression applies first conclusion of Theorem 11. One can plug in the second as well.

$R > n_2$. The expression of risk in the underparameterized case is different, and covered by the second case of (Wu & Xu, 2020, Eq(4.4)). We plot it in Fig 5.1(b) on the left side of the peak as a comparison.

Risk with respect to PCA level R. In Fig. 5.5, we plot the error of the whole meta-learning algorithm. We simulate representation learning and get $\hat{\boldsymbol{M}}$, use it to compute $\boldsymbol{\Lambda}$ and plot the theoretical downstream risk (experiments match, see Fig. 5.1 (b)). Mainly, we compare the behavior of Theorem 11 with different R. When R grows, we search $\boldsymbol{\Lambda}$ in a larger space. The optimal representation $\boldsymbol{\Lambda}$ in a feasible *sub*set is always no better than searching in a larger space, thus the risk decreases with R increasing. At the same time, representation learning error increases with R since we need to fit a matrix in a larger space. In essence, this result provides a theoretical justification on a sweet-spot for the optimal representation. $d = R$ is optimal when $N = \infty$, i.e., representation learning error is 0. As N decreases, there is a tradeoff between learning error and truncating small eigenvalues. Thus choosing R adaptively with N can strike the right bias-variance tradeoff between the excess risk (variance) and the risk due to suboptimal representation.

5.6 Conclusion and future directions

We study the sample efficiency of meta-learning with linear representations. We show that the optimal representation is typically overparameterized and outperforms subspace-based representations for general data distributions and refine the sample complexity analysis for learning arbitrary distributions and show the importance of inductive bias of feature and task. Finally we provide an end-to-end bound for the meta-learning algorithm showing the tradeoff of choosing larger representation dimension v.s. robustness against representation learning error.

Our optimal representation works with jointly diagonalizable covariances and is asymptotic (although this is also the case in literature such as Chang et al. (2020); Wu & Xu (2020)). The setting is limited to mixed linear regression while linearized settings (such as neural tangent kernel Jacot et al. (2018); Du et al. (2018); Arora et al. (2019)) are helpful for understanding nonlinear models, and the nonlinear meta-learning can be the next step.

BIBLIOGRAPHY

Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 1–26, 2011.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009a.

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009b.

Agarwal, N., Boumal, N., Bullins, B., and Cartis, C. Adaptive regularization with cubics on manifolds. *arXiv preprint arXiv:1806.00065*, 2018.

Agarwal, N., Bullins, B., Hazan, E., Kakade, S. M., and Singh, K. Online control with adversarial disturbances. *arXiv preprint arXiv:1902.08721*, 2019.

Alonso, C. A., Yang, F., and Matni, N. Data-driven distributed and localized model predictive control. *arXiv preprint arXiv:2112.12229*, 2021.

Ambrose, W. and Singer, I. M. A theorem on holonomy. *Transactions of the American Mathematical Society*, 75(3):428–443, 1953.

Anderson, J., Doyle, J. C., Low, S. H., and Matni, N. System level synthesis. *Annual Reviews in Control*, 47:364–393, 2019.

Anderson, T. W. et al. Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. *Essays in probability and statistics*, pp. 1–24, 1970.

Argyriou, A., Evgeniou, T., and Pontil, M. Convex multi-task feature learning. *Machine learning*, 73(3):243–272, 2008.

Arias-Castro, E., Candes, E. J., and Davenport, M. A. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2012.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332, 2019.

Avdiukhin, D., Jin, C., and Yaroslavtsev, G. Escaping saddle points with inequality constraints via noisy sticky projected gradient descent. In *11th Annual Workshop on Optimization for Machine Learning*, 2019.

Ayazoglu, M. and Sznaier, M. An algorithm for fast constrained nuclear norm minimization and applications to systems identification. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 3469–3475. IEEE, 2012.

Bahmani, S. and Romberg, J. Convex programming for estimation in nonlinear recurrent models. *arXiv preprint arXiv:1908.09915*, 2019.

Balakrishnan, V. and Vandenberghe, L. Semidefinite programming duality and linear time-invariant systems. *IEEE Transactions on Automatic Control*, 48(1):30–41, 2003.

Balcan, M.-F., Blum, A., and Vempala, S. Efficient representations for lifelong learning and autoencoding. In *Conference on Learning Theory*, pp. 191–210, 2015.

Banerjee, A., Chen, S., Fazayeli, F., and Sivakumar, V. Estimation with norm regularization. In *Advances in Neural Information Processing Systems*, pp. 1556–1564, 2014.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Baxter, J. A model of inductive bias learning. *Journal of artificial intelligence research*, 12: 149–198, 2000.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Blomberg, N. *On nuclear norm minimization in system identification.* PhD thesis, KTH Royal Institute of Technology, 2016.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pp. 12–58, 2014.

Bonnabel, S. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Boumal, N. and Absil, P.-a. Rtrmc: A riemannian trust-region method for low-rank matrix completion. In *Advances in neural information processing systems*, pp. 406–414, 2011.

Boumal, N., Absil, P.-A., and Cartis, C. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 2016a.

Boumal, N., Voroninski, V., and Bandeira, A. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pp. 2757–2765, 2016b.

Boumal, N., Absil, P.-A., and Cartis, C. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, pp. drx080, 2018. doi: 10.1093/imanum/drx080. URL `http://dx.doi.org/10.1093/imanum/drx080`.

Bouniot, Q., Redko, I., Audigier, R., Loesch, A., Zotkin, Y., and Habrard, A. Towards better understanding meta-learning methods through multi-task representation learning theory. *arXiv preprint arXiv:2010.01992*, 2020.

Boyd, S., El Ghaoui, L., Feron, E., and Balakrishnan, V. *Linear matrix inequalities in system and control theory*. SIAM, 1994.

Bradtke, S. J., Ydstie, B. E., and Barto, A. G. Adaptive linear quadratic control using policy iteration. In *Proceedings of 1994 American Control Conference-ACC'94*, volume 3, pp. 3475–3479. IEEE, 1994.

Bu, J., Mesbahi, A., Fazel, M., and Mesbahi, M. Lqr through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019a.

Bu, J., Ratliff, L. J., and Mesbahi, M. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv preprint arXiv:1911.04672*, 2019b.

Bu, J., Mesbahi, A., and Mesbahi, M. Policy gradient-based algorithms for continuous-time linear quadratic control. *arXiv preprint arXiv:2006.09178*, 2020.

Cadzow, J. A. Signal enhancement-a composite property mapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(1):49–62, 1988.

Cai, J.-F., Qu, X., Xu, W., and Ye, G.-B. Robust recovery of complex exponential signals from random gaussian projections via low rank hankel matrix reconstruction. *Applied and computational harmonic analysis*, 41(2):470–490, 2016.

Candes, E. J. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 98(6): 925–936, 2010.

Carmon, Y. and Duchi, J. C. Gradient descent efficiently finds the cubic-regularized nonconvex newton step. *arXiv preprint arXiv:1612.00547*, 2017.

Caruana, R. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. Linear algorithms for online multitask classification. *The Journal of Machine Learning Research*, 11:2901–2934, 2010.

Chang, X., Li, Y., Oymak, S., and Thrampoulidis, C. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. *arXiv preprint arXiv:2012.08749*, 2020.

Cheeger, J. and Ebin, D. G. *Comparison Theorems in Riemannian Geometry*. AMS Chelsea Publishing, Providence, RI, 2008.

Chen, S., Li, J., and Song, Z. Learning mixtures of linear regressions in subexponential time via fourier moments. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 587–600, 2020.

Chen, Y. and Poor, H. V. Learning mixtures of linear dynamical systems, 2022.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

Collins, L., Mokhtari, A., Oh, S., and Shakkottai, S. MAML and ANIL provably learn representations. *arXiv preprint arXiv:2202.03483*, 2022.

Costa, O. L. d. V. *Discrete-time Markov jump linear systems*. Probability and its applications (Springer-Verlag). Springer, London, 2005. ISBN 1852337613.

Criscitiello, C. and Boumal, N. Efficiently escaping saddle points on manifolds. *arXiv preprint arXiv:1906.04321*, 2019.

Davis, C. and Kahan, W. M. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

De Moor, B., De Gersem, P., De Schutter, B., and Favoreel, W. Daisy: A database for identification of systems. *JOURNAL A*, 38:4–5, 1997.

Dean, S., Tu, S., Matni, N., and Recht, B. Safely learning to control the constrained linear quadratic regulator. In *2019 American Control Conference (ACC)*, pp. 5582–5588. IEEE, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ding, T., Sznaier, M., and Camps, O. I. A rank minimization approach to video inpainting. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. IEEE, 2007.

Do Carmo, M. P. *Differential Geometry of Curves and Surfaces.* Courier Dover Publications, 2016.

Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Poczos, B. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pp. 1067–1077, 2017.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Du, Z., Ozay, N., and Balzano, L. Mode clustering for markov jump systems. In *2019*

IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pp. 126–130. IEEE, 2019.

Du, Z., Sattar, Y., Tarzanagh, D. A., Balzano, L., Oymak, S., and Ozay, N. Certainty equivalent quadratic control for markov jump systems. *arXiv preprint arXiv:2105.12358*, 2021.

Duan, J., Li, J., and Zhao, L. Optimization landscape of gradient descent for discrete-time static output feedback. *arXiv preprint arXiv:2109.13132*, 2021.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Dullerud, G. E. and Paganini, F. *A course in robust control theory: a convex approach*, volume 36. Springer Science & Business Media, 2013.

Durmus, A., Jiménez, P., Moulines, É., Said, S., and Wai, H.-T. Convergence analysis of riemannian stochastic approximation schemes. *arXiv preprint arXiv:2005.13284*, 2020.

Edelman, A., Arias, T. A., and Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Elad, M., Milanfar, P., and Golub, G. H. Shape from moments-an estimation theory perspective. *IEEE Transactions on Signal Processing*, 52(7):1814–1829, 2004.

Fattahi, S. Learning partially observed linear dynamical systems from logarithmic number of samples. *arXiv preprint arXiv:2010.04015*, 2020.

Fazel, M. Matrix rank minimization with applications. 2002.

Fazel, M., Hindi, H., and Boyd, S. P. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, volume 6, pp. 4734–4739. IEEE, 2001.

Fazel, M., Pong, T. K., Sun, D., and Tseng, P. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013.

Fazel, M., Ge, R., Kakade, S. M., and Mesbahi, M. Global convergence of policy gradient methods for linearized control problems. *arXiv preprint arXiv:1801.05039*, 2018.

Feng, H. and Lavaei, J. Connectivity properties of the set of stabilizing static decentralized controllers. *SIAM Journal on Control and Optimization*, 58(5):2790–2820, 2020.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135, 2017.

Foster, D. J., Rakhlin, A., and Sarkar, T. Learning nonlinear dynamical systems from a single trajectory. *arXiv preprint arXiv:2004.14681*, 2020.

Furieri, L., Zheng, Y., and Kamgarpour, M. Learning the globally optimal distributed lq regulator. In *Learning for Dynamics and Control*, pp. 287–297. PMLR, 2020.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pp. 797–842, 2015.

Gillard, J. Cadzow's basic algorithm, alternating projections and singular spectrum analysis. *Statistics and its Interface*, 3(3):335–343, 2010.

Gordon, Y. On milman's inequality and random subspaces which escape through a mesh in r n. In *Geometric aspects of functional analysis*, pp. 84–106. Springer, 1988.

Grossmann, C., Jones, C. N., and Morari, M. System identification with missing data via nuclear norm regularization. In *2009 European Control Conference (ECC)*, pp. 448–453. IEEE, 2009.

Gulluk, H. I., Sun, Y., Oymak, S., and Fazel, M. Sample efficient subspace-based representations for nonlinear meta-learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3685–3689. IEEE, 2021.

Hansson, A., Liu, Z., and Vandenberghe, L. Subspace system identification via weighted nuclear norm optimization. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 3439–3444. IEEE, 2012.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Hazan, E., Lee, H., Singh, K., Zhang, C., and Zhang, Y. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, pp. 4634–4643, 2018.

Hewer, G. An iterative technique for the computation of the steady state gains for the discrete optimal regulator. *IEEE Transactions on Automatic Control*, 16(4):382–384, 1971.

Ho, B. and Kálmán, R. E. Effective construction of linear state-variable models from input/output functions. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

Hu, J., Milzarek, A., Wen, Z., and Yuan, Y. Adaptive quadratically regularized Newton method for Riemannian optimization. *SIAM J. Matrix Anal. Appl.*, 39(3):1181–1207, 2018.

Ishteva, M., Absil, P.-A., Van Huffel, S., and De Lathauwer, L. Best low multilinear rank approximation of higher-order tensors, based on the riemannian trust-region scheme. *SIAM Journal on Matrix Analysis and Applications*, 32(1):115–135, 2011.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Jansch-Porto, J. P., Hu, B., and Dullerud, G. E. Convergence guarantees of policy optimization methods for markovian jump linear systems. In *2020 American Control Conference (ACC)*, pp. 2882–2887. IEEE, 2020.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1724–1732. JMLR. org, 2017a.

Jin, C., Netrapalli, P., and Jordan, M. I. Accelerated gradient descent escapes saddle points faster than gradient descent. *arXiv preprint arXiv:1711.10456*, 2017b.

Kakade, S. M. A natural policy gradient. In *Advances in neural information processing systems*, pp. 1531–1538, 2002.

Kalman, R. E. et al. Contributions to the theory of optimal control. *Bol. soc. mat. mexicana*, 5(2):102–119, 1960.

Karcher, H. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.

Kasai, H. and Mishra, B. Inexact trust-region algorithms on riemannian manifolds. In *Advances in Neural Information Processing Systems 31*, pp. 4254–4265. 2018.

Khosravi, M. and Smith, R. S. Nonlinear system identification with prior knowledge of the region of attraction. *arXiv preprint arXiv:2003.12330*, 2020.

Khuzani, M. B. and Li, N. Stochastic primal-dual method on riemannian manifolds with bounded sectional curvature. *arXiv preprint arXiv:1703.08167*, 2017.

Kong, W., Somani, R., Kakade, S., and Oh, S. Robust meta-learning for mixed linear regression with small batches. *arXiv preprint arXiv:2006.09702*, 2020a.

Kong, W., Somani, R., Song, Z., Kakade, S., and Oh, S. Meta-learning for mixed linear regression. *arXiv preprint arXiv:2002.08936*, 2020b.

Krahmer, F., Mendelson, S., and Rauhut, H. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904, 2014.

Lancaster, P. and Rodman, L. *Algebraic riccati equations.* Clarendon press, 1995.

Lasserre, J. B. Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization*, 11(3):796–817, 2001.

Lee, D. and Hu, J. Primal-dual q-learning framework for lqr design. *IEEE Transactions on Automatic Control*, 64(9):3756–3763, 2018.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32:8572–8583, 2019.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. *Conference on Learning Theory*, pp. 1246–1257, 2016.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

Lee, J. M. *Riemannian manifolds : an introduction to curvature.* Graduate texts in mathematics ; 176. Springer, New York, 1997. ISBN 9780387227269.

Lee, J. Y., Park, J. B., and Choi, Y. H. Integral q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems. *Automatica*, 48(11):2850–2859, 2012.

Lewis, F. L. and Vrabie, D. Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE circuits and systems magazine*, 9(3):32–50, 2009.

Li, Y. and Liang, Y. Learning mixtures of linear regressions with nearly optimal complexity. In *Conference On Learning Theory*, pp. 1125–1144, 2018.

Li, Y., Tang, Y., Zhang, R., and Li, N. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. *IEEE Transactions on Automatic Control*, 2021.

Liu, Z., Hansson, A., and Vandenberghe, L. Nuclear norm system identification with missing inputs and outputs. *Systems & Control Letters*, 62(8):605–612, 2013.

Ljung, L. System identification: theory for the user. *PTR Prentice Hall, Upper Saddle River, NJ*, pp. 1–14, 1999.

Lojasiewicz, S. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117:87–89, 1963.

Lounici, K., Pontil, M., Van De Geer, S., Tsybakov, A. B., et al. Oracle inequalities and optimal inference under group sparsity. *The annals of statistics*, 39(4):2164–2204, 2011.

Lu, S., Razaviyayn, M., Yang, B., Huang, K., and Hong, M. Snap: Finding approximate second-order stationary solutions efficiently for non-convex linearly constrained problems. *arXiv preprint arXiv:1907.04450*, 2019a.

Lu, S., Zhao, Z., Huang, K., and Hong, M. Perturbed projected gradient descent converges to approximate second-order points for bound constrained nonconvex problems. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5356–5360. IEEE, 2019b.

Lu, Y. and Mo, Y. Non-episodic learning for online lqr of unknown linear gaussian system. *arXiv preprint arXiv:2103.13278*, 2021.

Lucas, J., Ren, M., Kameni, I., Pitassi, T., and Zemel, R. Theoretical bounds on estimation error for meta-learning. *arXiv preprint arXiv:2010.07140*, 2020.

Malik, D., Pananjady, A., Bhatia, K., Khamaru, K., Bartlett, P., and Wainwright, M. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems.

In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2916–2925. PMLR, 2019.

Mangoubi, O., Smith, A., et al. Rapid mixing of geodesic walks on manifolds with positive curvature. *The Annals of Applied Probability*, 28(4):2501–2543, 2018.

Mania, H., Tu, S., and Recht, B. Certainty equivalent control of lqr is efficient. *arXiv preprint arXiv:1902.07826*, 2019.

Mania, H., Jordan, M. I., and Recht, B. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.

Mårtensson, K. *Gradient methods for large-scale and distributed linear quadratic control.* PhD thesis, Lund University, 2012.

Maurer, A., Pontil, M., and Romera-Paredes, B. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1):2853–2884, 2016.

McCoy, M. B. and Tropp, J. A. The achievable performance of convex demixing. *arXiv preprint arXiv:1309.7478*, 2013.

McKelvey, T., Akçay, H., and Ljung, L. Subspace-based multivariable system identification from frequency response data. *IEEE Transactions on Automatic Control*, 41(7):960–979, 1996.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Mhammedi, Z., Foster, D. J., Simchowitz, M., Misra, D., Sun, W., Krishnamurthy, A., Rakhlin, A., and Langford, J. Learning the linear quadratic regulator from nonlinear observations. *arXiv preprint arXiv:2010.03799*, 2020.

Mohammadi, H., Zare, A., Soltanolkotabi, M., and Jovanović, M. R. Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem. *arXiv preprint arXiv:1912.11899*, 2019a.

Mohammadi, H., Zare, A., Soltanolkotabi, M., and Jovanović, M. R. Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 7474–7479. IEEE, 2019b.

Mokhtari, A., Ozdaglar, A., and Jadbabaie, A. Escaping saddle points in constrained optimization. *arXiv preprint arXiv:1809.02162*, 2018.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime, 2020.

Muthukumar, V., Vodrahalli, K., and Sahai, A. Harmless interpolation of noisy data in regression. *CoRR*, abs/1903.09139, 2019. URL `http://arxiv.org/abs/1903.09139`.

Muthukumar, V., Narang, A., Subramanian, V., Belkin, M., Hsu, D., and Sahai, A. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv preprint arXiv:2005.08054*, 2020.

Nakkiran, P., Venkat, P., Kakade, S., and Ma, T. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

Nouiehed, M., Lee, J. D., and Razaviyayn, M. Convergence to second-order stationarity for constrained non-convex optimization. *arXiv preprint arXiv:1810.02024*, 2018.

Oymak, S. Stochastic gradient descent learns state equations with nonlinear activations. In *Conference on Learning Theory*, pp. 2551–2579, 2019.

Oymak, S. and Ozay, N. Non-asymptotic identification of lti systems from a single trajectory. *arXiv preprint arXiv:1806.05722*, 2018.

Oymak, S., Thrampoulidis, C., and Hassibi, B. Simple bounds for noisy linear inverse problems with exact side information. *arXiv preprint arXiv:1312.0641*, 2013.

Oymak, S., Fabian, Z., Li, M., and Soltanolkotabi, M. Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *arXiv preprint arXiv:1906.05392*, 2019.

Pemantle, R. Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, pp. 698–712, 1990.

Perdomo, J. C., Umenberger, J., and Simchowitz, M. Stabilizing dynamical systems via policy gradient methods. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Rajeswaran, A., Lowrey, K., Todorov, E. V., and Kakade, S. M. Towards generalization and simplicity in continuous control. In *Advances in Neural Information Processing Systems*, pp. 6550–6561, 2017.

Rapcsák, T. Sectional curvatures in nonlinear optimization. *Journal of Global Optimization*, 40(1-3):375–388, 2008.

Rawlings, J. B., Mayne, D. Q., and Diehl, M. *Model predictive control: theory, computation, and design*, volume 2. Nob Hill Publishing Madison, WI, 2017.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Reyhanian, N. and Haupt, J. Online stochastic gradient descent learns linear dynamical systems from a single trajectory. *arXiv preprint arXiv:2102.11822*, 2021.

Roberts, J. W., Manchester, I. R., and Tedrake, R. Feedback controller parameterizations for reinforcement learning. In *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pp. 310–317. IEEE, 2011.

Rutledge, K., Yong, S. Z., and Ozay, N. Finite horizon constrained control and bounded-error estimation in the presence of missing data. *Nonlinear Analysis: Hybrid Systems*, 36:100854, 2020.

Sakai, T. *Riemannian Geometry*, volume 149 of *Translations of Mathematical Monographs*. American Mathematical Society, 1996.

Sanchez-Pena, R. S. and Sznaier, M. *Robust systems theory and applications*. Wiley-Interscience, 1998.

Sarkar, T. and Rakhlin, A. Near optimal finite time identification of arbitrary linear dynamical systems. *arXiv preprint arXiv:1812.01251*, 2019.

Sarkar, T., Rakhlin, A., and Dahleh, M. A. Finite-time system identification for partially observed lti systems of unknown order. *arXiv preprint arXiv:1902.01848*, 2019.

Sarkar, T. K. and Pereira, O. Using the matrix pencil method to estimate the parameters of a sum of complex exponentials. *IEEE Antennas and Propagation Magazine*, 37(1):48–55, 1995.

Sattar, Y. and Oymak, S. Non-asymptotic and accurate learning of nonlinear dynamical systems. *arXiv preprint arXiv:2002.08538*, 2020.

Sattar, Y., Du, Z., Tarzanagh, D. A., Balzano, L., Ozay, N., and Oymak, S. Identification and adaptive control of markov jump systems: Sample complexity and regret bounds. *arXiv preprint arXiv:2111.07018*, 2021.

Scherer, C. and Weiland, S. Linear matrix inequalities in control. *Lecture Notes, Dutch Institute for Systems and Control, Delft, The Netherlands*, 3(2), 2000.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pp. 439–473, 2018.

Simchowitz, M., Boczar, R., and Recht, B. Learning linear dynamical systems with semi-parametric least squares. *arXiv preprint arXiv:1902.00768*, 2019.

Sontag, E. D. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer Science & Business Media, 2013.

Stengel, R. F. *Optimal control and estimation.* Courier Corporation, 1994.

Sun, J., Qu, Q., and Wright, J. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2): 885–914, 2017.

Sun, Y. and Fazel, M. Escaping saddle points efficiently in equality-constrained optimization problems. In *Workshop on Modern Trends in Nonconvex Optimization for Machine Learning, International Conference on Machine Learning*, 2018.

Sun, Y. and Fazel, M. Learning optimal controllers by policy gradient: Global optimality via convex parameterization. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 4576–4581. IEEE, 2021.

Sun, Y., Flammarion, N., and Fazel, M. Escaping from saddle points on riemannian manifolds. In *Advances in Neural Information Processing Systems*, pp. 7276–7286, 2019.

Sun, Y., Oymak, S., and Fazel, M. Finite sample system identification: Optimal rates and the role of regularization. In *Learning for Dynamics and Control*, pp. 16–25. PMLR, 2020.

Sun, Y., Narang, A., Gulluk, H. I., Oymak, S., and Fazel, M. Towards sample-efficient overparameterized meta-learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=-KU_e4Biu0`.

Talebi, S., Alemzadeh, S., Rahimi, N., and Mesbahi, M. Online regulation of unstable linear systems from a single trajectory. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 4784–4789. IEEE, 2020.

Tang, Y., Zheng, Y., and Li, N. Analysis of the optimization landscape of linear quadratic gaussian (lqg) control. In *Learning for Dynamics and Control*, pp. 599–610. PMLR, 2021.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Sample efficient linear meta-learning by alternating minimization. *arXiv preprint arXiv:2105.08306*, 2021.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*, pp. 3420–3428, 2015.

Tripuraneni, N., Flammarion, N., Bach, F., and Jordan, M. I. Averaging Stochastic Gradient Descent on Riemannian Manifolds. *arXiv preprint arXiv:1802.09128*, 2018.

Tripuraneni, N., Jin, C., and Jordan, M. I. Provable meta-learning of linear representations. *arXiv preprint arXiv:2002.11684*, 2020.

Tsiamis, A. and Pappas, G. J. Finite sample analysis of stochastic system identification. *arXiv preprint arXiv:1903.09122*, 2019.

Tu, L. W. *Differential geometry : connections, curvature, and characteristic classes*. Graduate texts in mathematics ; 275. Springer, Cham, Switzerland, 2017. ISBN 9783319550848.

Tu, S., Boczar, R., Packard, A., and Recht, B. Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*, 2017.

Umenberger, J., Simchowitz, M., Perdomo, J. C., Zhang, K., and Tedrake, R. Globally convergent policy search over dynamic filters for output estimation. *arXiv preprint arXiv:2202.11659*, 2022.

Van Overschee, P. and De Moor, B. A unifying theorem for three subspace system identification algorithms. *Automatica*, 31(12):1853–1864, 1995.

Van Overschee, P. and De Moor, B. *Subspace identification for linear systems: Theory–Implementation–Applications*. Springer Science & Business Media, 2012.

Verhaegen, M. and Hansson, A. N2sid: Nuclear norm subspace identification of innovation models. *Automatica*, 72:57–63, 2016.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., and Wierstra, D. Matching networks for one shot learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf`.

Wagenmaker, A. and Jamieson, K. Active learning for identification of linear dynamical systems. *arXiv preprint arXiv:2002.00495*, 2020.

Wong, Y.-c. Sectional curvatures of Grassmann manifolds. *Proc. Nat. Acad. Sci. U.S.A.*, 60: 75–79, 1968.

Wu, D. and Xu, J. On the optimal weighted ℓ_2 regularization in overparameterized linear regression, 2020.

Xu, W., Yi, J., Dasgupta, S., Cai, J.-F., Jacob, M., and Cho, M. Sep] ration-free super-resolution from compressed measurements is possible: an orthonormal atomic norm minimization approach. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 76–80. IEEE, 2018.

Yang, J., Hu, W., Lee, J. D., and Du, S. S. Provable benefits of representation learning in linear bandits. *arXiv preprint arXiv:2010.06531*, 2020.

Yang, J., Hu, W., Lee, J. D., and Du, S. S. Impact of representation learning in linear bandits. In *International Conference on Learning Representations*, 2021.

Zare, A., Mohammadi, H., Dhingra, N. K., Georgiou, T. T., and Jovanović, M. R. Proximal algorithms for large-scale statistical modeling and sensor/actuator selection. *IEEE Transactions on Automatic Control*, 65(8):3441–3456, 2019.

Zhang, D. and Tajbakhsh, S. D. Riemannian stochastic variance-reduced cubic regularized newton method. *arXiv preprint arXiv:2010.03785*, 2020.

Zhang, H. and Sra, S. First-order methods for geodesically convex optimization. *arXiv:1602.06053*, 2016. *Preprint*.

Zhang, H., Reddi, S. J., and Sra, S. Riemannian svrg: fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*, pp. 4592–4600, 2016.

Zhang, J. and Zhang, S. A cubic regularized newton's method over riemannian manifolds. *arXiv preprint arXiv:1805.05565*, 2018.

Zhang, K., Yang, Z., and Basar, T. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. *Advances in Neural Information Processing Systems*, 32:11602–11614, 2019.

Zhang, K., Hu, B., and Basar, T. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. In *Learning for Dynamics and Control*, pp. 179–190, 2020.

Zheng, Y. and Li, N. Non-asymptotic identification of linear dynamical systems using multiple trajectories. *IEEE Control Systems Letters*, 5(5):1693–1698, 2020.

Zhong, K., Jain, P., and Dhillon, I. S. Mixed linear regression with multiple components. In *Advances in neural information processing systems*, pp. 2190–2198, 2016.

Appendix A

APPENDIX OF CHAPTER 2

A.1 Taylor expansions on Riemannian manifold

We provide here the Taylor expansion for functions and gradients of functions defined on a Riemannian manifold.

Taylor expansion for the gradient

For any point $x \in \mathcal{M}$ and $z \in \mathcal{M}$ be a point in the neighborhood of x where the geodesic $\gamma_{x\to z}$ is defined.

$$\begin{aligned}\Gamma_z^x(\mathrm{grad}f(z)) &= \mathrm{grad}f(x) + \nabla_{\gamma'_{x\to z}(0)}\mathrm{grad}f + \textstyle\int_0^1 (\Gamma^x_{\gamma_{x\to z}(\tau)}\nabla_{\gamma'_{x\to z}(\tau)}\mathrm{grad}f - \nabla_{\gamma'_{x\to z}(0)}\mathrm{grad}f)dx_\tau \\ &= \mathrm{grad}f(x) + \nabla_{\gamma'_{x\to z}(0)}\mathrm{grad}f + \Delta(z), \qquad\qquad (A.1)\end{aligned}$$

where $\Delta(z) := \int_0^1 (\Gamma^x_{\gamma_{x\to z}(\tau)}\nabla_{\gamma'_{x\to z}(\tau)}\mathrm{grad}f - \nabla_{\gamma'_{x\to z}(0)}\mathrm{grad}f)d\tau$. The Taylor approximation in Eq. (A.1) is proven by Absil et al. (2009a, Lemma 7.4.7).

Taylor expansion for the function

Taylor expansion of the gradient enables us to approximate the iterations of the main algorithm, but obtaining the convergence rate of the algorithm requires proving that the function value decreases following the iterations. We need to give the Taylor expansion of f with the parallel translated gradient on LHS of Eq. (A.1). To simplify the notation, let γ

denote the $\gamma_{x\to z}$.

$$\begin{aligned}
f(z)-f(x) &= \int_0^1 \frac{d}{d\tau} f(\gamma(\tau))d\tau && \text{(A.2a)}\\
&= \int_0^1 \langle \gamma'(\tau), \mathrm{grad} f(\gamma(\tau))\rangle d\tau && \text{(A.2b)}\\
&= \int_0^1 \langle \Gamma^x_{\gamma(\tau)}\gamma'(\tau), \Gamma^x_{\gamma(\tau)}\mathrm{grad} f(\gamma(\tau))\rangle d\tau && \text{(A.2c)}\\
&= \int_0^1 \langle \gamma'(0), \Gamma^0_{\gamma(\tau)}\mathrm{grad} f(\gamma(\tau))\rangle d\tau && \text{(A.2d)}\\
&= \int_0^1 \langle \gamma'(0), \mathrm{grad} f(x) + \nabla_{\tau\gamma'(0)}\mathrm{grad} f + \Delta(\gamma(\tau))\rangle d\tau && \text{(A.2e)}\\
&= \langle \gamma'(0), \mathrm{grad} f(x) + \tfrac{1}{2}\nabla_{\gamma'(0)}\mathrm{grad} f + \bar{\Delta}(z)\rangle. && \text{(A.2f)}
\end{aligned}$$

$\Delta(z)$ is defined in Eq. (A.1). $\bar{\Delta}(z) = \int_0^1 \Delta(\gamma(\tau))d\tau$. The second line is just rewriting by definition. Eq. (A.2c) means the parallel translation preserves the inner product (Tu, 2017, Prop. 14.16). Eq. (A.2d) uses $\Gamma^x_{\gamma(t)}\gamma'(t) = \gamma'(0)$, meaning that the velocity stays constant along a geodesic (Absil et al., 2009a, (5.23)). Eq. (A.2e) uses Eq. (A.1). In Euclidean space, the Taylor expansion is

$$f(z) - f(x) = \langle z, \nabla f(x) + \nabla^2 f(x)z + \int_0^1 (\nabla^2 f(\tau z) - \nabla^2 f(x))z d\tau\rangle. \tag{A.3}$$

Compare Eq. (A.2) and Eq. (A.3), z is replaced by $\gamma'(0) := \gamma'_{x\to z}(0)$ and τz is replaced by $\tau\gamma'_{x\to z}(0)$ or $\gamma_{x\to z}(\tau)$.

Now we have

$$f(u_t) = f(x) + \langle \gamma'(0), \mathrm{grad} f(x)\rangle + \frac{1}{2}H(x)[\gamma'(0), \gamma'(0)] + \langle \gamma'(0), \bar{\Delta}(u_t)\rangle.$$

A.2 Linearization of the iterates in a fixed tangent space

In this section we linearize the progress of the iterates of our algorithm in a fixed tangent space $\mathcal{T}_x\mathcal{M}$. We always assume here that all points are within a region of diameter $R := 12\mathscr{S} \le \mathfrak{I}$.

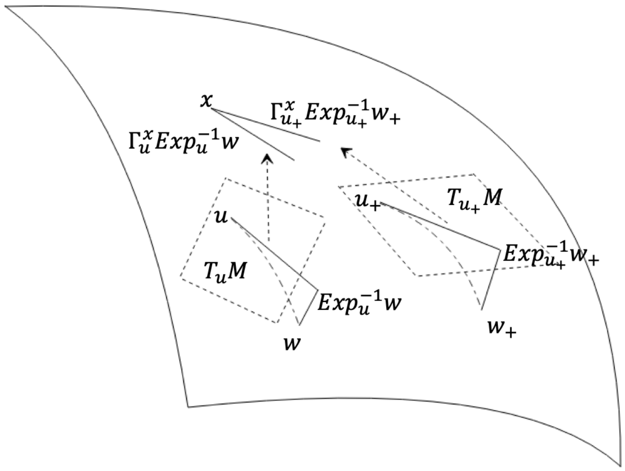


Figure A.1: Lemma 10. First map w and w_+ to $\mathcal{T}_u\mathcal{M}$ and $\mathcal{T}_{u_+}\mathcal{M}$, and transport the two vectors to $\mathcal{T}_x\mathcal{M}$, and get their relation.

In the course of the proof we need several auxilliary lemmas which are stated in the last two subsections of this section.

Evolution of $\mathrm{Exp}_u^{-1}(w)$

We first consider the evolution of $\mathrm{Exp}_u^{-1}(w)$ in a fixed tangent space $\mathcal{T}_x\mathcal{M}$. We show in the following lemma that it approximately follows a linear reccursion.

Lemma 10. *Define* $\gamma = \sqrt{\hat{\rho}\epsilon}$*,* $\kappa = \frac{\beta}{\gamma}$*, and* $\mathscr{S} = \sqrt{\eta\beta}\frac{\gamma}{\hat{\rho}}\log^{-1}(\frac{d\kappa}{\delta})$*. Let us consider* x *be a* $(\epsilon, -\sqrt{\hat{\rho}\epsilon})$ *saddle point, and define* $u^+ = \mathrm{Exp}_u(-\eta \mathrm{grad} f(u))$ *and* $w^+ = \mathrm{Exp}_w(-\eta \mathrm{grad} f(w))$*. Under Assumptions 1, 2, 3, if all pairwise distances between* u, w, u^+, w^+, x *are less than* $12\mathscr{S}$*, then for some explicit constant* $C_1(K, \rho, \beta)$ *depending only on* K, ρ, β*, there is*

$$\|\Gamma_{u^+}^{x}\mathrm{Exp}_{u^+}^{-1}(w^+) - (I - \eta H(x))\Gamma_u^x \mathrm{Exp}_u^{-1}(w)\|$$
$$\leq C_1(K, \rho, \beta) d(u, w) \left(d(u, w) + d(u, x) + d(w, x)\right).$$

for some explicit function C_1*.*

This lemma is illustrated in Fig. A.1.

Proof. Denote $-\eta\text{grad}f(u) = v_u$, $-\eta\text{grad}f(w) = v_w$. v is a smooth map. We first prove the following claim.

Claim 1.

$$d(u_+, w_+) \leq c_6(K) d(u, w),$$

where $c_6(K) = c_4(K) + 1 + c_2(K)R^2$.

To show this, note that

$$d(u_+, w_+) \leq d(u_+, \tilde{w}_+) + d(\tilde{w}_+, w_+),$$

and using Lemma 14 with $\tilde{w}_+ = \text{Exp}_w(\Gamma_u^w v_u)$,

$$\begin{aligned} d(\tilde{w}_+, w_+) &= d(\text{Exp}_w(v_w), \text{Exp}_w(\Gamma_u^w v_u)) \\ &\leq (1 + c_2(K)R^2)\|v_w - \Gamma_u^w v_u\| \\ &\leq \beta(1 + c_2(K)R^2) d(u, w). \end{aligned}$$

Using Lemma 14,

$$d(\tilde{w}_+, u_+) \leq c_4(K) d(u, w). \tag{A.4}$$

Adding the two inequalities proves the claim.

We use now Lemma 12 between (u, w, u_+, w_+) in two different ways. First let us use it for $a = \text{Exp}_u^{-1}(w)$ and $y = \Gamma_w^u v_w$. We obtain:

$$d(w_+, \text{Exp}_u(\text{Exp}_u^{-1}(w) + \Gamma_w^u v_w)) \leq c_1(K) d(u, w)(d(u, w)^2 + \|v_w\|^2). \tag{A.5}$$

Then we use it for $a = \mathrm{Exp}_u^{-1}(v_u)$ and $y = \Gamma_{u_+}^u \mathrm{Exp}_{u_+}^{-1}(w_+)$ which yields

$$\begin{aligned}
&d(w_+, \mathrm{Exp}_u(v_u + \Gamma_{u_+}^u \mathrm{Exp}_{u_+}^{-1}(w_+))) \\
\le\ &c_1(K)d(u_+, w_+)(d(u_+, w_+)^2 + \|v_u\|^2) \\
\le\ &c_1(K)c_5(K, \|v_u\|, \|v_w\|)d(u, w) \cdot \Big[c_5(K, \|v_u\|, \|v_w\|)^2 d(u, w)^2 + \|v_u\|^2\Big].
\end{aligned}$$

Using the triangular inequality we have

$$\begin{aligned}
&d(\mathrm{Exp}_u(\mathrm{Exp}_u^{-1}(w) + \Gamma_w^u v_w), \mathrm{Exp}_u(v_u + \Gamma_{u_+}^u \mathrm{Exp}_{u_+}^{-1}(w_+))) \\
\le\ &d(w_+, \mathrm{Exp}_u(\mathrm{Exp}_u^{-1}(w) + \Gamma_w^u v_w)) + d(w_+, \mathrm{Exp}_u(v_u + \Gamma_{u_+}^u \mathrm{Exp}_{u_+}^{-1}(w_+))) \\
\le\ &c_7 d(u, w)
\end{aligned}$$

with c_7 defined as

$$c_7 = c_1(K)c_6(K) \cdot [c_5(K, \|v_u\|, \|v_w\|)^2 d(u, w)^2 + \|v_u\|^2 + \|v_w\|^2\Big].$$

We use again Lemma 13,

$$\|\Gamma_{u_+}^u \mathrm{Exp}_{u_+}^{-1}(w_+)) - \mathrm{Exp}_u^{-1}(w) - [v_u - \Gamma_w^u v_w]\| \le (1 + c_3(K)R^2) \cdot c_7 d(u, w).$$

Therefore we have linearized the iterate in $T_u\mathcal{M}$. We should see how to transport it back to $T_x\mathcal{M}$. With Lemma 15 we have

$$\|[\Gamma_u^x \Gamma_{u_+}^u - \Gamma_{u_+}^x]\mathrm{Exp}_{u_+}^{-1}(w_+))\| = c_5(K)d(u, x)d(u_+, w_+)\|v_u\|.$$

Note v_u and v_w are $-\eta \mathrm{grad} f(u)$ and $-\eta \mathrm{grad} f(w)$, we define $\nabla v(x)$ the gradient of v, i.e.,

$-\eta H$. Using Hessian Lipschitz,

$$\begin{aligned}
&\|v_u - \Gamma_w^u v_w + \eta H(u)\mathrm{Exp}_u^{-1}(w)\| \\
&= \|v_u - \Gamma_w^u v_w - \nabla v(u)\mathrm{Exp}_u^{-1}(w)\| \\
&\le \rho d(u,w)^2,
\end{aligned}$$

and

$$\|\nabla v(u)\mathrm{Exp}_u^{-1}(w) - \Gamma_x^u \nabla v(x)\Gamma_u^x \mathrm{Exp}_u^{-1}(w)\| \le \rho d(u,w)d(u,x).$$

So we have

$$\begin{aligned}
&\|\Gamma_{u_+}^x \mathrm{Exp}_{u_+}^{-1}(w_+) - (I + \nabla v(x))\Gamma_u^x \mathrm{Exp}_u^{-1}(w)\| \\
&\le c_7 d(u,w) + \rho d(u,w)(d(u,w) + d(u,x)) + c_5(K)d(u,x)d(u_+,w_+)\|v_u\| := D_1 \qquad \text{(A.6)}
\end{aligned}$$

□

Evolution of $\mathrm{Exp}_x^{-1}(w) - \mathrm{Exp}_x^{-1}(u)$

We consider now the evolution of $\mathrm{Exp}_x^{-1}(w) - \mathrm{Exp}_x^{-1}(u)$ in the fixed tangent space $\mathcal{T}_x\mathcal{M}$. We show in the following lemma that it also approximately follows a linear iteration.

Lemma 11. *Define* $\gamma = \sqrt{\hat\rho\epsilon}$, $\kappa = \frac{\beta}{\gamma}$, *and* $\mathscr{S} = \sqrt{\eta\beta}\frac{\gamma}{\hat\rho}\log^{-1}(\frac{d\kappa}{\delta})$. *Let us consider* x *be a* $(\epsilon, -\sqrt{\hat\rho\epsilon})$ *saddle point, and define* $u^+ = \mathrm{Exp}_u(-\eta \mathrm{grad} f(u))$ *and* $w^+ = \mathrm{Exp}_w(-\eta \mathrm{grad} f(w))$. *Under Assumptions 1, 2, 3, if all pairwise distances between* u, w, u^+, w^+, x *are less than* $12\mathscr{S}$, *then for some explicit constant* $C(K,\rho,\beta)$ *depending only on* K, ρ, β, *there is*

$$\begin{aligned}
&\|\mathrm{Exp}_x^{-1}(w^+) - \mathrm{Exp}_x^{-1}(u^+) - (I - \eta H(x))(\mathrm{Exp}_x^{-1}(w) - \mathrm{Exp}_x^{-1}(u))\| \qquad \text{(A.7)} \\
&\le C(K,\rho,\beta)d(u,w)\left(d(u,w) + d(u,x) + d(w,x)\right).
\end{aligned}$$

This lemma controls the error of the linear approximation of the iterates hen mapped in $\mathcal{T}_x\mathcal{M}$ and largely follows from Lemma 10.

Proof. We have that

$$w = \mathrm{Exp}_x(\mathrm{Exp}_x^{-1}(w)) \tag{A.8}$$

$$= \mathrm{Exp}_u(\mathrm{Exp}_u^{-1}(w)). \tag{A.9}$$

Use Eq. (A.9), let $a = \mathrm{Exp}_x^{-1}(u)$ and $v = \Gamma_u^x \mathrm{Exp}_u^{-1}(w)$, Lemma 12 suggests that

$$\begin{aligned} & d(\mathrm{Exp}_u(\mathrm{Exp}_u^{-1}(w)), \mathrm{Exp}_x(\mathrm{Exp}_x^{-1}(u) + \Gamma_u^x \mathrm{Exp}_u^{-1}(w))) \\ & \leq c_1(K)\|\mathrm{Exp}_u^{-1}(w)\|(\|\mathrm{Exp}_u^{-1}(w)\| + \|\mathrm{Exp}_x^{-1}(u)\|)^2. \end{aligned}$$

Compare with Eq. (A.8), we have

$$\begin{aligned} & d(\mathrm{Exp}_x(\mathrm{Exp}_x^{-1}(w)), \mathrm{Exp}_x(\mathrm{Exp}_x^{-1}(u) + \Gamma_u^x \mathrm{Exp}_u^{-1}(w))) \\ & \leq c_1(K)\|\mathrm{Exp}_u^{-1}(w)\|(\|\mathrm{Exp}_u^{-1}(w)\| + \|\mathrm{Exp}_x^{-1}(u)\|)^2 \\ & := D. \end{aligned} \tag{A.10}$$

Denote the quantity above by D. Now use Lemma 13

$$\|\mathrm{Exp}_x^{-1}(w) - (\mathrm{Exp}_x^{-1}(u) + \Gamma_u^x \mathrm{Exp}_u^{-1}(w))\| \leq (1 + c_3(K)R^2)D.$$

Analogously

$$\|\mathrm{Exp}_x^{-1}(w_+) - (\mathrm{Exp}_x^{-1}(u_+) + \Gamma_{u_+}^x \mathrm{Exp}_{u_+}^{-1}(w_+))\| \leq (1 + c_3(K)R^2)D_+$$

where

$$D_+ = c_1(K)\|\mathrm{Exp}_{u_+}^{-1}(w_+)\|(\|\mathrm{Exp}_{u_+}^{-1}(w_+)\| + \|\mathrm{Exp}_x^{-1}(u_+)\|)^2 \tag{A.11}$$

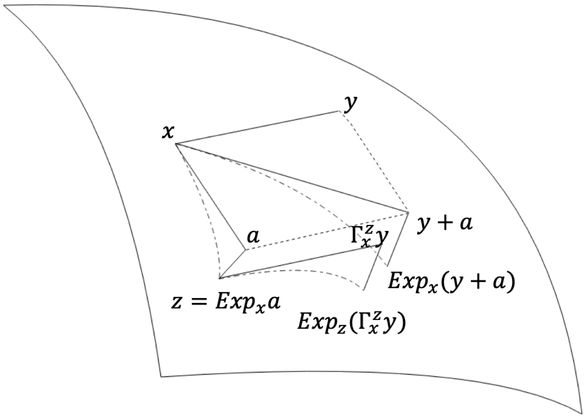


Figure A.2: Lemma 12 bounds the difference of two steps starting from x: (1) take $y + a$ step in $\mathcal{T}_x\mathcal{M}$ and map it to manifold, and (2) take a step in $\mathcal{T}_x\mathcal{M}$, map to manifold, call it z, and take $\Gamma_x^z y$ step in $\mathcal{T}_x\mathcal{M}$, and map to manifold. $\mathrm{Exp}_z(\Gamma_x^z y)$ is close to $\mathrm{Exp}_x(y + a)$.

And we can compare $\Gamma_u^x \mathrm{Exp}_u^{-1}(w)$ and $\Gamma_{u_+}^x \mathrm{Exp}_{u_+}^{-1}(w_+)$ using Eq. (A.6). In the end we have

$$
\begin{aligned}
&\|\mathrm{Exp}_x^{-1}(w^+) - \mathrm{Exp}_x^{-1}(u^+) - (I - \eta H(x))(\mathrm{Exp}_x^{-1}(w) - \mathrm{Exp}_x^{-1}(u))\| \\
\le\ & \|\mathrm{Exp}_x^{-1}(w_+) - (\mathrm{Exp}_x^{-1}(u_+) + \Gamma_{u_+}^x \mathrm{Exp}_{u_+}^{-1}(w_+))\| \\
&+ \|\mathrm{Exp}_x^{-1}(w) - (\mathrm{Exp}_x^{-1}(u) + \Gamma_u^x \mathrm{Exp}_u^{-1}(w))\| \\
&+ \|\Gamma_{u_+}^x \mathrm{Exp}_{u_+}^{-1}(w_+) - \Gamma_u^x \mathrm{Exp}_u^{-1}(w) - \nabla v(x)\Gamma_u^x \mathrm{Exp}_u^{-1}(w)\| \\
&+ \|\nabla v(x)(\Gamma_u^x \mathrm{Exp}_u^{-1}(w) - (\mathrm{Exp}_x^{-1}(w) - \mathrm{Exp}_x^{-1}(u)))\| \\
\le\ & (1 + c_3(K)R^2)(D_+ + D) + D_1 + \eta\|H(x)\|D.
\end{aligned}
$$

D, D_+ and D_1 are defined in Eq. (A.10), Eq. (A.11) and Eq. (A.6), they are all order $d(u,w)\big(d(u,w) + d(u,x) + d(w,x)\big)$ so we get the correct order in Eq. (2.3). □

Control of two-steps iteration

In the following lemma we control the distance between the point obtained after moving along the sum of two vectors in the tangent space, and the point obtained after moving a first time along the first vector and then a second time along the transport of the second vector. This is illustrated in Fig. A.2.

Lemma 12. *Let $x \in \mathcal{M}$ and $y, a \in T_x\mathcal{M}$. Let us denote by $z = \mathrm{Exp}_x(a)$ then under Assumption 3*

$$d(\mathrm{Exp}_x(y+a), \mathrm{Exp}_z(\Gamma_x^z y)) \leq c_1(K) \min\{\|a\|, \|y\|\}(\|a\| + \|y\|)^2. \tag{A.12}$$

This lemma which is crucial in the proofs of Lemma 11 and Lemma 10 tightens the result of Karcher (1977, C2.3), which only shows an upper-bound $O(\|a\|(\|a\| + \|y\|)^2)$.

Proof. We adapt the proof of Karcher (1977, Eq. (C2.3) in App C2.2), the only difference being that we bound more carefully the initial normal component. We restate here the whole proof for completeness.

Let $x \in \mathcal{M}$ and $y, a \in T_x\mathcal{M}$. We denote by $\gamma(t) = \mathrm{Exp}_x(ta)$. We want to compare the point $\mathrm{Exp}_x(r(y+a))$ and $\mathrm{Exp}_\gamma(1)(\Gamma_x^{\gamma(1)y})$. These two points , for a fixed r are joined by the curve

$$t \mapsto c(r,t) = \mathrm{Exp}_{\gamma(t)}(r\Gamma_x^{\gamma(t)}(y + (1-t)a)).$$

We note that $\frac{d}{dt}c(r,t)$ is a Jacobi field along the geodesic $r \mapsto c(r,t)$, which we denote by $J_t(r)$. We importantly remark that the length of the geodesic $r \mapsto c(r,t)$ is bounded as $\|\frac{d}{dr}c(r,t)\| \leq \|y + (1-t)a\|$. We denote this quantity by $\rho_t = \|y + (1-t)a\|$. The initial condition of the Jacobi field J_t are given by:

$$\begin{aligned} J_t(0) &= \frac{d}{dt}\gamma(t) = \Gamma_x^{\gamma(t)}a \\ \frac{D}{dr}J_t(0) &= \frac{D}{dr}\Gamma_x^{\gamma(t)}(y + (1-t)a) = -\Gamma_x^{\gamma(t)}a. \end{aligned}$$

These two vectors are linearly dependent and it is therefore possible to apply Karcher (1977, Proposition A6) to bound J_t^{norm}. Moreover, following Karcher (1977, App A0.3), the tangential component of the Jacobi field is known explicitly, independent of the metric, by

$$J_t^{\tan}(r) = \left(J_t^{\tan}(0) + r\frac{D}{dr}J_t^{\tan}(0)\right)\frac{d}{dr}c(r,t)$$

where the initial conditions of the tangential component of the Jacobi fields are given by $J_t^{\tan}(0) = \langle J_t(0), \frac{\frac{d}{dr}c(r,t)}{\|\frac{d}{dr}c(r,t)\|}\rangle$ and $\frac{D}{dr}J_t^{\tan}(0) = \langle \frac{D}{dr}J_t(0), \frac{\frac{d}{dr}c(r,t)}{\|\frac{d}{dr}c(r,t)\|}\rangle = -J_t^{\tan}(0)$. Therefore

$$J_t^{\tan}(r) = (1-r)J_t^{\tan}(0)\frac{d}{dr}c(r,t),$$

and $J_t^{\tan}(1) = 0$.

We estimate now the distance $d(\mathrm{Exp}_x(y+a), \mathrm{Exp}_z(\Gamma_x^z y))$ by the length of the curve $t \mapsto c(r,t)$ as follows:

$$d(\mathrm{Exp}_x(y+a), \mathrm{Exp}_z(\Gamma_x^z y)) \leq \int_0^1 \|\frac{d}{dt}c(1,t)\|dt = \int_0^1 \|J_t^{\mathrm{norm}}(1)\|dt,$$

where we use crucially that $J_t^{tan}(1) = 0$.

We utilize (Karcher, 1977, Proposition A.6) to bound $\|J_t^{\mathrm{norm}}(1)\|$ as

$$\|J_t^{\mathrm{norm}}(1)\| \leq \|J_t^{\mathrm{norm}}(0)\|(\cosh(\sqrt{K}\rho_t) - \frac{\sinh(\sqrt{K}\rho_t)}{\sqrt{K}\rho_t})$$

using (Karcher, 1977, Equation (A6.3)) with $\kappa = 0$, $f_\kappa(1) = 0$ and recalling that the geodesics $r \mapsto c(r,t)$ have length ρ_t.

In particular for small value $\|a\| + \|y\|$ we have for some constant $c_1(K)$,

$$\|J_t^{\mathrm{norm}}(1)\| \leq \|J_t^{\mathrm{norm}}(0)\|c_1(K)\rho_t^2.$$

We bound $\|J_t^{\mathrm{norm}}(0)\|$ now. This is the main difference with the original proof of Karcher (1977) who directly bounded $\|J_t^{\mathrm{norm}}(0)\| \leq \|J_t(0)\| = \|a\|$ and $\rho_t \leq \|a\| + \|y\|$. Therefore his proof does not lead to the correct dependence in $\|y\|$.

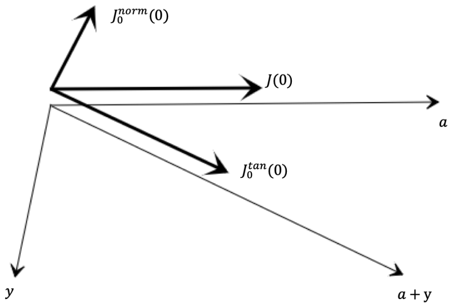


Figure A.3: Figure for Lemma 12.

We have $J_t^0 = \Gamma_x^{\gamma(t)} a$, and the tangential component (velocity of $r \to c(r,t)$) is in the $\Gamma_x^{\gamma(t)}(y+(1-t)a)$ direction. Let $\tilde{z} = \Gamma_x^{\gamma(t)}(y+(1-t)a)$ and $\mathcal{P}_{\tilde{z}^\perp}$ and $\mathcal{P}_{a^\perp}$ denote the projection onto orthogonal complement of $\tilde{z}$ and a.

$$\begin{aligned}
\|J_t^{\text{norm}}(0)\|^2 &= \|\mathcal{P}_{\tilde{z}^\perp}(a)\|^2 \\
&= \|a\|^2 - \frac{(a^T\tilde{z})^2}{\|\tilde{z}\|^2} \\
&= \frac{\|a\|^2}{\|\tilde{z}\|^2}\left(\|\tilde{z}\|^2 - \frac{(a^T\tilde{z})^2}{\|\tilde{z}\|^2}\right) \\
&\le \frac{\|a\|^2}{\|\tilde{z}\|^2}\|\mathcal{P}_{a^\perp}(\Gamma_x^{\gamma(t)}(y+(1-t)a))\|^2 \\
&\le \frac{\|a\|^2}{\|\tilde{z}\|^2}\|\mathcal{P}_{a^\perp}(\Gamma_x^{\gamma(t)}((1-t)a)) + \mathcal{P}_{a^\perp}(\Gamma_x^{\gamma(t)}y)\|^2 \\
&= \frac{\|a\|^2}{\|\tilde{z}\|^2}\|\mathcal{P}_{a^\perp}(\Gamma_x^{\gamma(t)}y)\|^2 \\
&\le \frac{\|a\|^2\|y\|^2}{\|\tilde{z}\|^2}.
\end{aligned}$$

So

$$\begin{aligned}
\|J_t^{\text{norm}}(1)\| &\le \|J_t^{\text{norm}}(0)\|c_1(K)\rho_t^2 \\
&\le \frac{\|a\|\cdot\|y\|}{\|\tilde{z}\|}c_1(K)\|\tilde{z}\|^2 \\
&\le c_1(K)\|a\|\cdot\|y\|(\|a\|+\|y\|),
\end{aligned}$$

and

$$d(\mathrm{Exp}_x(y+a), \mathrm{Exp}_z(\Gamma_x^z y)) \leq c_1(K)\|a\| \cdot \|y\|(\|a\| + \|y\|).$$

□

A.3 Auxilliary lemmas

In the proofs of Lemma 10 and Lemma 11 we needed numerous auxiliary lemmas we are stating here.

We needed the following lemma which shows that both the exponential map and its inverse are Lipschitz.

Lemma 13. *Let $x, y, z \in M$, and the distance of each two points is no bigger than R. Then under Assumption 3*

$$(1 + c_2(K)R^2)^{-1} d(y,z) \leq \|\mathrm{Exp}_x^{-1}(y) - \mathrm{Exp}_x^{-1}(z)\| \leq (1 + c_3(K)R^2) d(y,z).$$

Intuitively this lemma relates the norm of the difference of two vectors of $\mathcal{T}_x\mathcal{M}$ to the distance between the corresponding points on the manifold $\mathcal{M}$ and follows from bounds on the Hessian of the square-distance function (Sakai, 1996, Ex. 4 p. 154).

Proof. The upper-bound is directly proven in Karcher (1977, Proof of Cor. 1.6), and we prove the lower-bound via Lemma 12. Let $b = \mathrm{Exp}_y(\Gamma_x^y(\mathrm{Exp}_x^{-1}(z) - \mathrm{Exp}_x^{-1}(y)))$. Using $d(y,b) = \|\mathrm{Exp}_y^{-1}(b)\|$ and Lemma 12,

$$\begin{aligned} d(y,z) &\leq d(y,b) + d(b, \mathrm{Exp}_x(\mathrm{Exp}_x^{-1}(z))) \\ &\leq \|\mathrm{Exp}_x^{-1}(y) - \mathrm{Exp}_x^{-1}(z)\| \\ &\quad + c_1(K)\|\mathrm{Exp}_x^{-1}(y) - \mathrm{Exp}_x^{-1}(z)\|(\|\mathrm{Exp}_x^{-1}(y) - \mathrm{Exp}_x^{-1}(z)\| + \|\mathrm{Exp}_x^{-1}(y)\|)^2 \end{aligned}$$

□

The following contraction result is fairly classical and is proven using the Rauch comparison theorem from differential geometry (Cheeger & Ebin, 2008).

Lemma 14. *(Mangoubi et al., 2018, Lemma 1) Under Assumption 3, for $x, y \in \mathcal{M}$ and $w \in T_x\mathcal{M}$,*

$$d(\mathrm{Exp}_x(w), \mathrm{Exp}_y(\Gamma_x^y w)) \leq c_4(K) d(x, y).$$

Eventually we need the following corollary of the famous Ambrose-Singer holonomy theorem (Ambrose & Singer, 1953).

Lemma 15. *(Karcher, 1977, Section 6) Under Assumption 3, for $x, y, z \in \mathcal{M}$ and $w \in T_x\mathcal{M}$,*

$$\|\Gamma_y^z \Gamma_x^y w - \Gamma_x^z w\| \leq c_5(K) d(x, y) d(y, z) \|w\|.$$

A.4 Proof of Lemma 7 and 8

In this section we prove two important lemmas from which the proof of our main result mainly comes out. Then we show, in the last subsection, how to combine them to prove this main result.

Proof of Lemma 7

Suppose $f(u_{t+1}) - f(u_t) \leq -\frac{\eta}{2}\|\text{grad} f(u_t)\|^2$.

$$\begin{aligned}
d(u_{\hat{c}\mathscr{T}}, u_0)^2 &\leq (\sum_{0}^{\hat{c}\mathscr{T}-1} d(u_{t+1}, u_t))^2 \\
&\leq \hat{c}\mathscr{T} \sum_{0}^{\hat{c}\mathscr{T}-1} d(u_{t+1}, u_t)^2 \\
&\leq \eta^2 \hat{c}\mathscr{T} \sum_{0}^{\hat{c}\mathscr{T}-1} \|\text{grad} f(u_t)\|^2 \\
&\leq 2\eta \hat{c}\mathscr{T} \sum_{0}^{\hat{c}\mathscr{T}-1} f(u_t) - f(u_{t+1}) \\
&= 2\eta \hat{c}\mathscr{T}(f(u_0) - f(u_{\hat{c}\mathscr{T}})) \\
&\leq 6\eta \hat{c}\mathscr{T}\mathscr{F} = 6\hat{c}\mathscr{S}^2.
\end{aligned}$$

Proof of Lemma 8

Note that, for any points inside a region with diameter R, under the assumption of Lemma 7, we have $\max\{c_2(K), c_3(K)\}R^2 \leq 1/2$.

Define $v_t = \text{Exp}_{\tilde{x}}^{-1}(w_t) - \text{Exp}_{\tilde{x}}^{-1}(u_t)$, let $v_0 = e_1$ be the smallest eigenvector of $H(\tilde{x})$, then let $\hat{y}_{2,t}$ be a unit vector, we have

$$v_{t+1} = (I - \eta H(\tilde{x}))v_t + C(K, \rho, \beta)d(u_t, w_t) \cdot (d(u_t, \tilde{x}) + d(w_t, \tilde{x}) + d(\tilde{x}, u_0))\hat{y}_{2,t}. \tag{A.15}$$

Let $C := C(K, \rho, \beta)$. Suppose Lemma 7 is false, then $0 \leq t \leq T$, $d(u_t, \tilde{x}) \leq 3\hat{c}\mathscr{S}$, $d(w_t, \tilde{x}) \leq 3\hat{c}\mathscr{S}$, so $d(u_t, w_t) \leq 6\hat{c}\mathscr{S}$, and the norm of the last term in Eq. (A.15) is smaller than $14\eta C\hat{c}\mathscr{S}\|v_t\|$.

Lemma 4 indicates that

$$\|v_t\| \in [1/2, 2] \cdot d(u_t, w_t) = [3/2, 6] \cdot \hat{c}\mathscr{S}. \tag{A.16}$$

Let ψ_t be the norm of v_t projected onto e_1, the smallest eigenvector of $H(0)$, and φ_t be the norm of v_t projected onto the remaining subspace. Then Eq. (A.15) is

$$\psi_{t+1} \geq (1+\eta\gamma)\psi_t - \mu\sqrt{\psi_t^2+\phi_t^2},$$
$$\phi_{t+1} \leq (1+\eta\gamma)\phi_t + \mu\sqrt{\psi_t^2+\phi_t^2}.$$

Prove that for all $t \leq T$, $\phi_t \leq 4\mu t\psi_t$. Assume it is true for t, we have

$$4\mu(t+1)\psi_{t+1} \geq 4\mu(t+1)\cdot\left((1+\eta\gamma)\psi_t - \mu\sqrt{\psi_t^2+\phi_t^2}\right),$$
$$\phi_{t+1} \leq 4\mu t(1+\eta\gamma)\phi_t + \mu\sqrt{\psi_t^2+\phi_t^2}.$$

So we only need to show that

$$(1+4\mu(t+1))\sqrt{\psi_t^2+\phi_t^2} \leq (1+\eta\gamma)\psi_t.$$

By choosing $\sqrt{c_{\max}} \leq \frac{1}{56\hat{c}^2}$ and $\eta \leq c_{\max}/\beta$, we have

$$4\mu(t+1) \leq 4\mu T \leq 4\eta C\mathscr{S}\cdot 14\hat{c}^2\mathscr{T} = 56\hat{c}^2\frac{C}{\hat{\rho}}\sqrt{\eta\beta} \leq 1.$$

This gives

$$4(1+\eta\gamma)\psi_t \geq 2\sqrt{2\psi_t^2} \geq (1+4\mu(t+1))\sqrt{\psi_t^2+\phi_t^2}.$$

Now we know $\phi_t \leq 4\mu t\psi_t \leq \psi_t$, so $\psi_{t+1} \geq (1+\eta\gamma)\psi_t - \sqrt{2}\mu\psi_t$, and

$$\mu = 14\hat{c}\eta C\mathscr{S} \leq 14\hat{c}\sqrt{c_{\max}}\eta\gamma C\log^{-1}(\frac{d\kappa}{\delta})/\hat{\rho} \leq \eta\gamma/2,$$

so $\psi_{t+1} \geq (1+\eta\gamma/2)\psi_t$.

We also know that $\|v_t\| \leq 6\hat{c}\mathscr{S}$ for all $t \leq T$ from Eq. (A.16), so

$$\begin{aligned} 6\hat{c}\mathscr{S} &\geq \|v_t\| \geq \psi_t \geq (1+\eta\gamma/2)^t \psi_0 \\ &= (1+\eta\gamma/2)^t \frac{\mathscr{S}}{\kappa} \log^{-1}(\frac{d\kappa}{\delta}) \\ &\geq (1+\eta\gamma/2)^t \frac{\delta\mathscr{S}}{2\sqrt{d}\kappa} \log^{-1}(\frac{d\kappa}{\delta}). \end{aligned}$$

This implies

$$\begin{aligned} T &< \frac{\log(12\frac{\kappa\sqrt{d}}{\delta}\hat{c}\log(\frac{d\kappa}{\delta}))}{2\log(1+\eta\gamma/2)} \\ &\leq \frac{\log(12\frac{\kappa\sqrt{d}}{\delta}\hat{c}\log(\frac{d\kappa}{\delta}))}{\eta\gamma} \\ &\leq (2+\log(12\hat{c}))\mathscr{T}. \end{aligned}$$

By choosing $\hat{c}$ such that $2+\log(12\hat{c}) < \hat{c}$, we have $T \leq \hat{c}\mathscr{T}$, which finishes the proof.

Proof of function value decrease at an approximate saddle point

With Lemma 7 and 7 proved, we can lower bound the function value in $O(\mathscr{T})$ iterations decrease by $\Omega(\mathscr{F})$, thus match the convergence rate in the main theorem. Let $T' := \inf_t \left\{t|\tilde{f}_{u_0}(u_t) - f(u_0) \leq -3\mathscr{F}\right\}$. Let $\check{}$ denote the operator $\mathrm{Exp}_{u_0}^{-1}(\cdot)$. If $T' \leq T$,

$$\begin{aligned} & f(u_{T'}) - f(u_0) \\ \leq\ & \nabla f(u_0)^T(u_{T'} - u_0) + \frac{1}{2}H(u_0)[\check{u}_{T'} - u_0, \check{u}_{T'} - u_0] + \frac{\rho}{6}\|\check{u}_{T'} - u_0\|^3 \\ \leq\ & \tilde{f}_{u_0}(u_t) - f(u_0) + \frac{\rho}{2}d(u_0,\tilde{x})\|\check{u}_{T'} - u_0\|^2 \\ \leq\ & -3\mathscr{F} + O(\rho\mathscr{S}^3) \leq -2.5\mathscr{F}. \end{aligned}$$

If $T' > T$, then $\inf_t \left\{t|\tilde{f}_{w_0}(w_t) - f(w_0) \leq -3\mathscr{F}\right\} \leq T$, and we know $f(w_T) - f(w_0) \leq -2.5\mathscr{F}$.

Remark 5. *What is left is bounding the volume of the stuck region, to get the probability of getting out of the stuck region by the perturbation. The procedure is the same as in Jin et al. (2017a). We sample from a unit ball in $\mathcal{T}_x\mathcal{M}$, where x is the approximate saddle point. In Lemma 7 and 7, we study the inverse exponential map at the approximate saddle point x, and the coupling difference between $\mathrm{Exp}_x^{-1}(w)$ and $\mathrm{Exp}_x^{-1}(u)$. The iterates we study and the noise are all in the tangent space $\mathcal{T}_x\mathcal{M}$ which is a Euclidean space, so the probability bound is same as the one in Jin et al. (2017a).*

Appendix B

APPENDIX OF CHAPTER 3

B.1 Proof of the main theorems

Theorem 12. *Suppose assumptions 4,5 hold, and consider the two problems* (3.9) *and* (3.10). *Let K^* denote the global minimizer of $\mathcal{L}(K)$ in S_K. Then there exist constants $C_1, C_2 > 0$ independent of the suboptimality $\mathcal{L}(K) - \mathcal{L}(K^*)$, and a direction V, with $\|V\|_F = 1$, in the descent cone of $\mathcal{S}_K$ at K such that,*

1. if f is convex, the gradient of $\mathcal{L}$ satisfies[1]

$$\nabla\mathcal{L}(K)[V] \leq -C_1(\mathcal{L}(K) - \mathcal{L}(K^*)). \tag{B.1}$$

2. if f is μ-strongly convex, the gradient satisfies

$$\nabla\mathcal{L}(K)[V] \leq -C_2(\mu(\mathcal{L}(K) - \mathcal{L}(K^*)))^{1/2}. \tag{B.2}$$

where (the constants can be bounded with simple constraints bounding norms of L, P or K)

$$\begin{aligned} C_1 &= (2\max\{\|L - L^*\|_F\sigma_{\min}^{-1}(P), \|P - P^*\|_F\sigma_{\min}^{-2}(P)\sigma_{\max}(L)\})^{-1}, \\ C_2 &= (2\max\{\sigma_{\min}^{-1}(P), \sigma_{\min}^{-2}(P)\sigma_{\max}(L)\})^{-1}. \end{aligned}$$

Proof. Let $f(x)$ be any convex function. Denote $\mathcal{P}_{\mathcal{S}}(\nabla f(x))$ as the projection of $\nabla f(x)$ onto the descent cone of $\mathcal{S}$ at x, and we know $\|\mathcal{P}_{\mathcal{S}}(\nabla f(x))\| \geq \nabla f(x)[\frac{\Delta}{\|\Delta\|}]$ for any $-\Delta$ in the descent cone of $\mathcal{S}$ at x. We will find the direction Δ and bound the directional deravative. First, for any convex function $f(x)$, let the minimum be x^*, and $x - x^* = \Delta$. Let $\nabla f(x) = g$.

[1] We always consider the directional derivative of a feasible direction within descent cone.

For any non-stationary point, $f(x) \leq f(x^*) + g^\top \Delta$. Since $\mathcal{S}$ is a convex set, $-\Delta$ belongs to the descent cone of $\mathcal{S}$ at x, so the direction $-\frac{\Delta}{\|\Delta\|}$ is feasible, $f(x) - f(x - t\frac{\Delta}{\|\Delta\|}) \leq tg^\top \frac{\Delta}{\|\Delta\|}$ when $t \to_+ 0$, so that $f(x)[\frac{\Delta}{\|\Delta\|}] = g^\top \frac{\Delta}{\|\Delta\|} \geq \frac{f(x)-f(x^*)}{\|x-x^*\|}$. We will apply the inequality for $f(L, P, Z)$.

Let K^* be the optimal K and (L^*, P^*, Z^*) be the optimal point in the parameterized space. We have $\mathcal{L}(K^*) = f(L^*, P^*, Z^*)$.

We denote $\mathscr{Z}(L, P) \in \text{argmin}_Z f(L, P, Z)$ subject to $(L, P, Z) \in \mathcal{S}$ (if there are multiple minimizers we pick any one). With either Assumption 6 or 5, we can define the mapping from K to (L, P, Z) respectively in one of the following ways:

1. (Assumption 6) let K map to (L, P) with $K = LP^{-1}$ and $Z = \mathscr{Z}(L, P)$.
2. (Assumption 5) let

$$\begin{aligned} &(L, P, Z) = \text{argmin}_{L',P',Z'}\ f(L', P', Z') \\ &\text{s.t. } (L', P', Z') \in \mathcal{S},\ P' \succ 0,\ L'P'^{-1} = K. \end{aligned}$$

Note f is convex, so

$$\begin{aligned} &\nabla f(L, P, Z)[(L, P, Z) - (L^*, P^*, Z^*)] \\ &\geq f(L, P, Z) - f(L^*, P^*, Z^*) \\ &= f(L, P, \mathscr{Z}(L, P)) - f(L^*, P^*, \mathscr{Z}(L^*, P^*)) \\ &= \mathcal{L}(K) - \mathcal{L}(K^*). \end{aligned} \tag{B.3}$$

Now we consider the directional derivative in K space. By definition,

$$\nabla \mathcal{L}(K)[V] = \lim_{t \to 0^+} (\mathcal{L}(K + tV) - \mathcal{L}(K))/t.$$

Let $\Delta L = L^* - L$, $\Delta P = P^* - P$, and $V = \Delta L P^{-1} - LP^{-1}\Delta P P^{-1}$. Then

$$\begin{aligned}\nabla\mathcal{L}(K)[V] &= \lim_{t\to 0^+}(\mathcal{L}(K+tV)-\mathcal{L}(K))/t\\ &= \lim_{t\to 0^+}(\mathcal{L}(LP^{-1}+t(\Delta LP^{-1}-LP^{-1}\Delta PP^{-1}))-\mathcal{L}(LP^{-1}))/t\\ &= \lim_{t\to 0^+}(\mathcal{L}((L+t\Delta L)(P+t\Delta P)^{-1})-\mathcal{L}(LP^{-1}))/t.\end{aligned}$$

The last line uses $(P+t\Delta P)^{-1} = P^{-1} - tP^{-1}\Delta PP^{-1} + o(t)$. Denote $\Delta(L,P,Z) = (L^*,P^*,Z^*) - (L,P,Z)$, $\Delta(L,P,Z)$ is in the descent cone of $\mathcal{S}$ at (L,P,Z) due to the convexity of $\mathcal{S}$. With Assumption 6, we continue with

$$\begin{aligned}\nabla\mathcal{L}(K)[V] &= \lim_{t\to 0^+}(f(L+t\Delta L, P+t\Delta P, \mathscr{Z}(L+t\Delta L, P+t\Delta P)) - f(L,P,\mathscr{Z}(L,P)))/t\\ &\le \lim_{t\to 0^+}(f(L+t\Delta L, P+t\Delta P, \mathscr{Z}(L,P)+t\Delta Z) - f(L,P,\mathscr{Z}(L,P)))/t\\ &= \nabla f(L,P,Z)[\Delta(L,P,Z)].\end{aligned}$$

With Assumption 5, we continue with

$$\begin{aligned}\nabla\mathcal{L}(K)[V] = \lim_{t\to 0^+}\min_{L',P',Z'}\ & f(L',P',Z') - f(L,P,Z)\\ \text{s.t. }& (L',P',Z')\in\mathcal{S},\ \ P'\succ 0,\\ & L'P'^{-1} = (L+t\Delta L)(P+t\Delta P)^{-1}.\end{aligned}$$

$(L+t\Delta L, P+t\Delta P, \mathscr{Z}(L,P)+t\Delta Z)$ is a feasible point of the optimization problem, thus is less than or equal to the minimum, and then

$$\begin{aligned}\nabla\mathcal{L}(K)[V] &\le \lim_{t\to 0^+}(f(L+t\Delta L, P+t\Delta P, \mathscr{Z}(L,P)+t\Delta Z) - f(L,P,\mathscr{Z}(L,P)))/t\\ &= \nabla f(L,P,Z)[\Delta(L,P,Z)].\end{aligned}$$

So the following inequality holds.

$$\begin{aligned}\nabla\mathcal{L}(K)[V] &\leq \nabla f(L,P,Z)[\Delta(L,P,Z)]\\ &\leq -(f(L,P,Z)-f(L^*,P^*,Z^*)) = -(\mathcal{L}(K)-\mathcal{L}(K^*)) < 0.\end{aligned}$$

After normalization, we have

$$\nabla\mathcal{L}(K)[\frac{V}{\|V\|_F}] \geq \frac{1}{\|V\|_F}(\mathcal{L}(K)-\mathcal{L}(K^*)). \tag{B.4}$$

With $V = \Delta LP^{-1} - LP^{-1}\Delta PP^{-1}$, we can get $\|V\|_F \leq 1/C_1$.

If $f(L,P,Z)$ is μ strongly convex, then we can restrict f in the line segment $(L,P,Z) - (L^*,P^*,Z^*)$ and get

$$\begin{aligned}(\frac{\nabla\mathcal{L}(K)[V]}{\|V\|_F})^2 &\geq \frac{1}{\|V\|_F^2}(\nabla f(L,P,Z)[\Delta(L,P,Z)])^2\\ &\geq \frac{\mu\|\Delta(L,P,Z)\|_F^2}{\|V\|_F^2}\cdot(f(L,P,Z)-f(L^*,P^*,Z^*))\\ &= \frac{\mu(\|L^*-L\|_F^2+\|P^*-P\|_F^2+\|Z^*-Z\|_F^2)}{\|(L^*-L)P^{-1}-LP^{-1}(P^*-P)P^{-1}\|_F^2}\cdot(f(L,P,Z)-f(L^*,P^*,Z^*))\\ &\geq \frac{\mu(\|L^*-L\|_F^2+\|P^*-P\|_F^2)}{\|(L^*-L)P^{-1}-LP^{-1}(P^*-P)P^{-1}\|_F^2}\cdot(f(L,P,Z)-f(L^*,P^*,Z^*))\\ &\geq \frac{\mu(f(L,P,Z)-f(L^*,P^*,Z^*))}{(2\max\{\sigma_{\min}^{-1}(P),\sigma_{\min}^{-2}(P)\sigma_{\max}(L)\})^2}.\end{aligned}$$

Now we will prove with the following assumption that is weaker than μ strong convexity: let $\mathcal{P}_{\mathcal{S}}(-\nabla f(L,P,Z))$ be the projection of $-\nabla f(L,P,Z)$ in the descent cone of $\mathcal{S}$ at (L,P,Z), if for any

$$(L,P,Z) = \arg\min_{L',P',Z'} f(L',P',Z'),\ \text{s.t.}\ (L',P',Z')\in\mathcal{S},\ L'(P')^{-1}=K,$$

we have $\|\mathcal{P}_{\mathcal{S}}(-\nabla f(L,P,Z))\|_F^2 \geq \mu(f(L,P,Z)-f(L^*,P^*,Z^*))$.

Now we denote

$$\Delta(L,P,Z) = (\Delta L, \Delta P, \Delta Z) = \frac{\mathcal{P}_{\mathcal{S}}(-\nabla f(L,P,Z))}{\|\mathcal{P}_{\mathcal{S}}(-\nabla f(L,P,Z))\|}$$

and $V = \Delta L P^{-1} - LP^{-1}\Delta P P^{-1}$. The proof is similar to strongly convex case:

$$\begin{aligned}
(\frac{\nabla\mathcal{L}(K)[V]}{\|V\|_F})^2 &\geq \frac{1}{\|V\|_F^2}(\nabla f(L,P,Z)[\Delta(L,P,Z)])^2 \\
&\geq \frac{\mu\|\Delta(L,P,Z)\|_F^2}{\|V\|_F^2}\cdot(f(L,P,Z)-f(L^*,P^*,Z^*)) \\
&= \frac{\mu(\|\Delta L\|_F^2+\|\Delta P\|_F^2+\|\Delta Z\|_F^2)}{\|(\Delta L)P^{-1}-LP^{-1}(\Delta P)P^{-1}\|_F^2}\cdot(f(L,P,Z)-f(L^*,P^*,Z^*)) \\
&\geq \frac{\mu(\|\Delta L\|_F^2+\|\Delta P\|_F^2)}{\|(\Delta L)P^{-1}-LP^{-1}(\Delta P)P^{-1}\|_F^2}\cdot(f(L,P,Z)-f(L^*,P^*,Z^*)) \\
&\geq \frac{\mu(f(L,P,Z)-f(L^*,P^*,Z^*))}{(2\max\{\sigma_{\min}^{-1}(P),\sigma_{\min}^{-2}(P)\sigma_{\max}(L)\})^2}.
\end{aligned}$$

And we get the same gradient dominance parameter as strongly convex case.

□

Theorem 3. *Denote* $\Delta K = \Psi(P)[P^*-P]$. *Let* $\nabla\mathcal{L}(K)[\Delta K]$ *be the directional derivative of* $\mathcal{L}(K)$ *in direction* ΔK. *Then with Assumptions 7, 8 we have*

$$\nabla\mathcal{L}(K)[\Delta K] \leq \mathcal{L}(K^*) - \mathcal{L}(K).$$

Proof. Suppose $f(P)$ is convex in P, and the optimizer of (3.27) is P^*. Denote

$$P = \text{argmin}_{P'}\ f(P'),\ \text{s.t. } P' \in \mathcal{S},\ K = \Phi(P'),$$

and

$$\Delta P = P^* - P,\ \Delta K = \Psi(P)[\Delta P].$$

We take the directional derivative and get (explanation of key steps below the last line)

$$\begin{aligned}
\nabla\mathcal{L}(K)[\Delta K] &= \lim_{t\to 0^+} \frac{\mathcal{L}(K+t\Delta K)-\mathcal{L}(K)}{t} \\
&= \lim_{t\to 0^+} \frac{\mathcal{L}(K+t\Psi(P)[\Delta P])-f(P)}{t} && \text{(B.5)} \\
&= \lim_{t\to 0^+} \frac{\mathcal{L}(\Phi(P)+t\Psi(P)[\Delta P])-f(P)}{t} && \text{(B.6)} \\
&= \lim_{t\to 0^+} \frac{\mathcal{L}(\Phi(P+t\Delta P)-o(t))-f(P)}{t} && \text{(B.7)} \\
&= \lim_{t\to 0^+} \frac{\mathcal{L}(\Phi(P+t\Delta P))-f(P)}{t} \\
&= \lim_{t\to 0^+} \frac{\min_{P'\in\mathcal{S},\ \Phi(P+t\Delta P)=\Phi(P')}\ f(P')-f(P)}{t} && \text{(B.8)} \\
&= \lim_{t\to 0^+} \frac{\min_{P'\in\mathcal{S},\ \Phi(P+t\Delta P)=\Phi(P')}\ f(P')-f(P+t\Delta P)+f(P+t\Delta P)-f(P)}{t} \\
&= \lim_{t\to 0^+} \frac{\min_{P'\in\mathcal{S},\ \Phi(P+t\Delta P)=\Phi(P')}\ f(P')-f(P+t\Delta P)}{t} + \nabla f(P)[\Delta P]. && \text{(B.9)}
\end{aligned}$$

(B.5) and (B.6) replace ΔK and K with expressions in P and ΔP. (B.7) applies the Taylor expansion of Φ:

$$\Phi(P+t\Delta P)-(\Phi(P)+t\Psi(P)[\Delta P]) = o(t).$$

(B.8) applies Assumption 8, and we plug in $K=\Phi(P+t\Delta P)$. (B.9) applied the definition of directional derivative

$$\nabla f(P)[\Delta P] = \lim_{t\to 0^+} \frac{f(P+t\Delta P)-f(P)}{t}.$$

Now we bound the first term of (B.9). Note that, since $P+t\Delta P$ for $t>0$ and $t\to 0^+$ belongs to the line segment from P to P^*. Since $\mathcal{S}$ is a convex set, we know that the line segment between to feasible points P^* and P is in $\mathcal{S}$. then

$$P+t\Delta P \in \left\{P' \mid P'\in\mathcal{S},\ \Phi(P+t\Delta P)=\Phi(P')\right\},$$

so that $f(P + t\Delta P)$ is no less than the minimum of the optimization problem (3.28),

$$\lim_{t\to 0^+} \frac{\min_{P'\in\mathcal{S},\ \Phi(P+t\Delta P)=\Phi(P')}\ f(P') - f(P+t\Delta P)}{t} \le 0.$$

$\nabla f(P)[\Delta P]$ is the directional derivative of $f(P)$ in the direction of $P^* - P$, for a convex function f, if P is not an optimizer, $\nabla f(P)[\Delta P]$ is upper bounded by $f(P^*) - f(P) = \mathcal{L}(K^*) - \mathcal{L}(K) < 0$. □

B.2 Constants for continuous time LQR

Theorem 12 asks for two constants C_1, C_2. They are bounded differently for different examples. As an instance, we will calculate the constants for continuous time LQR, quoted from (Mohammadi et al., 2019b, Appendix B). First $P \succ 0$, so we replace singular value by eigenvalue with P,

$$\begin{aligned} C_1 &= (2\max\{\|L - L^*\|_F \lambda_{\min}^{-1}(P), \|P - P^*\|_F \lambda_{\min}^{-2}(P)\sigma_{\max}(L)\})^{-1}, \\ C_2 &= (2\max\{\lambda_{\min}^{-1}(P), \lambda_{\min}^{-2}(P)\sigma_{\max}(L)\})^{-1}. \end{aligned}$$

We need upper bounds for P, L and a lower bound for $\lambda_{\min}(P)$ to guarantee C_1, C_2 being finite. We will show the bounds within the sublevel set that $\{K : \mathcal{L}(K) \le a\}$. Since we can randomly initialize a feasible K_0 and run (projected) gradient descent method with respect to K, if $\mathcal{L}(K)$ is gradient dominant, it is reasonable to assume that during all iterations of the optimization algorithm, the function value is always upper bounded by $\mathcal{L}(K_0)$, or some values not too larger than $\mathcal{L}(K_0)$. So our derivation with a finite sublevel set is reasonable.

Suppose the matrices $Q, R \succ 0$, and we consider the sublevel set when $\mathcal{L}(K) \le a$. The sublevel set gives $\mathbf{tr}(QP) + \mathbf{tr}(LP^{-1}L^\top R) \leq a$, so

$$\begin{aligned}
\lambda_{\min}(R)\lambda_{\max}^{-1}(P)\|L\|_F^2 &\leq \lambda_{\min}(R)\|LP^{-1/2}\|_F^2 \\
&\leq \mathbf{tr}(LP^{-1}L^\top R) \\
&\leq \mathbf{tr}(QP) + \mathbf{tr}(LP^{-1}L^\top R) \leq a.
\end{aligned}$$

So $\|L\|_F \leq a(\lambda_{\max}(P)\lambda_{\min}^{-1}(R))^{1/2}$, and we know from (Mohammadi et al., 2019b, eq(34)) $\mathbf{tr}(P) \leq a\lambda_{\min}^{-1}(Q)$. So we can bound P, L

$$\begin{aligned}
&\mathbf{tr}(P) \leq a\lambda_{\min}^{-1}(Q), \\
&\|L\|_F \leq a(\lambda_{\min}(Q)\lambda_{\min}(R))^{-1/2}.
\end{aligned}$$

Define

$$\nu = \frac{\lambda_{\min}^2(\Sigma)}{4}\left(\sigma_{\max}(A)\lambda_{\min}^{-1/2}(Q) + \sigma_{\max}(B)\lambda_{\min}^{-1/2}(R)\right)^{-2}$$

(Zare et al., 2019, eq(38,40)) suggests $\lambda_{\min}(P) \geq \nu/a$. In summary, we upper bounded L, and upper and lower bounded P in the sublevel set $\mathcal{L}(K) \leq a$, and those bounds are also true for L^*, P^*. We can complete the calculation by inserting the bounds into C_1.

$$\begin{aligned}
C_1 &= (2\max\{\|L - L^*\|_F\lambda_{\min}^{-1}(P), \|P - P^*\|_F\lambda_{\min}^{-2}(P)\sigma_{\max}(L)\})^{-1} \\
&\geq \frac{\nu\lambda_{\min}^{1/2}(Q)\lambda_{\min}^{1/2}(R)}{4a^4} \cdot \min\left\{a^2,\ \nu\lambda_{\min}(Q)\right\}.
\end{aligned}$$

C_2 is calculated similarly with upper bound on P, L, P^{-1}.

$$\begin{aligned}
C_2 &= (2\max\{\lambda_{\min}^{-1}(P), \lambda_{\min}^{-2}(P)\sigma_{\max}(L)\})^{-1} \\
&\geq \frac{\nu}{2a^3}\min\left\{a^2,\ \nu\lambda_{\min}^{1/2}(Q)\lambda_{\min}^{1/2}(R)\right\}.
\end{aligned}$$

B.2.1 Strongly convex parameter of continuous time LQR

In our previous convex formulation of continuous time LQR (3.8), we translate the objective function as a linear function in the new variables L, P, Z. The problem (3.8) can be slightly reformulated as

$$\min_{L,P} \; f(L,P) := \mathbf{tr}(QP) + \mathbf{tr}(LP^{-1}L^\top R), \tag{B.10a}$$

$$\text{s.t. } \mathcal{A}(P) + \mathcal{B}(L) + \Sigma = 0, \; P \succ 0. \tag{B.10b}$$

Compared with (3.8), (B.10) does not contain the variable Z. Below, we will prove that the new objective function $f(L,P)$, restricted within the feasible set, is a strongly convex function, which is not the case for the linear objective (3.8). In Theorem 2, there is another result with strongly convex f and the gradient domminance parameter depends on the strongly convex parameter μ. We also calculate μ of $f(L,P)$ below.

Lemma 16. *Define a sublevel set of of f at level a, consisting of all L, P such that $f(L,P) \leq a$. Define*

$$\begin{aligned}
\nu &= \frac{\lambda_{\min}^2(\Sigma)}{4}\left(\sigma_{\max}(A)\lambda_{\min}^{-1/2}(Q) + \sigma_{\max}(B)\lambda_{\min}^{-1/2}(R)\right)^{-2},\\
\eta &= \|\mathcal{B}\|\left(\nu^{1/2}\lambda_{\min}(\Sigma)\lambda_{\min}(Q)\lambda_{\min}^{1/2}(R)\right)^{-1},\\
\mu_0 &= \frac{2\lambda_{\min}(Q)\lambda_{\min}(R)}{a(1+a^2\eta)^2}, \;\; \mu \geq (\|\mathcal{A}^{-1}\circ\mathcal{B}\| + 1)^{-1}\mu_0.
\end{aligned}$$

The function $f(L,P)$ restricted within the feasible sublevel set (B.10) *is μ strongly convex.*

Proof. Denote $\mathcal{A}^{-1}$ as the inverse of $\mathcal{A}$, a linear operator such that $\mathcal{A}^{-1}(\mathcal{A}(P)) = P$. (Mohammadi et al., 2019b, Proposition 1) concludes that the following function $h(\cdot)$ is μ_0 strongly convex.

$$h(L) = f(L, -\mathcal{A}^{-1}(\mathcal{B}(L) + \Sigma)). \tag{B.11}$$

Define a perturbation direction $(\tilde{L}, \tilde{P})$ such that $(L+\tilde{L}, P+\tilde{P})$ is feasible. Any feasible perturbation at the point L, P will satisfy $\mathcal{A}(\tilde{P}) + \mathcal{B}(\tilde{L}) = 0$, so $\tilde{P} = -\mathcal{A}^{-1}(\mathcal{B}(\tilde{L}))$.

Let the strongly convex parameter of f in the feasible directions be μ, we will show its connection with μ_0.

Let L be perturbed by $\tilde{L}$. Apply chain rule to (B.11),

$$\nabla^2 h(L)[\tilde{L}, \tilde{L}] = \nabla^2 f(L,P)[(\tilde{L}, -\mathcal{A}^{-1}(\mathcal{B}(\tilde{L}))), (\tilde{L}, -\mathcal{A}^{-1}(\mathcal{B}(\tilde{L})))], \tag{B.12}$$

Here $\nabla^2 h(L)[\tilde{L}, \tilde{L}]$ is the Hessian operator of h at L acting on $\tilde{L}, \tilde{L}$, which equals $\langle \tilde{L}, \nabla^2 h(L) \tilde{L} \rangle$. The right hand side is defined similarly. Due to the strong convexity of h,

$$\nabla^2 h(L)[\tilde{L}, \tilde{L}] \geq \frac{\mu_0 \|\tilde{L}\|_F^2}{2}. \tag{B.13}$$

We perturb f at (L, P) in direction $(\tilde{L}, \tilde{P}) = (\tilde{L}, -\mathcal{A}^{-1}(\mathcal{B}(\tilde{L})))$. The strongly convex parameter of f in feasible directions is defined as the positive number μ such that

$$\nabla^2 f(L,P)[(\tilde{L}, \tilde{P}), (\tilde{L}, \tilde{P})] \geq \frac{\mu(\|\tilde{P}\|_F^2 + \|\tilde{L}\|_F^2)}{2}$$

for all $(\tilde{L}, \tilde{P})$ such that $\tilde{P} = -\mathcal{A}^{-1}(\mathcal{B}(\tilde{L}))$. The directional Hessian is

$$\nabla^2 f(L,P)[(\tilde{L}, \tilde{P}), (\tilde{L}, \tilde{P})] = \nabla^2 f(L,P)[(\tilde{L}, -\mathcal{A}^{-1}(\mathcal{B}(\tilde{L}))), (\tilde{L}, -\mathcal{A}^{-1}(\mathcal{B}(\tilde{L})))]. \tag{B.14}$$

(B.14) equals (B.12). So that we apply (B.13),

$$\begin{aligned} \nabla^2 f(L,P)[(\tilde{L}, \tilde{P}), (\tilde{L}, \tilde{P})] &\geq \frac{\mu_0 \|\tilde{L}\|_F^2}{2} \\ = \frac{\|\tilde{P}\|_F^2 + \|\tilde{L}\|_F^2}{2} &\cdot \frac{\mu_0 \|\tilde{L}\|_F^2}{\|\tilde{P}\|_F^2 + \|\tilde{L}\|_F^2} \\ = \frac{\|\tilde{P}\|_F^2 + \|\tilde{L}\|_F^2}{2} &\cdot \frac{\mu_0 \|\tilde{L}\|_F^2}{\|\mathcal{A}^{-1}(\mathcal{B}(\tilde{L}))\|_F^2 + \|\tilde{L}\|_F^2}. \end{aligned}$$

So

$$\mu \geq (\|\mathcal{A}^{-1} \circ \mathcal{B}\| + 1)^{-1}\mu_0.$$

□

B.3 Checking the assumptions for examples

B.3.1 Markov jump linear system

Example 2. *(Assumptions 4,5) We study the system*

$$x(t+1) = A_{w(t)}x(t) + B_{w(t)}u(t),\ w(t) \in [N].$$

The transition model is

$$\mathbf{Pr}(w(t+1) = j|w(t) = i) = \rho_{ij} \in [0,1],\ \forall t \geq 0.$$

Let $\mathbf{Pr}(w(0) = i) = p_i,\ K = [K_1, ..., K_N]$. *Define the cost as*

$$\mathcal{L}(K) = \mathbf{E}_{w,x_0} \sum_{t=0}^{\infty} x(t)^\top Qx(t) + u(t)^\top Ru(t),\ u(t) = K_{w(t)}x(t),\ \mathbf{Pr}(w(0) = i) = p_i.$$

Let the convex formulation be

$$\min\ \mathbf{tr}(Q\boldsymbol{X}_0) + \mathbf{tr}(Z_0R), \tag{B.15a}$$

$$s.t.\ \boldsymbol{X}_0 = \sum_{i=1}^{N} \boldsymbol{X}_i,\ Z_0 = \sum_{i=1}^{N} Z_i,\ \begin{bmatrix} Z_i & L_i \\ L_i^\top & \boldsymbol{X}_i \end{bmatrix} \succeq 0, \tag{B.15b}$$

$$\boldsymbol{X}_i - p_i\Sigma = \sum_{j=1}^{N} U_{ji},\ \begin{bmatrix} \rho_{ji}^{-1}U_{ji} & A_j\boldsymbol{X}_j + B_jL_j \\ (A_j\boldsymbol{X}_j + B_jL_j)^\top & \boldsymbol{X}_j \end{bmatrix} \succeq 0,\ \forall i,j \in [N]. \tag{B.15c}$$

Then the pair of problems satisfy Assumptions 4,5.

Proof. We start from the Grammian matrices below. Let $\boldsymbol{Y}_i(t) = \boldsymbol{E}(x(t)x(t)^\top \mathbf{1}_{w(t)=i})$, and $\boldsymbol{X}_i = \sum_{t=0}^{\infty} \boldsymbol{Y}_i(t)$. Then Jansch-Porto et al. (2020) suggests

$$\boldsymbol{Y}_i(t+1) = \sum_{j=1}^{N} \rho_{ji}(A_j + B_j K_j)\boldsymbol{Y}_j(t)(A_j + B_j K_j)^\top.$$

Then we can sum over the equation over time t,

$$\begin{aligned}
&\sum_{j=1}^{N} \rho_{ji}(A_j + B_j K_j)(\sum_{t=0}^{\infty} \boldsymbol{Y}_j(t))(A_j + B_j K_j)^\top \\
=&\sum_{t=0}^{\infty}\sum_{j=1}^{N} \rho_{ji}(A_j + B_j K_j)\boldsymbol{Y}_j(t)(A_j + B_j K_j)^\top \\
=&\sum_{t=0}^{\infty} \boldsymbol{Y}_i(t+1) = \sum_{t=1}^{\infty} \boldsymbol{Y}_i(t) \\
=&\sum_{t=0}^{\infty} \boldsymbol{Y}_i(t) - \boldsymbol{Y}_i(0)
\end{aligned}$$

So that

$$\sum_{j=1}^{N} \rho_{ji}(A_j + B_j K_j)\boldsymbol{X}_j(A_j + B_j K_j)^\top = \boldsymbol{X}_i - \boldsymbol{Y}_i(0).$$

Let $L_i = K_i \boldsymbol{X}_i$. We will relax the recursion as

$$\sum_{j=1}^{N} \rho_{ji}(A_j \boldsymbol{X}_j + B_j L_j)\boldsymbol{X}_j^{-1}(A_j \boldsymbol{X}_j + B_j L_j)^\top \preceq \boldsymbol{X}_i - \boldsymbol{Y}_i(0). \tag{B.16}$$

In our setting $\boldsymbol{E}(x(0)x(0)^\top) = \Sigma$ so that $\boldsymbol{Y}_i(0) = p_i \Sigma$.

Next, we will show that, if we solve the problem (B.15) with the extra constraints $K_i = L_i \boldsymbol{X}_i^{-1}$, then the function value is equal to the LQ cost of the system with controllers K_i's.

First, if we minimize over Z_i's, then we have $Z_i = L_i \boldsymbol{X}_i L_i^\top$. Moreover, the constraints

(B.15c) are equivalent to the relaxation (B.16). Suppose the equal signs are achieved in (B.16), then $\boldsymbol{X}_i$'s will be the Grammian of the system $\sum_{t=0}^{\infty} \boldsymbol{E}(x(t)x(t)^\top \mathbf{1}_{w(t)=i})$ and hence the function value is equal to the LQ cost (Costa, 2005, §4.4.2, Prop. 4.8). Now, it remains to show that, if not all (B.16) (with enumerating different j's) achieve equal signs, then the function value will only increase and not be optimal.

We define N matrices $\boldsymbol{W}_1, ..., \boldsymbol{W}_N$, such that $\boldsymbol{W}_i \succeq Y_i(0) = p_i\Sigma$, and

$$\sum_{j=1}^{N} \rho_{ji}(A_j + B_jK_j)\boldsymbol{X}_j(A_j + B_jK_j)^\top = \boldsymbol{X}_i - \boldsymbol{W}_i.$$

This corresponds to the Markov jump system

$$\tilde{x}(t+1) = A_{w(t)}\tilde{x}(t) + B_{w(t)}u(t),\ w(t) \in [N].$$

with the same parameters, transition probability matrix, controllers and a different initial condition

$$\boldsymbol{E}(\tilde{x}(t)\tilde{x}(t)^\top \mathbf{1}_{w(t)=i}) = \boldsymbol{W}_i \succeq p_i\Sigma = \boldsymbol{E}(x(t)x(t)^\top \mathbf{1}_{w(t)=i}). \tag{B.17}$$

Let $\tilde{\boldsymbol{Y}}_i(t) = \boldsymbol{E}(\tilde{x}(t)\tilde{x}(t)^\top \mathbf{1}_{w(t)=i})$ (so that $\tilde{\boldsymbol{Y}}_i(0) = \boldsymbol{W}_i$), and let $\tilde{\boldsymbol{X}}_i = \sum_{t=0}^{\infty} \tilde{\boldsymbol{Y}}_i(t)$. We will show that $\tilde{\boldsymbol{Y}}_i(t) \succeq \boldsymbol{Y}_i(t)$ for all $i = 1, ..., N$ and all $t \geq 0$.

We use induction over t. When $t = 0$, we assumed in (B.17) that $\tilde{\boldsymbol{Y}}_i(0) \succeq \boldsymbol{Y}_i(0)$ hold for all $i \in [N]$. And we have the recursions

$$\tilde{\boldsymbol{Y}}_i(t+1) = \sum_{j=1}^{N} \rho_{ji}(A_j + B_jK_j)\tilde{\boldsymbol{Y}}_j(t)(A_j + B_jK_j)^\top,$$
$$\boldsymbol{Y}_i(t+1) = \sum_{j=1}^{N} \rho_{ji}(A_j + B_jK_j)\boldsymbol{Y}_j(t)(A_j + B_jK_j)^\top.$$

If $\tilde{\boldsymbol{Y}}_i(t) \succeq \boldsymbol{Y}_i(t)$ for a certain $t \geq 0$ and for all $i \in [N]$, then the recursion implies that

$\tilde{\boldsymbol{Y}}_i(t+1) \succeq \boldsymbol{Y}_i(t+1)$ for all $i \in [N]$. We sum over t and get $\tilde{\boldsymbol{X}}_i \succeq \boldsymbol{X}_i$, so that the objective function with $\tilde{\boldsymbol{X}}_i$'s is larger than with $\boldsymbol{X}_i$'s unless $\tilde{\boldsymbol{X}}_i = \boldsymbol{X}_i$ for all $i \in [N]$.

As a result, the optimization problem (B.15) with the extra constraints $K_i = L_i \boldsymbol{X}_i^{-1}$ achieves minimum when $Z_i = L_i \boldsymbol{X}_i L_i^\top$ and (B.16) achieves equality for all $i \in [N]$. This means all $\boldsymbol{X}_i$'s are the Grammians $\sum_{t=0}^{\infty} \boldsymbol{E}(x(t)x(t)^\top \mathbf{1}_{w(t)=i})$ of the system, so that the objective function value is equal to LQ cost.

□

B.3.2 Minimizing $\mathcal{L}_2$ gain

Example 3. *(Assumptions 4,5) We consider minimizing the $\mathcal{L}_2$ gain of a closed loop system. The input output system is*

$$\dot{x} = Ax + Bu + B_w w, \ y = Cx + Du \tag{B.18}$$

and we use the state feedback controller $u = Kx$, and let

$$\mathcal{L}(K) := (\sup_{\|w\|_2 = 1} \|y\|_2)^2.$$

If we minimize the function $\mathcal{L}(K)$, this problem can be reformulated as

$$\begin{aligned} \min_{L,P,\gamma} \ & f(L,P,\gamma) := \gamma \\ s.t. \ & \begin{bmatrix} AP + PA^\top + BL + L^\top B^\top + B_w B_w^\top & (CP+DL)^\top \\ CP + DL & -\gamma I \end{bmatrix} \preceq 0. \end{aligned}$$

And $K^* = L^* P^{*-1}$. This pair of problems satisfy Assumptions 4,5.

Proof. We will check Assumption 5, which means checking

$$\mathcal{L}(K) \overset{?}{=} \min_{L,P,\gamma} \gamma \tag{B.19a}$$

$$\text{s.t.} \begin{bmatrix} AP + PA^\top + BL + L^\top B^\top + B_w B_w^\top & (CP + DL)^\top \\ CP + DL & -\gamma I \end{bmatrix} \preceq 0,\ LP^{-1} = K. \tag{B.19b}$$

Note that, the intermediate step in (Boyd et al., 1994, Sec 7.5.1) is

$$\mathcal{L}(K) = \min_{P,\gamma} \gamma, \quad \text{s.t.} \tag{B.20a}$$

$$\begin{bmatrix} (A + BK)P + P(A + BK)^\top + B_w B_w^\top & P^\top (C + DK)^\top \\ (C + DK)P & -\gamma I \end{bmatrix} \preceq 0. \tag{B.20b}$$

Denote the optimizer of (B.19) by $\hat{L}, \hat{P}, \hat{\gamma}$, and the optimizer of (B.20) by $\check{P}, \check{\gamma}$.

Note $\hat{\gamma} \leq \check{\gamma}$ because $(\gamma, L, P) = (\check{\gamma}, K\check{P}, \check{P})$ is feasible in (B.19). If (B.19) is not true (the equal sign cannot be satisfied), then $\hat{\gamma} < \check{\gamma}$, we can replace $\check{P}, \check{\gamma}$ with $\hat{P}, \hat{\gamma}$ in (B.20) and it's still feasible due to the feasibility in (B.19). Thus the optimality condition of $\check{P}, \check{\gamma}$ in (B.20) is violated, which contradicts the assumption that (B.19) is not true. Then we claim that (B.19) is true. The dissipativity uses the same change of variable and we omit the proof in Boyd et al. (1994). □

B.4 Multi-objective and mixed controller design

In this part, we study a few synthesis problems with dynamical controllers, where the objectives are about (e.g., norms of) transfer functions of the closed form system. We study the dynamical system with state, disturbance, input, output, and controller's input x, w, u, z, y

with the following dynamics

$$\begin{bmatrix} \dot{x} \\ z \\ y \end{bmatrix} = \begin{bmatrix} A & B_w & B \\ C_z & D & E \\ C & F & 0 \end{bmatrix} \begin{bmatrix} x \\ w \\ u \end{bmatrix}.$$

The controller follows

$$\begin{bmatrix} \dot{x}_c \\ u \end{bmatrix} = \begin{bmatrix} A_c & B_c \\ C_c & D_c \end{bmatrix} \begin{bmatrix} x_c \\ y \end{bmatrix}. \tag{B.21}$$

We will denote the transfer function of the closed loop system as $\mathscr{T}$, and the control problems below are typically related to $\mathscr{T}$.

In the next few subsections, we will present a few control problems:

1. The **variables** are the controller parameters $\begin{bmatrix} A_c & B_c \\ C_c & D_c \end{bmatrix}$.
2. The **objective functions** are $\mathcal{H}_2$ norm, $\mathcal{H}_\infty$ norm of $\mathscr{T}$ and the weighted sum of norms.
3. The book (Scherer & Weiland, 2000, eq(4.2.15)) defines the parameterization of the problem, by introducing the **variables that typically make the objective functions convex**:

$$v = [X, Y, K, L, M, N].$$

4. **Mapping of the variables.** Define invertible matrices U, V such that $UV^\top = I - XY$. The matrices A_c, B_c, C_c, D_c are the unique solution of

$$\begin{bmatrix} K & L \\ M & N \end{bmatrix} = \begin{bmatrix} U & XB \\ 0 & I \end{bmatrix} \begin{bmatrix} A_c & B_c \\ C_c & D_c \end{bmatrix} \begin{bmatrix} V^\top & 0 \\ CY & I \end{bmatrix} + \begin{bmatrix} XAY & 0 \\ 0 & 0 \end{bmatrix}. \tag{B.22}$$

The change of variable enables us to make some control problems as convex optimization, listed below. For simplicity of notation, let

$$\mathscr{X} = \begin{bmatrix} Y & I \\ I & X \end{bmatrix}, \ \mathscr{A} = \begin{bmatrix} AY + BM & A + BNC \\ K & AX + LC \end{bmatrix}, \tag{B.23}$$

$$\mathscr{B} = \begin{bmatrix} B_w + BNF \\ XB_w + LF \end{bmatrix}, \ \mathscr{C} = \begin{bmatrix} C_zY + EM & C_z + ENC \end{bmatrix}, \ \mathscr{D} = D + ENF. \tag{B.24}$$

Remark 6. *The mapping in* (B.22) *can be written as*

$$\begin{bmatrix} A_c & B_c \\ C_c & D_c \end{bmatrix} = \Phi(v) := \begin{bmatrix} U & XB \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} K - XAY & L \\ M & N \end{bmatrix} \begin{bmatrix} V^\top & 0 \\ CY & I \end{bmatrix}^{-1}$$

where Φ plays the role in (3.28)*. We propose a few control problems with convex forms in the next few subsections. The variables of nonconvex objective functions are A_c, B_c, C_c, D_c, the new objective functions with respect to $v = [X, Y, K, L, M, N]$ are convex, and the two forms satisfy Assumptions 7, 8. Thus the cost functions with respect to matrix $\begin{bmatrix} A_c & B_c \\ C_c & D_c \end{bmatrix}$ are gradient dominant.*

Remark 7. *In the following subsections, we refer to the result of Scherer & Weiland (2000) that, the optimal $\mathcal{H}_\infty$ design, $\mathcal{H}_2$ design and the multi-objective and robust designs, can be made convex optimization problems with the proposed way. However, this map is not guaranteed to be smooth. When matrices U, V are close to singular, the inverses of*

$$\begin{bmatrix} U & XB \\ 0 & I \end{bmatrix}, \ \begin{bmatrix} V^\top & 0 \\ CY & I \end{bmatrix}$$

are not well-defined. This makes the nonconvex objective function not gradient dominant.

For example, Tang et al. (2021) discusses the LQG problem. The dynamics takes the form

$$\dot{x}(t) = Ax(t) + Bu(t) + w(t),\ y(t) = Cx(t) + v(t)$$

where $w(t) \sim \mathcal{N}(0, W)$*,* $v(t) \sim \mathcal{N}(0, V)$*. The controller takes the form* (B.21)*. The cost function is*

$$\lim_{T\to\infty} \frac{1}{T}\boldsymbol{E}\left(\int_{t=0}^{T} (x(t)^\top Q x(t) + u(t)^\top R u(t)) dt\right).$$

The paper suggests that the set of stabilizing controllers of LQG problem can be non-connected, and the cost has saddle points. Thus, to apply our theorem and claim that the objectives with respect to controller K *are all gradient dominant, we have to restrict the problem in the set where the map is smooth, typically around the global minimum. We will review some controller design problems based on this map in the following subsections.*

B.4.1 $\mathcal{H}_\infty$ *design*

(Scherer & Weiland, 2000, §4.2.3) The goal in this part is to minimize the $\mathcal{H}_\infty$ norm of the closed loop system's transfer function by designing the optimal controller. Let the transfer function of the closed form system be $\mathscr{T}$. The problem with its raw, nonconvex form is to minimize the $\|\mathscr{T}\|_{\mathcal{H}_\infty}$ over A_c, B_c, C_c, D_c, and we will propose the convex formulation – the change of variable trick such that the argument becomes v. The problem takes the form:

$$\begin{aligned} &\min\ \gamma, \\ &\text{s.t.}\quad \mathscr{X} \succeq 0, \begin{bmatrix} \mathscr{A}^\top + \mathscr{A} & \mathscr{B} & \mathscr{C}^\top \\ \mathscr{B}^\top & -\gamma I & \mathscr{D}^\top \\ \mathscr{C} & \mathscr{D} & -\gamma I \end{bmatrix} \preceq 0. \end{aligned}$$

If we fix all other parameters listed in v and optimize over γ, then γ^* (that depends on v) is the $\mathcal{H}_\infty$ norm value of the closed loop system with the mapping from v to controller by

(B.22). If we minimize over γ and v, then we can get optimal $\mathcal{H}_\infty$ design.

B.4.2 $\mathcal{H}_2$ design

(Scherer & Weiland, 2000, §4.2.5) This part is similar to $\mathcal{H}_\infty$ design. Suppose the goal is to minimize $\|\mathscr{T}\|_{\mathcal{H}_2}$, one can alternatively solve

$$\begin{aligned} &\min\ \gamma, \\ &\text{s.t.}\quad \begin{bmatrix} \mathscr{A}^\top + \mathscr{A} & \mathscr{B} \\ \mathscr{B}^\top & -\gamma I \end{bmatrix} \preceq 0,\ \mathscr{D} = 0,\ \begin{bmatrix} \mathscr{X} & \mathscr{C}^\top \\ \mathscr{C} & Z \end{bmatrix} \succeq 0,\ \mathbf{tr}(Z) \le \gamma. \end{aligned}$$

If we fix all other parameters and optimize over γ, Z, then γ^* (that depends on v) is the $\mathcal{H}_2$ norm value of the closed loop system with the mapping from v to controller by (B.22). If we minimize over γ, Z and v, then we can get optimal $\mathcal{H}_2$ design.

B.4.3 Multi-objective synthesis

(Scherer & Weiland, 2000, §4.3) Let the system be

$$\begin{bmatrix} \dot{x} \\ z_1 \\ z_2 \\ y \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 & B \\ C_1 & D_1 & D_{12} & E_1 \\ C_2 & D_{21} & D_2 & E_2 \\ C & F_1 & F_2 & 0 \end{bmatrix} \begin{bmatrix} x \\ w_1 \\ w_2 \\ u \end{bmatrix} \tag{B.25}$$

Now we study the mixed design for $\mathcal{H}_\infty$ design from z_1 to w_1 and $\mathcal{H}_2$ design from z_2 to w_2. We keep the mapping (B.22) and the change of parameter (B.23), but replace (B.24) by

$$\mathscr{B}_i = \begin{bmatrix} B_i + BNF_i \\ XB_i + LF_i \end{bmatrix},\ \mathscr{C}_i = \begin{bmatrix} C_iY + E_iM & C_i + E_iNC \end{bmatrix},\ \mathscr{D}_i = D_i + E_iNF_i.$$

for $i = 1, 2$. Suppose we are given a positive number λ and hope to study $\|\mathscr{T}_1\|_{\mathcal{H}_\infty} + \lambda\|\mathscr{T}_2\|_{\mathcal{H}_2}$

where $\mathscr{T}_i$ is the transfer function of the i-th system (z_1 to w_1, z_2 to w_2), then we can write

$$\min \ \gamma_1 + \lambda\gamma_2, \tag{B.26}$$

$$\text{s.t.} \quad \begin{bmatrix} \mathscr{A}^\top + \mathscr{A} & \mathscr{B}_1 & \mathscr{C}_1^\top \\ \mathscr{B}_1^\top & -\gamma_1 I & \mathscr{D}_1^\top \\ \mathscr{C}_1 & \mathscr{D}_1 & -\gamma_1 I \end{bmatrix} \preceq 0, \tag{B.27}$$

$$\begin{bmatrix} \mathscr{A}^\top + \mathscr{A} & \mathscr{B}_2 \\ \mathscr{B}_2^\top & -\gamma_2 I \end{bmatrix} \preceq 0, \ \mathscr{D}_2 = 0, \ \begin{bmatrix} \mathscr{X} & \mathscr{C}_2^\top \\ \mathscr{C}_2 & Z \end{bmatrix} \succeq 0, \ \mathbf{tr}(Z) \leq \gamma_2. \tag{B.28}$$

If we fix all other parameters and optimize over γ_1, γ_2, Z, then the function value is the mixed $\mathcal{H}_\infty/\mathcal{H}_2$ norm value of the closed loop system with the mapping from v to controller by (B.22). If we minimize over γ_1, γ_2, Z and v, then we can get the optimal mixed $\mathcal{H}_\infty/\mathcal{H}_2$ design.

B.4.4 Robust state feedback control

(Scherer & Weiland, 2000, §8.1.2) We study the robust state feedback control problem, where the robustness corresponds to a system with uncertain parameters, denoted by Δ below. We apply the system model (B.25). "State feedback" means that $C = I$ and $F_1, F_2 = 0$. Let N_Δ be a positive integer. The connection between w_1 and z_1 is an uncertain channel

$$w_1(t) = \Delta(t) z_1(t)$$

for any

$$\Delta(t) \in \Delta_c := \text{conv}\{0, \Delta_1, ..., \Delta_{N_\Delta}\}.$$

The goal is to minimize a certain norm of the transfer function from z_2 to w_2, which can be $\mathcal{H}_2$ norm, $\mathcal{H}_\infty$ norm studied in the previous part. We consider minimizing the norms under an extra constraint when the closed loop system achieves stability with z_1 to w_1 (z_1 with finite norm) and robust quadratic performance with z_2 to w_2 via a matrix P_p. The robust quadratic performance is defined as: there exists a matrix P_p,

$$P_p = \begin{bmatrix} \tilde{Q}_p & \tilde{S}_p \\ \tilde{S}_p^\top & \tilde{R}_p \end{bmatrix}, \; P_p^{-1} = \begin{bmatrix} Q_p & S_p \\ S_p^\top & R_p \end{bmatrix}$$

such that $\tilde{R}_p \succ 0, Q_p \prec 0$, and

$$\int_0^\infty \begin{bmatrix} w_2(t) \\ z_2(t) \end{bmatrix}^\top P_p \begin{bmatrix} w_2(t) \\ z_2(t) \end{bmatrix} \mathrm{d}t \le \epsilon \|w_2\|_{\mathcal{H}_2}^2$$

for some $\epsilon > 0$.

Define new variables Q, S, R in addition to $v = [X, Y, K, L, M, N]$, and let $\mathscr{M}$ replace

$$\mathscr{M} \leftarrow \begin{bmatrix} -(AY + BM)^\top & -(C_1Y + E_1M)^\top & -(C_2Y + E_2M)^\top \\ I & 0 & 0 \\ -B_1^\top & -D_1^\top & -D_{21}^\top \\ 0 & I & 0 \\ -B_2^\top & -D_{12}^\top & -D_2^\top \\ 0 & 0 & I \end{bmatrix}.$$

The constraints, which is proven to be convex in (Scherer & Weiland, 2000, §8.1.2) can be written as

$$R \succ 0,\ Q \prec 0,\ \begin{bmatrix} I \\ -\Delta_j \end{bmatrix}^\top \begin{bmatrix} Q & S \\ S^\top & R \end{bmatrix} \begin{bmatrix} I \\ -\Delta_j \end{bmatrix} \prec 0,\ \forall j \in [N_\Delta]$$

$$Y \succ 0,\ \mathscr{M}^\top \begin{bmatrix} 0 & I & 0 & 0 & 0 & 0 \\ I & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & Q & S & 0 & 0 \\ 0 & 0 & S^\top & R & 0 & 0 \\ 0 & 0 & 0 & 0 & Q_p & S_p \\ 0 & 0 & 0 & 0 & S_p^\top & R_p \end{bmatrix} \mathscr{M} \succ 0.$$

For example, if we aim to minimize the $\mathcal{H}_2$ norm of the transfer function from z_2 to w_2, then we can minimize γ_2 subject to (B.28) and the constraints above. With the convex formulation, if we apply policy gradient descent with respect to $\mathcal{H}_2$ norm of the transfer function from z_2 to w_2 with robust stability of system 1 (z_1 with finite $\mathcal{H}_2$ norm) and robust quadratic performance constraints of system 2, then policy gradient descent converges to globally optimal controller.

B.4.5 Discrete time system

(Scherer & Weiland, 2000, §4.6) Suppose we study the discrete time system, and we define the system in a similar way of defining the continuous time system:

$$\begin{bmatrix} x(t+1) \\ z_1(t) \\ z_2(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 & B \\ C_1 & D_1 & D_{12} & E_1 \\ C_2 & D_{21} & D_2 & E_2 \\ C & F_1 & F_2 & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ w_1(t) \\ w_2(t) \\ u(t) \end{bmatrix},\quad \begin{bmatrix} x_c(t+1) \\ u(t) \end{bmatrix} = \begin{bmatrix} A_c & B_c \\ C_c & D_c \end{bmatrix} \begin{bmatrix} x_c(t) \\ y(t) \end{bmatrix}.$$

Now we study the mixed design for $\mathcal{H}_\infty$ design from z_1 to w_1 and $\mathcal{H}_2$ design from z_2 to w_2. Suppose we are given a positive number λ and hope to study $\|\mathscr{T}_1\|_{\mathcal{H}_\infty} + \lambda \|\mathscr{T}_2\|_{\mathcal{H}_2}$ where $\mathscr{T}_i$ is the transfer function of the i-th system (z_1 to w_1, z_2 to w_2), then we can write

$$\begin{aligned}
&\min \ \gamma_1 + \lambda\gamma_2, \\
&\text{s.t.} \ \begin{bmatrix} \mathscr{X} & 0 & \mathscr{A}^\top & \mathscr{C}_1^\top \\ 0 & \gamma_1 I & \mathscr{B}_1^\top & \mathscr{D}_1 \\ \mathscr{A} & \mathscr{B}_1 & \mathscr{X} & 0 \\ \mathscr{C}_1 & \mathscr{D}_1 & 0 & \gamma_1 I \end{bmatrix} \succ 0, \ \mathbf{tr}(Z) \leq \gamma_2, \\
&\begin{bmatrix} \mathscr{X} & \mathscr{A} & \mathscr{B}_2 \\ \mathscr{A}^\top & \mathscr{X} & 0 \\ \mathscr{B}_2^\top & 0 & \gamma_2 I \end{bmatrix} \succ 0, \ \begin{bmatrix} \mathscr{X} & 0 & \mathscr{C}_2 \\ 0 & \mathscr{X} & \mathscr{D}_2 \\ \mathscr{C}_2 & \mathscr{D}_2 & Z \end{bmatrix} \succ 0.
\end{aligned}$$

If we fix all other parameters and optimize over γ_1, γ_2, Z, then the function value is the mixed $\mathcal{H}_\infty/\mathcal{H}_2$ value of the closed loop system with the mapping from v. If we minimize over γ_1, γ_2, Z and v, then we can get the optimal mixed $\mathcal{H}_\infty/\mathcal{H}_2$ design.

B.5 System level synthesis with infinite horizon

We studied the landscape of the optimal control problem where the variables are matrices (which are finite dimensional), and SLS for finite horizon problem was an example. Generally, SLS also works with the infinite horizon problem. In this regime, the variables are *transfer functions* and they are infinite dimensional. In practice, when the problem is made convex, one can parameterize the transfer function (say as finite impulse response) and minimize the cost with respect to the finite dimensional parameters. However, Theorem 2 does not apply to the infinite dimensional optimization problems, and it is not obvious that the finite dimensional parameterization satisfies the assumptions for our main theorem. We review the infinite horizon SLS here. A future direction is to judge whether the Lojasiewicz inequality holds in the space of transfer function or its parameterized form, and how to analyze it using SLS.

Example 4. *(System level synthesis with infinite horizon (Anderson et al., 2019)) Suppose* *one has a discrete time dynamical system with*

$$x(t+1) = Ax(t) + Bu(t) + w(t).$$

One can apply a dynamic controller $K(z)$. The goal is to find the optimial controller which minimizes the LQR cost where $u(z) = K(z)x(z)$

$$\mathcal{L}(K) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} x(t)^\top Q x(t) + u(t)^\top R u(t).$$

Suppose x_0, w_t are i.i.d. from $\mathcal{N}(0, \Sigma)$. The SLS defines two transfer functions $\Phi_X(z), \Phi_U(z)$, and solve the following convex optimization problem

$$\begin{aligned}
&\min_{\Phi_X(z),\Phi_U(z)} \left\| \begin{bmatrix} Q^{1/2}\Phi_X(z) \\ R^{1/2}\Phi_U(z) \end{bmatrix} \Sigma^{1/2} \right\|_{\mathcal{H}_2}, \\
&s.t. \ \begin{bmatrix} zI - A & -B \end{bmatrix} \begin{bmatrix} \Phi_X(z) \\ \Phi_U(z) \end{bmatrix} = I, \ \Phi_X(z), \Phi_U(z) \in \frac{1}{z}\mathcal{RH}_\infty.
\end{aligned}$$

Let the optimizer be $\Phi_U^*(z), \Phi_X^*(z)$. The optimal controller is $K^*(z) = \Phi_U^*(z)(\Phi_X^*(z))^{-1}$.

B.6 Conditions of convexifiable nonconvex cost

We consider the pair of problems in Theorem 2, and ask the question: what property of the nonconvex cost function $\mathcal{L}(K)$ allows us to reformulate the problem (3.9) as a *convex* optimization problem (3.10)? In this section we propose the following lemma.

Lemma 17. *Suppose Assumptions 4, 6 hold, and $\mathcal{L}(LP^{-1})$ as a function of L, P is differentiable.We define the notation $\nabla^2_{L,P}\mathcal{L}(LP^{-1})[\Gamma_1, \Gamma_2]$ as in* (B.29). *If $\nabla^2_{L,P}\mathcal{L}(LP^{-1})[\Gamma_1, \Gamma_2] > 0$ for all $(L, P) \in \mathcal{S}$ and all (Γ_1, Γ_2) such that $\mathcal{A}(\Gamma_2) + \mathcal{B}(\Gamma_1) = 0$, then we can define a convex function $f(L, P)$ so that Assumption 4 holds.*

For the convex formulation with the above lemma, we can apply Theorem 2 so that all stationary points of $\mathcal{L}(K)$ are global minimum.

Proof. Suppose we observe the simple version (3.11). We know from Assumption 6 that, $f(L, P) = \mathcal{L}(K) = \mathcal{L}(LP^{-1})$ is convex in L, P. We take the Hessian and ask for

$$\nabla \begin{bmatrix} \nabla \mathcal{L}(LP^{-1})P^{-1} \\ -P^{-1}L^\top \nabla \mathcal{L}(LP^{-1})P^{-1} \end{bmatrix} \succ 0.$$

Note that this is a tensor and it is positive definite. For simplicity, we analyze the directional Hessian as the following. We expand the left hand side of the inequality above and define $\nabla^2_{L,P}\mathcal{L}(LP^{-1})[\Gamma_1, \Gamma_2]$ as

$$\begin{aligned} &\nabla^2_{L,P}\mathcal{L}(LP^{-1})[\Gamma_1, \Gamma_2] \\ :=\; &\nabla^2\mathcal{L}(LP^{-1})[\Gamma_1 G^{-2}, \Gamma_1] - 2\nabla^2\mathcal{L}(LP^{-1})[\Gamma_1, LP^{-3}\Gamma_2] \\ &- 2\langle \Gamma_1, \nabla\mathcal{L}(LP^{-1})P^{-1}\Gamma_2 P^{-1}\rangle + 2\langle \Gamma_2, LP^{-1}\Gamma_2 P^{-1}\nabla\mathcal{L}(LP^{-1})P^{-1}\rangle \\ &+ \nabla^2\mathcal{L}(LP^{-1})[LP^{-2}\Gamma_2, LP^{-2}\Gamma_2]. \end{aligned} \tag{B.29}$$

This is the directional Hessian of $\mathcal{L}$ with respect to (L, P) in direction (Γ_1, Γ_2). Thus, if $\nabla^2_{L,P}\mathcal{L}(LP^{-1})[\Gamma_1, \Gamma_2] > 0$ for all $(L, P) \in \mathcal{S}$ and all (Γ_1, Γ_2) such that $\mathcal{A}(\Gamma_2) + \mathcal{B}(\Gamma_1) = 0$ (which is a condition on nonconvex cost $\mathcal{L}$), we know that $f(L, P)$ is convex in L, P and the convex formulation can be made. □

Appendix C

APPENDIX OF CHAPTER 4

C.1 Sample complexity for MISO and MIMO problems

For multi-rollout case, we only observe the output at time $2n-1$, and let $u_{2n}=0$, we have

$$y_{2n-1}=\sum_{i=1}^{2n-2} CA^{2n-2-i}Bu_i + Du_{2n-1}. \tag{C.1}$$

Denote the impulse response by $h \in \mathbb{R}^{p(2n-1)}$, which is a block vector

$$h = \begin{bmatrix} h^{(1)} \\ h^{(2)} \\ \dots \\ h^{(2n-1)} \end{bmatrix}$$

where each block $h^{(i)} \in \mathbb{R}^p$. $\beta \in \mathbb{R}^{p(2n-1)}$ is a weighted version of h, with weights

$$\beta^{(i)} = K_i h^{(i)}$$

and

$$\beta = \begin{bmatrix} \beta^{(1)} \\ \beta^{(2)} \\ \dots \\ \beta^{(2n-1)} \end{bmatrix}$$

Define the reweighted Hankel map for the same h by

$$\mathcal{G}(\beta) = \begin{bmatrix} \beta^{(1)}/K_1 & \beta^{(2)}/K_2 & \beta^{(3)}/K_3 & ... \\ \beta^{(2)}/K_2 & \beta^{(3)}/K_3 & \beta^{(4)}/K_2 & ... \\ ... & & & \end{bmatrix}^T \in \mathbb{R}^{n\times pn}$$

and $\mathcal{G}^*$ is the adjoint of $\mathcal{G}$. We define each rollout input $u_1, ..., u_{2n-1}$ as independent Gaussian vectors with

$$u_i \sim \mathcal{N}(0, K_i^2 \mathbf{I}) \tag{C.2}$$

Now let $\boldsymbol{U} \in \mathbb{R}^{T\times p(2n-1)}$, each entry is iid standard Gaussian. Let $y \in \mathbb{R}^T$ be the concatenation of outputs

$$y = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_T \end{bmatrix}$$

where $y_i \in \mathbb{R}^m$ is defined in (C.1). We consider the question

$$\begin{aligned} \min_{\beta'} \quad & \|\mathcal{G}(\beta')\|_* \\ \text{s.t.,} \quad & \|\boldsymbol{U}\beta' - y\|_2 \le \delta \end{aligned} \tag{C.3}$$

where the norm of overall (state and output) noise is bounded by δ. We will present the following theorem, which generalizes the result of Cai et al. (2016) from SISO case to MISO case.

Theorem 13. *Let β be the true impulse response. If $T = \Omega((\sqrt{pR}\log n + \epsilon)^2)$ is the number of output observations, C is some constant, the solution $\hat{\beta}$ to* (C.3) *satisfies $\|\beta - \hat{\beta}\|_2 \le 2\delta/\epsilon$*

with probability

$$1 - \exp\left(-\frac{1}{2}(\sqrt{T-1} - C(\sqrt{pR}\log n + \epsilon) - \epsilon)^2\right).$$

When the system output is $y = \boldsymbol{U}\beta + z$ *and* z *is i.i.d. Gaussian noise with variance* σ_z^2*, we have that* $\|\beta - \hat{\beta}\|_2 \lesssim (\sqrt{pR + \epsilon})\sigma_z \log n$ *with probability (Oymak et al., 2013, Thm 1)*

$$1 - 6\exp\left(-\frac{1}{2}(\sqrt{T-1} - C(\sqrt{pR}\log n + \epsilon) - \epsilon)^2\right).$$

This theorem says that when the input dimension is p, the sample complexity is $O(\sqrt{pR}\log n)$. The proof strongly depends on the following lemma (Cai et al. (2016); Gordon (1988)):

Lemma 18. *Define the Gaussian width*[1]

$$w(S) := E_g(\sup_{\gamma \in S} \gamma^T g)$$

where g *is standard Gaussian vector of size* p*. Let* $\Phi = \mathcal{I}(\beta) \cap \mathbb{S}$ *where* $\mathbb{S}$ *is unit sphere. We have*

$$P(\min_{z \in \Phi} \|\boldsymbol{U}z\|_2 < \epsilon) \leq \exp\left(-\frac{1}{2}(\sqrt{T-1} - w(\Phi) - \epsilon)^2\right).$$

We will present the proof in Appendix C.1.1.

MIMO. For MIMO case, we say output size is m. We take each channel of output as a

[1] The Gaussian width of the normal cone of (4.6) and (C.3) are different up to a constant Banerjee et al. (2014).

system of at most order R, and solve m problems

$$\begin{aligned} \boldsymbol{P}_i : \min_{\beta_i} \quad & \|\mathcal{G}(\beta_i)\|_* \\ \text{s.t.,} \quad & \|\boldsymbol{U}x_i - y_i\|_2 \leq \delta, \\ y_i \in \mathbb{R}^T & \text{ is the } i\text{th output.} \end{aligned}$$

and for each problem we have failure probability equal to (C.5), which means the total failure probability is

$$m \exp\left(-\frac{1}{2}(\sqrt{T-1} - w(\Phi) - \epsilon)^2\right)$$

so we need $T = O((\sqrt{pR}\log n + \log(m) + \epsilon)^2)$. Let the solution to those optimization problems be $[x_1^*, ..., x_m^*]$, and the true impulse response be $[\hat{x}_1, ..., \hat{x}_m]$, then $\|[x_1^*, ..., x_m^*] - [\hat{x}_1, ..., \hat{x}_m]\|_F \leq \sqrt{m}\delta/\epsilon$ with probability

$$1 - \exp\left(-\frac{1}{2}(\sqrt{T-1} - w(\Phi) - \epsilon)^2\right)$$

Another way is that, for each rollout of input data, the output is m dimensional, but we take 1 channel of output from the observation and throw away other $m-1$ output. And we uniformly pick among channels and get T observations for each channel, and in total mT observations/input rollouts. In this case, when the sample complexity is $m\sqrt{pR}\log n$ (m times of before), we can recover the impulse response with Frobenius norm $\sqrt{m}\delta/\epsilon$ with probability

$$1 - \exp\left(-\frac{1}{2}(\sqrt{T/m-1} - w(\Phi) - \epsilon)^2\right)$$

C.1.1 Proof of Theorem 13

We will only prove the first equation.

Proof. Let $\mathcal{I}(\beta)$ be the descent cone of $\|\mathcal{G}(\beta)\|_*$ at β, we have the following lemma:

Lemma 19. *Assume*

$$\min_{z\in\mathcal{I}(\beta)} \frac{\|\boldsymbol{U}z\|_2}{\|z\|_2} \geq \epsilon,$$

then $\|\beta - \hat{\beta}\|_2 \leq 2\delta/\epsilon$.

(Proof is same in (Cai et al., 2016, Lemma 1), we omit it here) To prove Theorem 13, we only need lower bound LHS with Lemma 19. The following lemma gives the probability that LHS is lower bounded.

Lemma 20. *Define the Gaussian width*

$$w(S) := E_g(\sup_{\gamma\in S} \gamma^T g) \tag{C.4}$$

where g *is standard Gaussian vector of size* p*. Let* $\Phi = \mathcal{I}(\beta) \cap \mathbb{S}$ *where* $\mathbb{S}$ *is unit sphere. We have*

$$P(\min_{z\in\Phi} \|\boldsymbol{U}z\|_2 < \epsilon) \leq \exp\left(-\frac{1}{2}(\sqrt{T-1} - w(\Phi) - \epsilon)^2\right). \tag{C.5}$$

Now we need to study $w(\Phi)$.

Lemma 21. *(Cai et al. (2016) eq. (17)) Let* $\mathcal{I}^*(\beta)$ *be the dual cone of* $\mathcal{I}(\beta)$*, then*

$$w(\Phi) \leq E(\min_{\gamma\in\mathcal{I}^*(\beta)} \|g - \gamma\|_2). \tag{C.6}$$

Note that $\mathcal{I}^*(\beta)$ is just the cone of subgradient of $\mathcal{G}(\beta)$, so it can be written as

$$\mathcal{I}^*(\beta) = \{\mathcal{G}^*(V_1V_2^T + W) | V_1^T W = 0, WV_2 = 0, \|W\| \leq 1\}$$

where $\mathcal{G}(\beta) = V_1 \Sigma V_2^T$ is the SVD of $\mathcal{G}(\beta)$. So

$$\min_{\gamma \in \mathcal{I}^*(\hat{x})} \|g - \gamma\|_2 = \min_{\lambda, W} \|\lambda \mathcal{G}^*(V_1 V_2^T + W) - g\|_2.$$

For RHS, we have

$$\begin{aligned}
\|\lambda \mathcal{G}^*(V_1 V_2^T + W) - g\|_2 &= \|\lambda \mathcal{G}\mathcal{G}^*(V_1 V_2^T + W) - \mathcal{G}(g)\|_F \\
&= \|\lambda(V_1 V_2^T + W) - \mathcal{G}(g)\|_F + \|\lambda(I - \mathcal{G}\mathcal{G}^*)(V_1 V_2^T + W)\|_F \\
&\leq \|\lambda(V_1 V_2^T + W) - \mathcal{G}(g)\|_F.
\end{aligned}$$

Let $\mathcal{P}_W$ be projection operator onto subspace spanned by W, i.e.,

$$\{W | V_1^T W = 0, W V_2 = 0\}$$

and $\mathcal{P}_V$ be projection onto its orthogonal complement. Choose $\lambda = \|\mathcal{P}_W(\mathcal{G}(g))\|$ and $W = \mathcal{P}_W(\mathcal{G}(g))/\lambda$.

$$\begin{aligned}
\|\lambda(V_1 V_2^T + W) - \mathcal{G}(g)\|_F &= \|\mathcal{G}(g) - \mathcal{P}_W(\mathcal{G}(g)) - \|\mathcal{P}_W(\mathcal{G}(g))\| V_1 V_2^T\|_F \\
&\leq \|\mathcal{P}_V(\mathcal{G}(g)) - \|\mathcal{P}_W(\mathcal{G}(g))\| V_1 V_2^T\|_F \\
&\leq \|\mathcal{P}_V(\mathcal{G}(g))\|_F + \|\mathcal{P}_W(\mathcal{G}(g))\| \|V_1 V_2^T\|_F \\
&= \|\mathcal{P}_V(\mathcal{G}(g))\|_F + \sqrt{R}\|\mathcal{P}_W(\mathcal{G}(g))\| \\
&= \|\mathcal{P}_V(\mathcal{G}(g))\|_F + \sqrt{R}\|\mathcal{G}(g)\|.
\end{aligned}$$

Bound the first term by (note V_1 and V_2 span R dimensional space, so $V_1 \in \mathbb{R}^{n \times R}$ and

$V_2 \in \mathbb{R}^{pn\times R}$)

$$\begin{aligned}\|\mathcal{P}_V(\mathcal{G}(g))\|_F &= \|V_1V_1^T\mathcal{G}(g) + (I - V_1V_1^T)\mathcal{G}(g)V_2V_2^T\|_F \\ &\leq \|V_1V_1^T\mathcal{G}(g)\|_F + \|\mathcal{G}(g)V_2V_2^T\|_F \\ &\leq 2\sqrt{R}\|\mathcal{G}(g)\|.\end{aligned}$$

we get

$$\begin{aligned}w(\Phi) &\leq E(\min_{\lambda,W}\|\lambda\mathcal{G}^*(V_1V_2^T + W) - g\|_2) \\ &\leq E(\|\lambda\mathcal{G}^*(V_1V_2^T + W) - g\|_2)\big|_{\lambda=\|\mathcal{P}_W(\mathcal{G}(g))\|, W=\mathcal{P}_W(\mathcal{G}(g))/\lambda} \\ &\leq 3\sqrt{R}\|\mathcal{G}(g)\|.\end{aligned}$$

We know that, if $p = 1$, then $E\|\mathcal{G}(g)\| = O(\log(n))$. For general p, let

$$g^{(i)} = [g_1^{(i)}, ..., g_p^{(i)}]^T,$$

we rearrange the matrix as

$$\begin{aligned}\bar{\mathcal{G}}(g) &= \begin{bmatrix}\begin{bmatrix} g_1^{(1)} & g_1^{(2)}/\sqrt{2} & ... \\ g_1^{(2)}/\sqrt{2} & g_1^{(3)}/\sqrt{3} & ... \\ ... & & \end{bmatrix} & \begin{bmatrix} g_2^{(1)} & g_2^{(2)}/\sqrt{2} & ... \\ g_2^{(2)}/\sqrt{2} & g_2^{(3)}/\sqrt{3} & ... \\ ... & & \end{bmatrix} & ...\end{bmatrix} \\ &= [G_1, ..., G_p]\end{aligned}$$

where expectation of operator norm of each block is $\log(n)$. Then (note v below also has a block structure $[v^{(1)}; ...; v^{(n)}]$)

$$\begin{aligned}\|\bar{\mathcal{G}}(g)\| &= \max_{u,v} \frac{u^T \bar{\mathcal{G}}(g) v}{\|u\|\|v\|} \\ &= \max_{u,v^1,\ldots,v^p} \sum_{i=1}^{p} \frac{u^T G_i v^{(i)}}{\|u\|\|v\|} \\ &\le \max_{v^1,\ldots,v^p} O(\log(n)) \frac{\sum_{i=1}^{p} \|v^{(i)}\|}{\sqrt{\sum_{i=1}^{p} \|v^{(i)}\|^2}} \\ &\le O(\sqrt{p}\log(n)).\end{aligned}$$

And $\|\bar{\mathcal{G}}(g)\| = \|\mathcal{G}(g)\|$. So we have $\|\mathcal{G}(g)\| = \sqrt{p}\log(n)$. So $w(\Phi) = C\sqrt{pR}\log(n)$. Get back to (C.5), we want the probability be smaller than 1, and we get

$$\sqrt{T-1} - C\sqrt{pR}\log n - \epsilon > 0$$

thus $T = O((\sqrt{pR}\log(n) + \epsilon)^2)$.

At the end, we give a different version of Theorem 13. Theorem 13 in Cai et al. (2016) works for the any noise with bounded norm. Here we consider the iid Gaussian noise, and use the result in Oymak et al. (2013), we have the following result.

Theorem 14. *Let the system output $y = \boldsymbol{U}\beta + z$ where $\boldsymbol{U}$ entries are iid Gaussian $\mathcal{N}(0, 1/T)$, β is the true system parameter and $z \sim \mathcal{N}(0, \sigma_z^2)$. Then (C.3) recovers $\hat{\beta}$ with error $\|\hat{\beta} - \beta\|_2 \le w(\Phi)\|z\|_2/\sqrt{T} \lesssim \sqrt{pR}\sigma_z \log n$ with high probability.*

Remark 8. *Since the power of $\boldsymbol{U}$ is n times of that of $\bar{\boldsymbol{U}}$ and the variance of $\boldsymbol{U}$ is $1/T$, $\sigma_z = \sqrt{n/T}\sigma$, we have $\|\hat{h} - h\|_2 \le \|\hat{\beta} - \beta\|_2 \lesssim \sqrt{\frac{pnR}{T}}\sigma \log n$.*

□

C.2 Proof of regularization algorithm's spectral norm error

Theorem 4. *Consider the problem* (HNN) *in the MISO (multi-input single-output) setting (m=1, p inputs). Suppose the system is order R, $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$, each row consists of* *an input rollout* $u^{(i)} \in \mathbb{R}^{(2n-1)p}$*, and the scaled* $\boldsymbol{U} = \bar{\boldsymbol{U}}K^{-1}$ *has i.i.d Gaussian entries. Let* $\boldsymbol{snr} = \mathbb{E}[\|u\|^2/n]/\,\mathbb{E}[\|z\|^2]$ *and* $\sigma = 1/\sqrt{\boldsymbol{snr}}$*. Let* $\lambda = \sigma\sqrt{\frac{pn}{T}}\log(n)$*. Then, the problem* (HNN) *returns* $\hat{h}$ *such that*

$$\frac{\|\hat{h}-h\|_2}{\sqrt{2}} \le \|\mathcal{H}(\hat{h}-h)\| \lesssim \begin{cases} \sqrt{\frac{np}{\boldsymbol{snr}\times T}}\log(n) & \text{if} \quad T \gtrsim \min(R^2, n) \\ \sqrt{\frac{Rnp}{\boldsymbol{snr}\times T}}\log(n) & \text{if} \quad R \lesssim T \lesssim \min(R^2, n). \end{cases} \tag{4.7}$$

We will prove the first case of (4.7). The second case is a direct application of Theorem 14.

Lemma 22. *Suppose* $\xi \sim \mathcal{N}(0, \sigma_\xi I)$*,* $T \lesssim pR^2\log^2 n$*, and* $\boldsymbol{U}$ *has iid Gaussian entries with* $\mathbf{E}(\boldsymbol{U}^\top\boldsymbol{U}) = 1$*. Then, we have that* $\mathbf{E}(\Gamma) < 0.5$*, and* $P(\Gamma < 0.5) \ge 1 - O(R\log n\sqrt{p/T})$*. In this case* $\|\mathcal{G}(\hat{\beta}-\beta)\| \lesssim \sigma_\xi\sqrt{p}\log n$*.*

Remark 9. *To be consistent with the main theorem, we need to find the relation between* σ_ξ *and SNR, or* σ*. We do the following computation: (1)* $\mathcal{G}(\hat{\beta}-\beta) = \mathcal{H}(\hat{h}-h)$*, so we are bounding the Hankel spectral norm error here; (2) Each column of the input is unit norm, so each input is* $\mathcal{N}(0, 1/T)$*, and the average power of input is* $1/T$*; (3) Because of the scaling matrix* K*, the actual input of* $\bar{\boldsymbol{U}}$ *is* n *times the power of entries in* $\boldsymbol{U}$*. With all above discussion, we have* $\sigma_\xi = \sigma\sqrt{n/T}$*, which results in* $\|\mathcal{G}(\hat{\beta}-\beta)\| \lesssim \sqrt{\frac{np}{T}}\sigma\log n$*.*

Proof. Now we bound $\|\mathcal{G}(\hat{\beta}-\beta)\|$ by partitioning it to $\|\mathcal{G}(I-\boldsymbol{U}^T\boldsymbol{U})(\hat{\beta}-\beta)\|$ and $\|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta}-\beta))\|$. We have

$$\begin{aligned} \|\mathcal{G}(I-\boldsymbol{U}^T\boldsymbol{U})(\hat{\beta}-\beta)\| &= \|\mathcal{G}(I-\boldsymbol{U}^T\boldsymbol{U})\mathcal{G}^*\mathcal{G}(\hat{\beta}-\beta)\| \\ &\le \|\mathcal{G}(I-\boldsymbol{U}^T\boldsymbol{U})\mathcal{G}^*\|_{2,\mathcal{G}\mathcal{J}(\beta)}\|\mathcal{G}(\hat{\beta}-\beta)\| \\ &= \Gamma\|\mathcal{G}(\hat{\beta}-\beta)\|. \end{aligned} \tag{C.7}$$

And then we also have

$$
\begin{aligned}
\|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta}-\beta))\| &= \|\mathcal{G}\boldsymbol{U}^T(\boldsymbol{U}\hat{\beta}-y+\xi)\| \\
&\leq \|\mathcal{G}\boldsymbol{U}^T(\boldsymbol{U}\hat{\beta}-y)\| + \|\mathcal{G}(\boldsymbol{U}^T\xi)\|.
\end{aligned}
$$

Since $\hat{\beta}$ is the optimizer, we have

$$
\boldsymbol{U}^T(\boldsymbol{U}\hat{\beta}-y) + \lambda\mathcal{G}^*(\hat{V}_1\hat{V}_2^T + \hat{W}) = 0,
$$

where $\mathcal{G}(\hat{\beta}) = \hat{V}_1\hat{\Sigma}\hat{V}_2^T$ is the SVD of $\mathcal{G}(\hat{\beta})$, $\hat{W} \in \mathbb{R}^{n\times n}$ where $\hat{V}_1^T\hat{W} = 0, \hat{W}\hat{V}_2 = 0, \|\hat{W}\| \leq 1$. We have

$$
\|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta}-\beta))\| \leq \|\mathcal{G}(\boldsymbol{U}^T\xi)\| + \lambda. \tag{C.8}
$$

Combining (C.7) and (C.8), we have

$$
\begin{aligned}
\|\mathcal{G}(\hat{\beta}-\beta)\| &\leq \|\mathcal{G}(I-\boldsymbol{U}^T\boldsymbol{U})(\hat{\beta}-\beta)\| + \|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta}-\beta))\| \\
&\leq \Gamma\|\mathcal{G}(\hat{\beta}-\beta)\| + \|\mathcal{G}(\boldsymbol{U}^T\xi)\| + \lambda
\end{aligned}
$$

or equivalently,

$$
\|\mathcal{G}(\hat{\beta}-\beta)\| \leq \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\| + \lambda}{1-\Gamma}, \quad \Gamma = \|\mathcal{G}(I-\boldsymbol{U}^T\boldsymbol{U})\mathcal{G}^*\|_{2,\mathcal{G}\mathcal{J}(\beta)}.
$$

Bounding Γ. Denote the SVD of $\mathcal{G}(\beta) = V_1\Sigma V_2^T$. Denote projection operators $\mathcal{P}_V(M) = V_1V_1^TM + MV_2V_2^T - V_1V_1^TMV_2V_2^T$ and $\mathcal{P}_W(M) = M - \mathcal{P}_V(M)$. First we prove some side results for later use. From optimality of $\hat{\beta}$, we have

$$\begin{aligned}
& \frac{1}{2}\|y-\boldsymbol{U}\hat{\beta}\|^2+\lambda\|\mathcal{G}\hat{\beta}\|_* \le \frac{1}{2}\|y-\boldsymbol{U}\beta\|^2+\lambda\|\mathcal{G}\beta\|_* = \frac{1}{2}\|\xi\|^2+\lambda\|\mathcal{G}\beta\|_* \\
\Rightarrow\quad & \frac{1}{2}\|\boldsymbol{U}\beta+\xi-\boldsymbol{U}\hat{\beta}\|^2+\lambda\|\mathcal{G}\hat{\beta}\|_* \le \frac{1}{2}\|\xi\|^2+\lambda\|\mathcal{G}\beta\|_* \\
\Rightarrow\quad & \frac{1}{2}\|\boldsymbol{U}(\beta-\hat{\beta})\|^2+\xi^T\boldsymbol{U}(\beta-\hat{\beta})+\lambda\|\mathcal{G}\hat{\beta}\|_* \le \lambda\|\mathcal{G}\beta\|_* \\
\Rightarrow\quad & \lambda\|\mathcal{G}\hat{\beta}\|_* \le \lambda\|\mathcal{G}\beta\|_*+\xi^T\boldsymbol{U}(\hat{\beta}-\beta) \\
\Rightarrow\quad & \|\mathcal{G}\hat{\beta}\|_*-\|\mathcal{G}\beta\|_* \le \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}\|\mathcal{G}(\hat{\beta}-\beta)\|_* \qquad (C.9)
\end{aligned}$$

Eq.(C.9) is an important result to note, and following that,

$$\begin{aligned}
& \|\mathcal{G}\hat{\beta}\|_*-\|\mathcal{G}\beta\|_* \le \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}\|\mathcal{G}(\hat{\beta}-\beta)\|_* \\
\Rightarrow\quad & \langle\mathcal{G}(\hat{\beta}-\beta),V_1V_2^T+W\rangle \le \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}\|\mathcal{G}(\hat{\beta}-\beta)\|_* \\
\Rightarrow\quad & \|\mathcal{P}_W\mathcal{G}(\hat{\beta}-\beta)\|_* \le -\langle\mathcal{G}(\hat{\beta}-\beta),V_1V_2^T\rangle+\frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}\|\mathcal{G}(\hat{\beta}-\beta)\|_* \\
\Rightarrow\quad & \|\mathcal{P}_W\mathcal{G}(\hat{\beta}-\beta)\|_* \le \|\mathcal{P}_V\mathcal{G}(\hat{\beta}-\beta)\|_*+\frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}(\|\mathcal{P}_V\mathcal{G}(\hat{\beta}-\beta)\|_*+\|\mathcal{P}_W\mathcal{G}(\hat{\beta}-\beta)\|_*) \\
\Rightarrow\quad & \|\mathcal{P}_W\mathcal{G}(\hat{\beta}-\beta)\|_* \le \frac{1+\frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}}{1-\frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}}\|\mathcal{P}_V\mathcal{G}(\hat{\beta}-\beta)\|_* \qquad (C.10)
\end{aligned}$$

Let $\boldsymbol{U}$ be iid Gaussian matrix with scaling $\boldsymbol{E}(\boldsymbol{U}^T\boldsymbol{U})=I$. Here we need to study the Gaussian width of the normal cone $w(\mathcal{J}(\beta))$ of (4.6). Banerjee et al. (2014) proves that, if (C.9) is true, and $\lambda \ge 2\|\mathcal{G}(\boldsymbol{U}^T\xi)\|$, then the Gaussian width of this set (intersecting with unit ball) is less than 3 times of Gaussian width of $\{\hat{\beta} : \|\mathcal{G}(\hat{\beta})\|_* \le \|\mathcal{G}(\beta)\|_*\}$, which is $O(\sqrt{R}\log n)$ (Cai et al., 2016). A simple bound is that, let $\delta=\hat{\beta}-\beta$, Γ can be replaced by

$$\max\|\mathcal{G}((I-\boldsymbol{U}^T\boldsymbol{U})\delta)\|/\|\mathcal{G}(\delta)\|$$

subject to $\hat{\beta}\in\mathcal{J}(\beta)$. With (C.10), we have $\|\mathcal{P}_W\mathcal{G}(\delta)\|_* \le 3\|\mathcal{P}_V\mathcal{G}(\delta)\|_*$. Denote $\sigma=\|\mathcal{G}(\delta)\|$, we know that $\sigma \ge \max\{\|\mathcal{P}_W\mathcal{G}(\delta)\|,\|\mathcal{P}_V\mathcal{G}(\delta)\|\}$ and $\|\mathcal{P}_V\mathcal{G}(\delta)\| \ge \|\mathcal{P}_V\mathcal{G}(\delta)\|_*/(2R)$. And simple algebra gives that

$$\max_{0<\sigma_i<\sigma, \sum_i \sigma = S} \sum_i \sigma_i^2 \leq S\sigma.$$

So let σ_i be singular values of $\mathcal{P}_V\mathcal{G}(\delta)$ or $\mathcal{P}_W\mathcal{G}(\delta)$, and $S = \|\mathcal{P}_V\mathcal{G}(\delta)\|_*$ or $\|\mathcal{P}_W\mathcal{G}(\delta)\|_*$,

$$\frac{\sigma}{\|\mathcal{P}_V\mathcal{G}(\delta)\|_F} \geq \sqrt{\frac{\|\mathcal{P}_V\mathcal{G}(\delta)\|_*}{2R\|\mathcal{P}_V\mathcal{G}(\delta)\|_*}} \geq \sqrt{1/2R}$$
$$\frac{\sigma}{\|\mathcal{P}_W\mathcal{G}(\delta)\|_F} \geq \sqrt{\frac{\|\mathcal{P}_V\mathcal{G}(\delta)\|_*}{2R\|\mathcal{P}_W\mathcal{G}(\delta)\|_*}} \geq \sqrt{1/6R}$$

the second last inequality comes from (C.10). Thus if $\|(I - \boldsymbol{U}^T\boldsymbol{U})\delta\| = O(1/\sqrt{R})\|\delta\|$, in other words, $\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F = O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F$, whenever δ in normal cone, we have

$$\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\| \leq \|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F \leq O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F \leq \|\mathcal{G}(\delta)\| \tag{C.11}$$

so $\Gamma < 1$. To get this, we need $\sqrt{T}/w(\mathcal{J}(\beta)) = O(\sqrt{R})$ where $T = O(pR^2 \log^2 n)$ (Vershynin, 2018, Thm 9.1.1), still not tight in R, but $O(\min\{n, R^2 \log^2 n\})$ is as good as Oymak & Ozay (2018) and better than Sarkar et al. (2019), which are $O(n)$ and $O(n^2)$ correspondingly. (Vershynin, 2018, Thm 9.1.1) is a bound in expectation, but it naively turns into high probability bound since $\Gamma \geq 0$. □

C.3 Bounding Γ*, where do we lose?*

The previous proof is not tight here.

$$\underbrace{\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\| \leq \|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F}_{\text{not tight}} \leq O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F \leq \|\mathcal{G}(\delta)\| \tag{C.12}$$

If we can show that, for all δ in the tangent cone (thus independent of $\boldsymbol{U}$), $\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\| = O(1/\sqrt{R})\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F$ for $\boldsymbol{U} \in \mathbb{R}^{O(R \log^2 n) \times n}$, then we can get the correct sample complexity. The difficulty is that, we do not know the distribution of $(I - \boldsymbol{U}^T\boldsymbol{U})\delta$. Let $M = I - \boldsymbol{U}^T\boldsymbol{U}$ and $g := M\delta$. Let $\tilde{g}$ be a Gaussian vector with same mean and covariance as g that will be studied later. We know that $g_i = \sum M_{ij}\delta_j$. Let $z_{ij} = U_{:,i}^T U_{:,j}$, u, v denote standard Gaussian vectors of dimension T, we have (the last equation: $i \neq j$)

$$\begin{aligned}
E((1 - z_{ii}^2)^2) &= E((1 - \frac{1}{T}u^T u)^2) \\
&= 1 - \frac{2}{T}\sum_{i=1}^{T} E(u_i^2) + \frac{1}{T^2}(\sum_{i=1}^{T} E(u_i^4) + \sum_{i\neq j}^{T} E(u_i^2 u_j^2)) = \frac{2}{T}. \\
E(z_{ij}^2) &= E((\frac{1}{T}u^T v)^2) \\
&= \frac{1}{T^2}E(\sum u_i^2 v_i^2) = \frac{1}{T}. \\
E(g_i) &= 0, \\
E(g_i^2) &= E((\sum M_{ij}\delta_j)^2) \\
&= \delta_i^2 E((1 - z_{ii}^2)^2) + \sum_{j\neq i} \delta_j^2 E(z_{ij}^2) + \sum_{j\neq k} \delta_j \delta_k E(M_{ij}M_{ik}) \\
&\leq \frac{1}{T}(\delta_i^2 + \|\delta\|^2). \\
E(g_i g_j) &= E((\sum M_{ik}\delta_k)(\sum M_{jl}\delta_l)) \\
&= \delta_i \delta_j E(M_{ij}M_{ji}) \\
&= \frac{1}{T}\delta_i \delta_j.
\end{aligned}$$

So

$$Cov(g) = \frac{1}{T}(\|\delta\|^2 I + \delta\delta^T).$$

The problem is that g is not Gaussian so even we know mean and variance it's still hard to deal with. Let's study Gaussian first. If $\tilde{g} = \tilde{g}_1 + \check{g}_2\delta$ where $\tilde{g}_1 \sim \mathcal{N}(0, \frac{\|\delta\|^2}{T}I)$ and $\check{g}_2 \sim \mathcal{N}(0, 1/T)$, then we have

$$\begin{aligned} E(\|\mathcal{G}(\tilde{g})\|) &\leq E(\|\mathcal{G}(\tilde{g}_1)\|) + E(|\check{g}_2|\|\mathcal{G}(\delta)\|) \\ &\leq \frac{1}{\sqrt{T}}(\|\delta\|\frac{\log n}{\sqrt{n}} + \|\mathcal{G}(\delta)\|) \\ &\leq \frac{1}{\sqrt{T}}(\underbrace{\frac{\sqrt{R}\log n}{\sqrt{n}}}_{\text{proven before}} + 1)\|\mathcal{G}(\delta)\| \\ &\leq \frac{2}{\sqrt{T}}\|\mathcal{G}(\delta)\|. \end{aligned}$$

If we have

$$P(\|\mathcal{G}(\tilde{g})\| > \alpha E(\|\mathcal{G}(\tilde{g})\|)) \leq \psi(\alpha),$$

then let $\alpha = \sqrt{T}/2$, we have

$$P(\|\mathcal{G}(\tilde{g})\| > E(\|\mathcal{G}(\delta)\|)) \leq \psi(\sqrt{T}/2)$$

We hope that $\psi(\alpha) = \exp(-O(\alpha^2))$ or $\log(\psi(\alpha)) = -O(\alpha^2)$. Then with a set of Gaussian width $\sqrt{R}\log n$, we use a union bound and have (if we ignore the difference between g and $\tilde{g}$)

$$P(\max_{\delta}\|\mathcal{G}(g)\| > \|\mathcal{G}(\delta)\|) \leq \psi(\sqrt{T}/2)\exp(O(R\log^2 n)) = \exp(O(R\log^2 n) + \log(\psi(\sqrt{T}/2))).$$

So if the derivation of a Gaussian vector can be applied to a non-Gaussian $g = (I - \boldsymbol{U}^T\boldsymbol{U})\delta$ with the same mean and variance, and $\|\mathcal{G}(g)\|$ is a subGaussian random variable, then we can get the tight bound.

C.4 Proof of suboptimal recovery guarantee with i.i.d. input

Theorem 6. *Suppose the system impulse response is h such that $h_t = 1$, $\forall t \geq 1$, which is order 1. The Gaussian width of the set $\{x \mid \|\mathcal{H}(h+x)\|_* \leq \|\mathcal{H}(h)\|_*\} \cap \mathbb{S}$ is lower bounded* *by $Cn^{1/6}$ for some constant C.*

Proof. We consider the Gaussian width $w(\Phi)$ defined in this specific case.

Let $V = \frac{1}{n}\mathbf{1}\mathbf{1}^T$, and

$$\mathcal{I}^*(h) = \{\mathcal{H}^*(V+W) | \mathbf{1}^T W = 0, W\mathbf{1} = 0, \|W\| \le 1\}$$

we have[2]

$$w(\Phi) = E(\min_{\lambda, W} \|\lambda \mathcal{H}^*(V+W) - g\|_2).$$

In the instance, $V = \frac{1}{n}\mathbf{1}\mathbf{1}^T$. and we take W such that $\|W\| \le 1$ and $W\mathbf{1} = W^T\mathbf{1} = 0$.

First, we note that

$$\begin{aligned}
&E(\min_{\lambda, W} \|\lambda \mathcal{H}^*(V+W) - g\|_2) \\
&= \frac{1}{2}\left(E(\min_{\lambda, W} \|\lambda \mathcal{H}^*(V+W) - g\|_2 \mid \mathbf{1}^T g \le 0) + E(\min_{\lambda, W} \|\lambda \mathcal{H}^*(V+W) - g\|_2 \mid \mathbf{1}^T g > 0)\right) \\
&\ge \frac{1}{2} E(\min_{\lambda, W} \|\lambda \mathcal{H}^*(V+W) - g\|_2 \mid \mathbf{1}^T g \le 0). \qquad \text{(C.13)}
\end{aligned}$$

Proof strategy: Based on the previous derivation, we focus on the case when $\mathbf{1}^T g \le 0$. Denote $z = \lambda \mathcal{H}^*(V+W) - g$, and the vector $z_{1:k}$ is the first 1 to k entries of z. Then we prove that

$$(1)\ \lambda \le \|z\|_2/\sqrt{n}, \quad (2)\ \|z_{1:1/\lambda}\|_2 = \Omega(\lambda^{-1/2}).$$

Then we have

$$\|z\|_2 = \Omega(\|z_{1:1/\lambda}\|_2) \overset{2}{=} \Omega(\lambda^{-1/2}) \overset{1}{=} \Omega((\|z\|_2/\sqrt{n})^{-1/2})$$

[2] We slightly change the definition of Gaussian width. We refer readers to (McCoy & Tropp, 2013, Thm 1). It is known to be as tight and the probability of failure is order constant if the number of measurements is smaller than order square of the quantity.

which suggests $\|z\|_2 = \Omega(n^{1/6})$.

Lemma 23. *Let g be a standard Gaussian vector of size $2n-1$ conditioned on $\mathbf{1}^T g \leq 0$. Let $z = \lambda\mathcal{H}^*(V+W) - g$ where $V = \frac{1}{n}\mathbf{1}\mathbf{1}^T$, and $W\mathbf{1} = W^T\mathbf{1} = 0$, $\|W\| \leq 1$. Then we have that $\lambda \leq \|z\|_2/\sqrt{n}$.*

We observe that $\mathbf{1}^T\mathcal{H}^*(X)$ is the summation of every entry in X for any matrix X. Thus $\mathbf{1}^T\mathcal{H}^*(W) = 0$ since $W\mathbf{1} = 0$. Conditioned on $\mathbf{1}^T g \leq 0$, we have

$$\mathbf{1}^T(\lambda\mathcal{H}^*(V+W) - g) \geq \lambda\mathbf{1}^T\mathcal{H}^*(V) = \lambda n.$$

And so that $\|\lambda\mathcal{H}^*(V+W) - g\|_2 \geq \lambda\sqrt{n}$. Then $\|z\|_2/\sqrt{n} \geq \lambda$, we have proven the first point.

Lemma 24. *Let g be a standard Gaussian vector of size $2n-1$ conditioned on $\mathbf{1}^T g \leq 0$. Let $z = \lambda\mathcal{H}^*(V+W) - g$ where $V = \frac{1}{n}\mathbf{1}\mathbf{1}^T$, and $W\mathbf{1} = W^T\mathbf{1} = 0$, $\|W\| \leq 1$. Let the vector $z_{1:k}$ is the first 1 to k entries of z.Then we have that $\|z_{1:1/\lambda}\|_2 = \Omega(\lambda^{-1/2})$.*

If $\|z\|_2 \leq \sqrt{n}$, we observe $z_{1:\sqrt{n}/\|z\|_2}$. When $i \leq \sqrt{n}/\|z\|_2$, the i-th entry of $\mathcal{H}^*(V+W)$, denoted as $(\mathcal{H}^*(V+W))_i$, is summation of $2i$ terns in V and W. Since these two matrices have bounded spectral norm 1, then every entry of V is $1/n$ and every entry of W is no bigger than 1. So

$$z_i = \lambda(\mathcal{H}^*(V+W))_i - g_i \in \pm(1+1/n)i\lambda - g_i \in \pm\frac{(1+1/n)i\|z\|_2}{\sqrt{n}} - g_i.$$

Thus

$$\begin{aligned}\|z_{1:\sqrt{n}/\|z\|_2}\|_2 &\geq -\frac{(1+1/n)\|z\|_2}{\sqrt{n}}\|[1,2,...,\sqrt{n}/\|z\|_2]\|_2 + \|g_{1:\sqrt{n}/\|z\|_2}\|_2 \\ &\geq -\frac{(1+1/n)n^{1/4}}{\sqrt{3}\|z\|_2^{1/2}} + \frac{n^{1/4}}{\|z\|_2^{1/2}}.\end{aligned}$$

Note that the first term is smaller than the second, so we have

$$\|z_{1:\sqrt{n}/\|z\|_2}\|_2 \geq C_1 \frac{n^{1/4}}{\|z\|_2^{1/2}}$$

for some constant $C_1 > 0$. Note this is the norm of a part of z, which is smaller than the norm of z, so we have

$$\frac{C_1 n^{1/4}}{\|z\|_2^{1/2}} \leq \|z\|_2$$

so that $\|z\|_2 = \Omega(n^{1/6})$, and we have bounded the quantity (C.13). □

C.5 Proof of least square spectral norm error

We first propose the following lemma.

Lemma 25. *Denote the discrete Fourier transform matrix by F. Denote $z_{(i)} \in \mathbb{R}^T, i = 1, ..., m$ as the noise that corresponds to each dimension of output. The solution $\hat{h}$ of*

$$\hat{h} := h + \bar{\boldsymbol{U}}^\dagger z = \min_{h'} \ \frac{1}{2}\|\bar{\boldsymbol{U}}h' - y\|_F^2. \tag{C.14}$$

obeys

$$\begin{aligned}\|\hat{h} - h\|_F &\leq \|z\|_F/\sigma_{\min}(\bar{\boldsymbol{U}})\\ \|\mathcal{H}(\hat{h} - h)\| &\leq \left\| \left[\|F\bar{\boldsymbol{U}}^\dagger z_{(1)}\|_\infty, ..., \|F\bar{\boldsymbol{U}}^\dagger z_{(m)}\|_\infty\right] \right\|.\end{aligned}$$

Proof. First we clarify the notation here. In regularization part, we only consider the MISO system, whereas we can prove the bound for MIMO system as well in least square. Here we assume the input is p dimension and output is m dimension, at each time. For the notation in (C.14), $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$, whose each row is the input in a time interval of length $2n-1$. The impulse response is $h \in \mathbb{R}^{(2n-1)p\times m}$ and output and noise are $y, z \in \mathbb{R}^{T\times m}$, where each column corresponds to one channel of the output. Each row of y is an output observation at a single time point. $z_{(i)} \in \mathbb{R}^T$ is a column of the noise, meaning one channel of the noise contaminating all observations at this channel.

(C.14) has close form solution and we have $\|\hat{h} - h\| = \|\bar{\boldsymbol{U}}^\dagger z\| \leq \|z\|/\sigma_{\min}(\bar{\boldsymbol{U}})$. To get the error bound in Hankel matrix, we can denote $\bar{z} = \bar{\boldsymbol{U}}^\dagger z = (\bar{\boldsymbol{U}}^T \bar{\boldsymbol{U}})^{-1} \bar{\boldsymbol{U}}^T z$, and

$$H_{\bar{z}} = \begin{bmatrix} \bar{z}_1 & \bar{z}_2 & \dots & \bar{z}_{2n-1} \\ \bar{z}_2 & \bar{z}_3 & \dots & \bar{z}_1 \\ \dots & & & \\ \bar{z}_{2n-1} & \bar{z}_1 & \dots & \bar{z}_{2n-2} \end{bmatrix}.$$

If $m = 1$, $\bar{z} \in \mathbb{R}^{(2n-1)p}$ is a vector (Krahmer et al., 2014, Section 4) proves that

$$H_{\bar{z}} = F^{-1} \text{diag} F \bar{z} F.$$

So the spectral norm error is bounded by $\|\text{diag} F \bar{z}\|_2 = \|F \bar{z}\|_\infty$.

If $m > 1$, all columns of z are independent, so $H_{\bar{z}}$ can be seen as concatenation of m independent noise matrices where each satisfies the previous derivation. □

Now we prove the following theorem.

Theorem 7. *Denote the solution to* (4.8) *as* $\hat{h}$. *Let* $\bar{\boldsymbol{U}} \in \mathbb{R}^{T \times (2n-1)p}$ *be input matrix obtained from multiple rollouts, with i.i.d. standard normal entries,* $y \in \mathbb{R}^{T \times m}$ *be the corresponding outputs and* $z \in \mathbb{R}^{T \times m}$ *be the noise matrix with i.i.d.* $\mathcal{N}(0, \sigma_z^2)$ *entries. Then the spectral norm error obeys* $\|\mathcal{H}(\hat{h} - h)\| \lesssim \sigma_z \sqrt{\frac{mnp}{T}} \log(np)$.

Proof. First let $m = 1$. The covariance of $F\bar{z} = F\bar{\boldsymbol{U}}^\dagger z$ is $F(\bar{\boldsymbol{U}}^\top \bar{\boldsymbol{U}})^{-1} F^\top$. If $T \gtrsim n$, it's proven in Vershynin (2018) that $\frac{TI}{2} \preceq \bar{\boldsymbol{U}}^\top \bar{\boldsymbol{U}} \preceq \frac{3TI}{2}$. Then $\frac{n}{2T} I \preceq F(\bar{\boldsymbol{U}}^\top \boldsymbol{U})^{-1} F^\top \preceq \frac{3n}{2T} I$. So $\|F\bar{z}\|_\infty$ should scale as $O(\sigma_z \sqrt{\frac{n}{T}} \log n)$, and then $\|\mathcal{H}(\bar{z})\|_2 \leq \|H_{\bar{z}}\|_2 \leq \|F\bar{z}\|_\infty = O(\sigma_z \sqrt{\frac{n}{T}} \log n)$.

If $m > 1$, then by concatenation we simply bound the spectral norm by m times MISO case. When $m > 1$, with previous discussion of concatenation, and each submatrix to be concatenated has the same distribution, so the spectral norm error is at most $\sqrt{m}$ times larger. □

C.6 Proof of model selection method

Theorem 8. *Consider the setting of Thm. 4. Sample T i.i.d. training rollouts $(\boldsymbol{U}, y)$ and T_{val} i.i.d. validation rollouts $(\boldsymbol{U}_{val}, y_{val})$. Set $\lambda^* = C\sigma\sqrt{\frac{pn}{T}}\log(n)$ which is the choice in Thm. 4. Fix failure probability $P \in (0,1)$. Suppose that:*

(a) There is a candidate $\hat{\lambda} \in \Lambda$ obeying $\lambda^*/2 \leq \hat{\lambda} \leq 2\lambda^*$.

(b) Validation set obeys $T_{val} \gtrsim \left(\frac{T\log^2(|\Lambda|/P)}{R\log^2(n)}\right)^{1/3}$.

Set $\bar{R} = \min(R^2, n)$. With probability at least $1-P$, Algorithm 2 achieves an estimation error equivalent to (4.7):

$$\|\mathcal{H}(\hat{h}-h)\| \lesssim \begin{cases} \sqrt{\frac{np}{\boldsymbol{snr}\times T}}\log(n), & \textit{if } T \gtrsim \bar{R}, \\ \sqrt{\frac{Rnp}{\boldsymbol{snr}\times T}}\log(n), & \textit{if } R \lesssim T \lesssim \bar{R}. \end{cases} \tag{4.9}$$

Proof. We select $\delta > 0$ such that $T_{\text{val}} \gtrsim \frac{1}{\delta^2}\log\frac{|\Lambda|}{P}$, and denote $a_1 = 1 - \frac{\delta}{\delta+2},\ a_2 = 1 + \frac{\delta}{\delta+2}$. Then we have $a_2/a_1 = 1+\delta$. Let $T_0 = \max\{1, T/(T_{\text{val}}R\log^2 n)\}$. We will show that

$$\begin{aligned} \frac{\|\hat{h}-h\|_2}{\sqrt{2}} &\leq \|\mathcal{H}(\hat{h}-h)\| \\ &\lesssim \begin{cases} (1+T_0^{1/4})\frac{a_2}{a_1}\sqrt{\frac{np}{\textbf{snr}\times T}}\log(n), & \text{if } T \gtrsim \min(R^2, n) \\ (1+T_0^{1/4})\frac{a_2}{a_1}\sqrt{\frac{Rnp}{\textbf{snr}\times T}}\log(n), & \text{if } R \lesssim T \lesssim \min(R^2, n). \end{cases} \end{aligned} \tag{C.15}$$

Note that we will need $T_0^{1/4}\delta \lesssim 1$ from our choice of T_{val} in the theorem, so the bound is sufficient for the theorem. This will be used later to calculate δ in (C.18).

We use the change of variable as in (4.6). We learn the parameter β with different λ, and get different estimations $\hat{\beta}$ which is a function of λ. To be more explicit, let $\hat{\beta}(\lambda)$ be the estimator associated with a certain regularization parameter λ. Among all the estimators, we choose the solution with the smallest validation error, which is denoted as

$$\hat{\beta} = \text{argmin}_{\hat{\beta}(\lambda)} \|\boldsymbol{U}_{\text{val}} \hat{\beta}(\lambda) - y_{\text{val}}\|_2^2$$

Denote the noise in validation data as ξ_{val}. We have that

$$\begin{aligned} \|\boldsymbol{U}_{\text{val}} \hat{\beta} - y_{\text{val}}\|_2^2 &= \|\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta) - \xi_{\text{val}}\|_2^2 \\ &= \|\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\|_2^2 + \|\xi_{\text{val}}\|_2^2 - 2\xi_{\text{val}}^\top \boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta). \end{aligned} \tag{C.16}$$

In this formulation, $\|\xi_{\text{val}}\|_2^2$ in (C.16) is regarded as fixed among all validation instances, and we study the other two terms. Since $\boldsymbol{U}_{\text{val}}$ is normalized that each entry is i.i.d. $\mathcal{N}(0, 1/T_{\text{val}})$, we have $\boldsymbol{E}\|\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\|_2^2 = \|\hat{\beta} - \beta\|_2^2$.

The quantity $\xi_{\text{val}}^\top \boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)$ is zero mean and we know that $\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta) \sim \mathcal{N}(0, \frac{\|\hat{\beta}-\beta\|_2^2}{T_{\text{val}}} I)$. Thus the variance of $\xi_{\text{val}}^\top \boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)$ is bounded by $O(\sigma_{\xi_{\text{val}}}^2 \|\hat{\beta} - \beta\|_2^2 / T_{\text{val}})$ (the distribution of the inner product is sub-exponential). We know that

$$\|\hat{\beta} - \beta\|_2 \approx \sqrt{\frac{R \log^2 n}{T}} \|\xi\|_2 = \sqrt{\frac{R \log^2 n}{T}} \sqrt{T_{\text{val}}} \sigma_{\xi_{\text{val}}}.$$

Case 1: If $T_{\text{val}} \gtrsim \frac{T}{R \log^2(n)}$, we have that $\|\hat{\beta} - \beta\|_2 \gtrsim \sigma_{\xi_{\text{val}}}$.

Suppose the number of validated parameters λ is $|\Lambda|$ and we need to choose the size of validation data. With different validation data size T_{val}, the variance of $\|\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\|_2^2$ decreases with rate $1/T_{\text{val}}$.

We fix factors a_1, a_2, such that with high probability, for all choices of λ, $\|\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\|_2^2 - 2\xi_{\text{val}}^\top \boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)$ is in the set $(a_1\|\hat{\beta} - \beta\|_2^2, a_2\|\hat{\beta} - \beta\|_2^2)$. We know that: the terms $\|\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\|_2^2$ and $2\xi_{\text{val}}^\top \boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)$ are subexponential; The mean of $\|\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\|_2^2$ is $\|\hat{\beta} - \beta\|_2^2$ and the variance is $O(\|\hat{\beta} - \beta\|_2^4 / T_{\text{val}})$; The mean of $2\xi_{\text{val}}^\top \boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)$ is 0 and the variance is $O(\|\hat{\beta} - \beta\|_2^4 / T_{\text{val}})$ (Note that $\|\hat{\beta} - \beta\|_2 \gtrsim \sigma_{\xi_{\text{val}}}$ in this case).

By Bernstein bound (Vershynin, 2010, Prop. 5.16), we know that the probability that the quantity of (C.17) is not between $(a_1, a_2) \cdot \|\hat{\beta} - \beta\|_2^2$ is $\exp(-\min_i (a_i - 1)^2 T_{\text{val}})$ where $(a_i - 1)^2 \approx \delta^2$.

Hence there exists a constant c such that for every choice of λ,

$$\begin{aligned}\mathbf{Pr}\Big(&\big|\|\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\|_2^2 - 2\xi_{\text{val}}^\top \boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\big| \\ &\notin (a_1, a_2) \cdot \|\hat{\beta} - \beta\|_2^2\Big) \\ &< \exp(-c\delta^2 T_{\text{val}}). \end{aligned} \tag{C.17}$$

We choose probability P that any of the event in (C.17) happens. If all $|\Lambda|$ validations corresponding to λ_i succeed, then we use the union bound on (C.17) and solve for $|\Lambda| \exp(-c\delta^2 T_{\text{val}}) < P$. Thus we set $T_{\text{val}} = \max\{\frac{T}{R\log^2(n)}, \frac{1}{c\delta^2} \log \frac{|\Lambda|}{P}\}$. so that (4.9) holds with probability $1 - P$.

Case 2: If $T_{\text{val}} \lesssim \frac{T}{R\log^2(n)}$, then we denote $T_0 = T/(T_{\text{val}} R \log^2 n)$, with similar derivation as above, we know that the mean of $\|\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\|_2^2$ is $\|\hat{\beta} - \beta\|_2^2$ and the variance is $O(\|\hat{\beta} - \beta\|_2^4 / T_{\text{val}})$; The mean of $2\xi_{\text{val}}^\top \boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)$ is 0 and the variance is $O(T_0 \|\hat{\beta} - \beta\|_2^4 / T_{\text{val}})$. Thus, similar to (C.17),

$$\begin{aligned}\mathbf{Pr}\Big(&\big|\|\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\|_2^2 - 2\xi_{\text{val}}^\top \boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\big| \\ &\notin (a_1, a_2) \cdot \sqrt{T_0} \|\hat{\beta} - \beta\|_2^2\Big) \\ &< \exp(-c\delta^2 T_{\text{val}}). \end{aligned}$$

The following steps are same as the first case, and the error is multiplied by $T_0^{1/4}$ compared to the first case.

At the end, we will need to argue about the lower bound for T_{val}. We used two inequalities in the proof above:

$$T_{\text{val}} \gtrsim \frac{1}{\delta^2} \log(\frac{|\Lambda|}{P}), \ T_0^{1/4} \delta \lesssim 1.$$

They are equivalent to

$$T_{\text{val}} \gtrsim \frac{1}{\delta^2} \log(\frac{|\Lambda|}{P}),\ T_{\text{val}} \gtrsim \frac{\delta^4 T}{R \log^2(n)}. \tag{C.18}$$

Setting the right hand side to be equal, we have

$$\delta^2 = \left(T^{-1} \log(\frac{|\Lambda|}{P}) R \log^2(n)\right)^{1/3}.$$

Plugging it into any lower bound for T_{val} in (C.18), we get the bound in the main theorem. □

Appendix D

APPENDIX OF CHAPTER 5

D.1 Numerical verification of inductive bias for representation learning

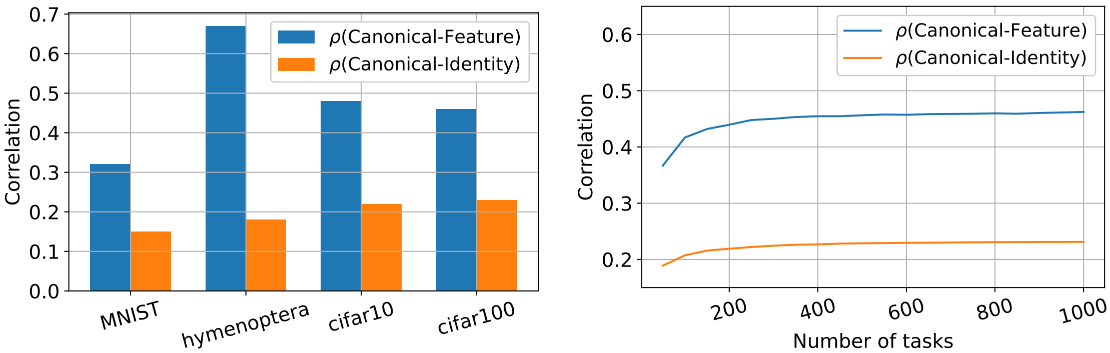


Figure D.1: (a) Alignment of feature-task on image classification models. For MNIST, we train 45 linear pairwise classifiers between each two classes. We apply the pretrained ResNet classification model on the other three datasets, compute the (last layer) feature/task covariances and get the alignments. The alignment is a measure of correlation which is denoted by ρ here. (b) We use the cifar100 dataset, take the pretrained ResNet18 network and vary the number of tasks (i.e., varying the number of output classes of the neural net, also equivalent to number of rows of the last layer matrix B defined below). The alignments increase with number of tasks.

We have figures with experiments on a few image datasets. We take the pretrained ResNet18 neural network, and feed the images into it. For every image, we take the last (closest to output) layer output as the feature $\boldsymbol{x}$, which is of dimension $d = 512$. The weights of the last layer are the tasks, which is a $T \times d$ matrix (We call it B). $T = 1000$, each row of B is a task vector. Then $Bx \in \mathbb{R}^T$ generates the label, whose each entry corresponds to each class. We calculate the feature and task covariance, as well as the alignments defined in Sec. 5.4. We can clearly see the inductive bias of every dataset.

D.2 Analysis of optimal representation

D.2.1 Proof of Observation 1 and equivalent noise

Observation 3. Let $\mathbf{\Lambda} \in \mathbb{R}^{d\times R}$, $\boldsymbol{X} \in \mathbb{R}^{n_2\times d}$ and $\boldsymbol{y} \in \mathbb{R}_2^n$, and define

$$\hat{\beta} = \mathbf{\Lambda}(\boldsymbol{X}\mathbf{\Lambda})^{\dagger}\boldsymbol{y}, \tag{D.1}$$

$$\hat{\beta}_1 = \lim_{t\to 0} \operatorname{argmin}_{\beta}\|\boldsymbol{X}\beta - \boldsymbol{y}\|^2 + t\beta^{\top}(\mathbf{\Lambda}\mathbf{\Lambda}^{\top})^{\dagger}\beta \tag{D.2}$$

Then $\hat{\beta}_1 = \hat{\beta}$.

Proof. Denote the SVD $(\boldsymbol{X}\mathbf{\Lambda})^{\top} = \boldsymbol{U}\mathbf{\Sigma}\boldsymbol{V}^{\top}$, where $\boldsymbol{U} \in \mathbb{R}^{R\times R}, \mathbf{\Sigma} \in \mathbb{R}^{R\times n_2}, \boldsymbol{V} \in \mathbb{R}^{n_2\times n_2}$.

$$\begin{aligned}
\hat{\beta}_1 &= \lim_{t\to 0} \operatorname{argmin}_{\beta}\|\boldsymbol{X}\beta - \boldsymbol{y}\|^2 + t\beta^{\top}(\mathbf{\Lambda}\mathbf{\Lambda}^{\top})^{\dagger}\beta \\
&= \lim_{t\to 0}(\boldsymbol{X}^{\top}\boldsymbol{X} + t(\mathbf{\Lambda}\mathbf{\Lambda}^{\top})^{\dagger})^{-1}\boldsymbol{X}\boldsymbol{y} \\
&= \lim_{s\to\infty} s\mathbf{\Lambda}(s\mathbf{\Lambda}^{\top}\boldsymbol{X}^{\top}\boldsymbol{X}\mathbf{\Lambda} + I)^{-1}\mathbf{\Lambda}^{\top}\boldsymbol{X}^{\top}\boldsymbol{y} \\
&= \lim_{s\to\infty} s\mathbf{\Lambda}(s\boldsymbol{U}\mathbf{\Sigma}\boldsymbol{V}^{\top}\boldsymbol{V}\mathbf{\Sigma}^{\top}\boldsymbol{U}^{\top} + I)^{-1}\boldsymbol{U}\mathbf{\Sigma}\boldsymbol{V}^{\top}\boldsymbol{y} \\
&= \lim_{s\to\infty} s\mathbf{\Lambda}(s\boldsymbol{U}\operatorname{diag}(\mathbf{\Sigma}^{\top}\mathbf{\Sigma} + I_{n_2}, I_{R-n_2})\boldsymbol{U}^{\top})^{-1}\boldsymbol{U}\mathbf{\Sigma}\boldsymbol{V}^{\top}\boldsymbol{y} \\
&= \lim_{s\to\infty} \mathbf{\Lambda}\boldsymbol{U}(\operatorname{diag}(\mathbf{\Sigma}^{\top}\mathbf{\Sigma}, I_{R-n_2}/s))^{-1}\mathbf{\Sigma}\boldsymbol{V}^{\top}\boldsymbol{y}. \\
&= \mathbf{\Lambda}(\boldsymbol{X}\mathbf{\Lambda})^{\dagger}\boldsymbol{y}
\end{aligned}$$

□

The risk of $\hat{\beta}$ is given by

$$\operatorname{risk}(\hat{\beta}) = \boldsymbol{E}(y - \boldsymbol{x}^{\top}\hat{\beta}) = \boldsymbol{E}(\hat{\beta} - \beta)^{\top}\mathbf{\Sigma}_F(\hat{\beta} - \beta) + \sigma^2.$$

In Sec. D.2.2, we study the asymptotic optimal representation. Below, we characterize the properties of the problem for fixed β and arbitrary input covariance $\boldsymbol{\Sigma}_F$. We first go over this and then discuss how to obtain the optimal representation $\boldsymbol{\Lambda}^*$ minimizing test risk.

Remark 10. ***Projection onto R dimensional subspace.*** *For the remaining proof after this part, we will mainly analyze the relation between $\boldsymbol{\Lambda}_R$ and $\boldsymbol{\theta}$ in Thm. 9, which lie in an R dimensional subspace. Here we will build the connection from the d dimensional problem to R dimensional, mainly computing the equivalent noise below. The equivalent noise consists of original noise and the extra noise caused by PCA truncation.*

Let $\boldsymbol{x}_R$ be the projection of $\boldsymbol{x}$ onto the R-dimensional subspace spanned by columns of $\boldsymbol{U}_1$, and $\boldsymbol{x}_{R^\perp}$ is the projection of $\boldsymbol{x}$ onto the orthogonal complement. Namely, $\boldsymbol{x}_R = \boldsymbol{U}_1^\top \boldsymbol{x} \in \mathbb{R}^R$ and $\boldsymbol{x}_{R^\perp} = \boldsymbol{U}_2^\top \boldsymbol{x} \in \mathbb{R}^{(d-R)}$. Similarly we can define β_R and $\beta_{R^\perp}$. Thus,

$$y = \boldsymbol{x}^\top \beta + \varepsilon = \boldsymbol{x}_R^\top \beta_R + \boldsymbol{x}_{R^\perp}^\top \beta_{R^\perp} + \varepsilon \tag{D.3}$$

We can treat $\varepsilon_R = \boldsymbol{x}_{R^\perp}^\top \beta_{R^\perp} + \varepsilon$ as the new noise, and try to solve for β_R. Then define $\boldsymbol{\Sigma}_{T,R^\perp}$ as the matrix containing the same eigenvectors as $\boldsymbol{\Sigma}_T$ and the top R eigenvalues are zeroed out, our noise variance becomes $\sigma_R^2 = \sigma^2 + \boldsymbol{E}(\|\boldsymbol{x}_{R^\perp}\|^2 \|\beta_{R^\perp}\|^2) = \sigma^2 + \mathbf{tr}(\tilde{\boldsymbol{\Sigma}}_T) - \mathbf{tr}(\tilde{\boldsymbol{\Sigma}}_T^R)$ in our algorithm. If we are still in overparameterized regime, namely $R > n_2$, then we define optimal representation on top of it.

In summary, the R-SVD truncation reduces the search space of $\boldsymbol{\Lambda}$ into R dimensional space, where the covariance of the noise in $\boldsymbol{y}$ increases from $\sigma^2 I$ to $\sigma_R^2 I$.

D.2.2 Distributional characterization of least norm solution

In this part, for simplicity of discussion, we focus on the R dimensional space while omitting the projection step, and the equivalence of a diagonal eigen-weighting matrix $\boldsymbol{\Lambda}_R \in \mathbb{R}^{R\times R}$ and $\boldsymbol{\theta} \in \mathbb{R}^R$ in Thm. 9. Here, we assume a truncated feature matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n\times R}$ where the feature is projected into an R dimensional space.

Define $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n\times R}$, $\tilde{\boldsymbol{y}} \in \mathbb{R}^n$. We study the following least norm solution of the least squares problem

$$\hat{\beta} = \arg\min_{\beta'} \ \|\beta'\|, \quad \text{s.t.}, \ \tilde{\boldsymbol{X}}\beta' = \tilde{\boldsymbol{y}} \tag{D.4}$$

Assumption 13. *Assume the rows of $\tilde{\boldsymbol{X}}$ are independently drawn from $\mathcal{N}(0, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}})$. We focus on a double asymptotic regime where $R, n \to \infty$ at fixed overparameterization ratio $\kappa := R/n > 0$.*

Assumption 14. *The covariance matrix $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}$ is diagonal and there exist constants $\Sigma_{\min}, \Sigma_{\max} \in (0, \infty)$ such that: $0 \prec \Sigma_{\min} I \preceq \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}} \preceq \Sigma_{\max} I$.*

Assumption 15. *The joint empirical distribution of $\{(\lambda_i(\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}), \beta_i)\}_{i \in [R]}$ converges in Wasserstein-k distance to a probability distribution μ on $\mathbb{R}_{>0} \times \mathbb{R}$ for some $T \geq 4$. That is $\frac{1}{R} \sum_{i \in [R]} \delta_{(\lambda_i(\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}), \beta_i)} \overset{W_k}{\Longrightarrow} \mu$.*

Definition 5 (Asymptotic distribution characterization – Overparameterized regime)**.** *Thrampoulidis et al. (2015) Let random variables $(\Sigma, B) \sim \mu$ (where μ is defined in Assumption 15) and fix $\kappa > 1$. Define parameter ξ as the unique positive solution to the following equation*

$$\mathbb{E}_{\mu}\left[\left(1 + (\xi \cdot \Sigma)^{-1}\right)^{-1}\right] = \kappa^{-1}. \tag{D.5}$$

Define the positive parameter γ as follows:

$$\gamma := \Big(\sigma^2 + \mathbb{E}_{\mu}\Big[\frac{B^2\Sigma}{(1 + \xi\Sigma)^2}\Big]\Big) \Big/ \Big(1 - \kappa\, \mathbb{E}_{\mu}\Big[\frac{1}{\left(1 + (\xi\Sigma)^{-1}\right)^2}\Big]\Big). \tag{D.6}$$

With these and $H \sim \mathcal{N}(0, 1)$, define the random variable

$$X_{\kappa,\sigma^2}(\Sigma, B, H) := \Big(1 - \frac{1}{1 + \xi\Sigma}\Big) B + \sqrt{\kappa} \frac{\sqrt{\gamma}\, \Sigma^{-1/2}}{1 + (\xi\Sigma)^{-1}} H, \tag{D.7}$$

and let Π_{κ,σ^2} be its distribution.

Theorem 15 (Asymptotic distribution characterization – Overparameterized linear Gaussian problem)**.** *Thrampoulidis et al. (2015) Fix $\kappa > 1$ and suppose Assumptions 14 and 15 hold. Let*

$$\frac{1}{R}\sum_{i=1}^{R}\delta_{\sqrt{R}\hat{\beta}_i,\sqrt{R}\beta_i,\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}_{i,i}}}$$

be the joint empirical distribution of $(\sqrt{R}\hat{\beta},\sqrt{R}\beta,\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}})$ and it converges to a fixed distribution as dimension grows. Let $f:\mathbb{R}^3\to\mathbb{R}$ be a function in PL(2)*. We have that*

$$\frac{1}{R}\sum_{i=1}^{R} f(\sqrt{R}\hat{\beta}_i,\sqrt{R}\beta_i,\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}_{i,i}}) \xrightarrow{P} \mathbb{E}\left[f(X_{\kappa,\sigma^2},B,\Sigma)\right]. \tag{D.8}$$

In particular, the risk is given by

$$risk(\hat{\beta}_n) \xrightarrow{P} \mathbb{E}[\Sigma(B-X_{\kappa,\sigma^2})]+\sigma_0^2 \tag{D.9}$$

$$=\mathbb{E}[\frac{\Sigma}{(1+\xi\Sigma)^2}B^2+\frac{\kappa\gamma}{(1+(\xi\Sigma)^{-1})^2}]+\sigma_0^2. \tag{D.10}$$

D.2.3 Finding Optimal Representation

Now, for simplicity (and actually without losing generality) assume $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = \boldsymbol{I}$. This means that empirical measure of $\boldsymbol{\Sigma}_F$ trivially converges to $\Sigma = 1$. With the representation $\boldsymbol{\Lambda}^*$ with asymptotic distribution Λ, the ML problem has the following mapping

$$\beta \to \boldsymbol{\Lambda}_R^{-1}\beta \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}} \to \boldsymbol{\Lambda}_R\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\boldsymbol{\Lambda}_R.$$

This means the empirical measure converges to the following mapped distributions

$$B \to \bar{B} = \Lambda^{-1}B \quad \text{and} \quad \Sigma = 1 \to \bar{\Sigma} = \Lambda^2\Sigma = \Lambda^2.$$

Our question: Craft the optimal distribution Λ to minimize the representation learning risk. Specifically, for a given (B, Λ) pair, we know from the theorem above that

$$\text{risk}^{\boldsymbol{\Lambda}_R}(\hat{\beta}_n) \xrightarrow{P} \mathbb{E}[\frac{\bar{\Sigma}}{(1+\xi\bar{\Sigma})^2}\bar{B}^2 + \frac{\kappa\gamma}{(1+(\xi\bar{\Sigma})^{-1})^2}] + \sigma_0^2 \tag{D.11}$$

$$= \mathbb{E}[\frac{B^2}{(1+\xi\Lambda^2)^2} + \frac{\kappa\gamma}{(1+(\xi\Lambda^2)^{-1})^2}] + \sigma_0^2. \tag{D.12}$$

Thus, the optimal weighting strategy (asymptotically) is given by the distribution

$$\Lambda^* = \arg\min_{\Lambda} \mathbb{E}[\frac{B^2}{(1+\xi\Lambda^2)^2} + \frac{\kappa\gamma}{(1+(\xi\Lambda^2)^{-1})^2}],$$

where γ, ξ are strictly positive scalars that are also functions of Λ.

D.2.4 Non-asymptotic Analysis (for simpler insights)

We apply the discussion iin Sec. D.2.2 non-asymptotically in few-shot learning. Remember we define $\boldsymbol{X} \in \mathbb{R}^{n_2 \times R}, \boldsymbol{y} \in \mathbb{R}^{n_2}$, each row of $\boldsymbol{X}$ is independently drawn from $\mathcal{N}(0, \boldsymbol{\Sigma}_F)$. We study the following least norm solution of the least squares problem

$$\hat{\beta} = \arg\min_{\beta'} \ \|\beta'\|, \quad \text{s.t.}, \boldsymbol{X}\beta' = \boldsymbol{y}. \tag{D.13}$$

Definition 6 (Non-asymptotic distribution characterization)**.** *Set $\kappa = R/n_2 > 1$. Given $\sigma_0 > 0$, covariance $\boldsymbol{\Sigma}_F$ and latent vector β and define the unique non-negative terms $\xi, \gamma, \boldsymbol{z} \in \mathbb{R}^R$ and $\boldsymbol{\phi} \in \mathbb{R}^R$ as follows:*

$$\xi > 0 \quad \textit{is the solution of} \quad \kappa^{-1} = R^{-1}\sum_{i=1}^{R}\left(1 + (\xi\boldsymbol{\Sigma}_{F,i})^{-1}\right)^{-1},$$

$$\gamma = \frac{\sigma_0^2 + \frac{1}{R}\sum_{i=1}^{R}\frac{\boldsymbol{\Sigma}_{F,i}\beta_i^2}{(1+\xi\boldsymbol{\Sigma}_F)^2}}{1 - \frac{\kappa}{R}\sum_{i=1}^{R}(1+(\xi\boldsymbol{\Sigma}_{F,i})^{-1})^{-2}}.$$

Let $\boldsymbol{h} \sim \mathcal{N}(0, \boldsymbol{I}/R)$. The non-asymptotic distributional prediction is given by the following *random vector*

$$\hat{\beta}(\boldsymbol{\Sigma}_F) = \frac{1}{1+(\xi\boldsymbol{\Sigma}_F)^{-1}} \odot \beta + \frac{\sqrt{\kappa\gamma}\boldsymbol{\Sigma}_F^{-1/2}}{1+(\xi\boldsymbol{\Sigma}_F)^{-1}} \odot \boldsymbol{h}.$$

Note that, the above formulas can be slightly simplified to have a cleaner look by introducing an additional variable $\boldsymbol{z} = \frac{1}{1+(\xi\boldsymbol{\Sigma}_F)^{-1}}$.

Also note that, the terms in the non-asymptotic distribution characterization and asymptotic distribution characterization have one to one correspondence. Non-asymptotic distribution characterization is essentially a discretized version of asymptotic DC where instead of expectations (which is integral over pdf) we have summations.

Now, we can use this distribution to predict the test risk by using Def. 6 in the risk expression.

Going back to representation question, without losing generality, assume $\boldsymbol{\Sigma}_F = \boldsymbol{I}$ and let us find optimal $\boldsymbol{\Lambda}_R$. Then

$$\hat{\beta} = \boldsymbol{\Lambda}_R \left[\frac{1}{1+(\xi\boldsymbol{\Lambda}_R^2)^{-1}} \odot \boldsymbol{\Lambda}_R^{-1}\beta + \frac{\sqrt{\kappa\gamma}\boldsymbol{\Lambda}_R^{-1}}{1+(\xi\boldsymbol{\Lambda}_R^2)^{-1}} \odot \boldsymbol{h} \right].$$

The risk is given by (using $\boldsymbol{h} \sim \mathcal{N}(0, \boldsymbol{I}_p)$)

$$\begin{aligned} \text{risk}^{\boldsymbol{\Lambda}_R}(\hat{\beta}_n) - \sigma_0^2 &= \mathbb{E}[(\hat{\beta}-\beta)^\top \boldsymbol{\Sigma}_F (\hat{\beta}-\beta)] && \text{(D.14)} \\ &= \sum_{i=1}^{R} \frac{\boldsymbol{\Sigma}_{T,i}}{(1+\xi(\boldsymbol{\Lambda}_{R,i})^2)^2} + \sum_{i=1}^{R} \frac{\kappa\gamma}{(1+(\xi(\boldsymbol{\Lambda}_{R,i})^2)^{-1})^2}. && \text{(D.15)} \end{aligned}$$

Here, note that ξ is function of $\boldsymbol{\Lambda}^*$ and γ is function of $\beta, \boldsymbol{\Lambda}^*$. If we don't know $\boldsymbol{\Sigma}_T$, we use the estimation from representation learning $\hat{\boldsymbol{\Sigma}}_T$ instead.

To find the optimal representation, we will solve the following optimization problem that minimizes the risk.

$$\min_{\mathbf{\Lambda}^*} \quad \textstyle\sum_{i=1}^{R} \dfrac{\beta_i^2}{(1+\xi(\mathbf{\Lambda}_{R,i})^2)^2} + \textstyle\sum_{i=1}^{R} \dfrac{\kappa\gamma}{(1+(\xi(\mathbf{\Lambda}_{R,i})^2)^{-1})^2}$$

$$\text{s.t.} \qquad \kappa^{-1} = \frac{1}{R}\textstyle\sum_{i=1}^{R}(1+(\xi(\mathbf{\Lambda}_{R,i})^2)^{-1})^{-1} \tag{D.16}$$

$$\gamma = \frac{\sigma_0^2 + \sum_{i=1}^{R} \frac{\beta_i^2}{(1+\xi(\mathbf{\Lambda}_{R,i})^2)^2}}{1 - \frac{\kappa}{R}\sum_{i=1}^{R}(1+(\xi(\mathbf{\Lambda}_{R,i})^2)^{-1})^{-2}}.$$

So we plug in the expression of γ and get

$$\kappa\gamma = \frac{\sigma_0^2 + \frac{1}{R}\sum_{i=1}^{R} \frac{\beta_i^2}{(1+\xi(\mathbf{\Lambda}_{R,i})^2)^2}}{\kappa^{-1} - \frac{1}{R}\sum_{i=1}^{R}(1+(\xi(\mathbf{\Lambda}_{R,i})^2)^{-1})^{-2}} = \frac{R\sigma_0^2 + \sum_{i=1}^{R} \frac{\beta_i^2}{(1+\xi(\mathbf{\Lambda}_{R,i})^2)^2}}{\sum \frac{\xi(\mathbf{\Lambda}_{R,i})^2}{(1+\xi(\mathbf{\Lambda}_{R,i})^2)^2}}. \tag{D.17}$$

Let $\boldsymbol{\theta}_i = \frac{\xi(\mathbf{\Lambda}_{R,i})^2}{1+\xi(\mathbf{\Lambda}_{R,i})^2}$, then the objective function becomes

$$\sum_{i=1}^{R} \mathbf{\Sigma}_{T,i}(1-\boldsymbol{\theta}_i)^2 + (\sum_{i=1}^{R} \boldsymbol{\theta}_i^2)\frac{R\sigma_0^2 + \sum \mathbf{\Sigma}_{T,i}(1-\boldsymbol{\theta}_i)^2}{\sum_{i=1}^{R}\boldsymbol{\theta}_i(1-\boldsymbol{\theta}_i)} = \frac{n_2(\sum_{i=1}^{R}\mathbf{\Sigma}_{T,i}(1-\boldsymbol{\theta}_i)^2) + R\sigma_0^2(\sum_{i=1}^{R}\boldsymbol{\theta}_i^2)}{n_2 - \sum_{i=1}^{R}\boldsymbol{\theta}_i^2}$$

such that $0 \le \boldsymbol{\theta}_i < 1$ and $\sum_{i=1}^{R}\boldsymbol{\theta}_i = \frac{R}{\kappa} = n_2$. This quantity is same as the objective (D.16). We divide this quantity by d to get the risk function, which is same as the definition of f in (5.4).

D.2.5 Solving the optimization problem.

Here, we propose the algorithm for minimizing $f(\boldsymbol{\theta})$. We explore the KKT condition for its optimality.

The objective function is

$$f(\boldsymbol{\theta}) = \sum_{i=1}^{R} \mathbf{\Sigma}_{T,i}(1-\boldsymbol{\theta}_i)^2 + (\sum_{i=1}^{R} \boldsymbol{\theta}_i^2)\frac{R\sigma_0^2 + \sum \mathbf{\Sigma}_{T,i}(1-\boldsymbol{\theta}_i)^2}{\sum_{i=1}^{R}\boldsymbol{\theta}_i(1-\boldsymbol{\theta}_i)}. \tag{D.18}$$

Lemma 26. *Let $C, S, V \in \mathbb{R}$. Define*

$$\phi(\mathbf{\Sigma}_{T,i}; C, V, S) := \frac{Cp(R - n_2 - S)^2}{2n_2(V + R\sigma_0^2 + (R - n_2 - S){\mathbf{\Sigma}_{T,i}}^2)}$$

and we find the root of the following equations:

$$\begin{aligned} \sum_{i=1}^{R} \phi(\mathbf{\Sigma}_{T,i}; C, V, S) &= R - n_2, \\ \sum_{i=1}^{R} \phi^2(\mathbf{\Sigma}_{T,i}; C, V, S) &= S - (2n_2 - R), \\ \sum_{i=1}^{R} \mathbf{\Sigma}_{T,i}\phi^2(\mathbf{\Sigma}_{T,i}; C, V, S) &= V. \end{aligned}$$

Let $\boldsymbol{\theta}_i = 1 - \phi(\mathbf{\Sigma}_{T,i}; C^*, V^*, S^*)$ where C^*, V^*, S^* are the roots, then

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}'} \ f(\boldsymbol{\theta}'), \quad s.t.,\ 0 \leq \boldsymbol{\theta}' < 1,\ \sum_{i=1}^{R} \boldsymbol{\theta}'_i = n_2.$$

Proof. Define $s = \sum_{i=1}^{R} \boldsymbol{\theta}_i^2$, $\phi_i = 1 - \boldsymbol{\theta}_i$. Define $Q = \frac{1}{R} \sum_{i=1}^{R} \mathbf{\Sigma}_{T,i}\phi_i^2$. Then

$$\begin{aligned} f(\phi) &= \sum_{i=1}^{R} \mathbf{\Sigma}_{T,i}\phi_i^2 + \frac{s}{n_2 - s}(R\sigma_0^2 + \sum_{i=1}^{R} \mathbf{\Sigma}_{T,i}\phi_i^2) \\ &= R(Q + \frac{s}{n_2 - s}(\sigma_0^2 + Q)) \\ &= \frac{Rn_2}{R - n_2 - \sum_{i=1}^{R} \phi_i^2}(Q + \sigma_0^2). \end{aligned}$$

The last line uses

$$s = \sum_{i=1}^{R}(1 - \phi^2) = R - 2\sum_{i=1}^{R} \phi_i + \sum_{i=1}^{R} \phi_i^2 = R - 2(R - n_2) + \sum_{i=1}^{R} \phi_i^2 = 2n_2 - R + \sum_{i=1}^{R} \phi_i^2.$$

Now define $\sum_{i=1}^{R}\phi_i^2 = S$, and we compute the gradient of f, we have

$$\frac{df}{R\phi_i} = \left(2n_2(\sum_{j=1}^{R}\boldsymbol{\Sigma}_{Tj}\phi_j^2 + (R-n_2-s)\boldsymbol{\Sigma}_{T,i}) + 2Rn_2\sigma_0^2\right)\phi_i.$$

Suppose $0<\phi_i<1$, then we need $\frac{df}{R\phi_i}$ equal to each other for all i. Suppose $\frac{df}{R\phi_i} = C$, and denote $\sum \boldsymbol{\Sigma}_{Tj}\phi_j^2 = V$, we can solve for ϕ_i from $\frac{df}{R\phi_i} = C$ as

$$\phi_i = \frac{Cd(R-n_2-S)^2}{2n_2(V+R\sigma_0^2+(R-n_2-S){\boldsymbol{\Sigma}_{T,i}}^2)} := \phi(\boldsymbol{\Sigma}_{T,i};C,V,S). \tag{D.19}$$

We define the function $\phi(\boldsymbol{\Sigma}_{T,i};C,V,S)$ as above, and use the fact that

$$\sum_{i=1}^{R}\phi(\boldsymbol{\Sigma}_{T,i};C,V,S) = R-n_2,$$
$$\sum_{i=1}^{R}\phi^2(\boldsymbol{\Sigma}_{T,i};C,V,S) = S-(2n_2-R),$$
$$\sum_{i=1}^{R}\boldsymbol{\Sigma}_{T,i}\phi^2(\boldsymbol{\Sigma}_{T,i};C,V,S) = V.$$

We can solve[1] C, V, S and retrieve ϕ_i by (D.19). $\boldsymbol{\theta}_i = 1-\phi_i$. □

D.3 Analysis of MoM estimators

D.3.1 Covariance estimator

We will first present the estimation error of the feature covariance $\boldsymbol{\Sigma}_F$. Note that if $\boldsymbol{\Sigma}_F$ is fully aligned with $\boldsymbol{\Sigma}_T$, e.g., $\boldsymbol{\Sigma}_F = \boldsymbol{\Sigma}_T$, then estimating $\boldsymbol{\Sigma}_F$ is enough for getting optimal representation, and we will show it has lower sample complexity and error compared to estimating canonical covariance $\tilde{\boldsymbol{\Sigma}}_T$. That is a naive case, if it does not work, this intermediate result will help in our latter proof.

[1] For the root of 3-dim problem, the worst case we can grid the space and search with time complexity $\mathcal{O}(\varepsilon^{-3})$.

We will use the following Bernstein type concentration lemma, generalized from (Tripuraneni et al., 2020, Lemma 29):

Lemma 27. *Let $\boldsymbol{Z} \in \mathbb{R}^{n_1 \times n_2}$. Choose T_0, σ^2 such that*

1. $\boldsymbol{P}(\|\boldsymbol{Z}\| \geq C_0 T_0 + t) \leq \exp(-c\sqrt{t/T_0})$.
2. $\|\boldsymbol{E}(\boldsymbol{Z}\boldsymbol{Z}^\top)\|, \|\boldsymbol{E}(\boldsymbol{Z}^\top \boldsymbol{Z})\| \leq \sigma^2$.

Then with probability at least $1 - (nT_0)^{-c}$, $c > 10$,

$$\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Z}_i - \boldsymbol{E}(\boldsymbol{Z}_i)\| \lesssim \log(nT_0)\left(\frac{T_0 \log(nT_0)}{n} + \frac{\sigma}{\sqrt{n}}\right).$$

Proof. Define $K = \log^2(C_K n T_0)$ for $C_K > 0$, $\boldsymbol{Z}' = \boldsymbol{Z}\mathbf{1}(\|\boldsymbol{Z}\| \leq KT_0)$, then

$$\begin{aligned}\|\boldsymbol{E}(\boldsymbol{Z} - \boldsymbol{Z}')\| \leq \int_{KT_0}^{\infty} \exp(-c\sqrt{t/T_0})dt &\lesssim (1 + \sqrt{K})\exp(-c\sqrt{K})T_0 \\ &\lesssim (1 + \log(C_K n T_0))(nT_0)^{-C}.\end{aligned}$$

We can choose C_K large enough so that $C > 10$. We will use (Tripuraneni et al., 2020, Lemma 29). Set $R = \log^2(C_K n T_0)T_0 + C_0 T_0$, $\Delta = (1 + \log(C_K n T_0))(nT_0)^{-C}$, $t = C_t \log(nT_0)(\frac{T_0 \log(nT_0)}{n} + \frac{\sigma}{\sqrt{n}})$ for some $C_t > 0$, plugging in the last inequality of (Tripuraneni et al., 2020, Lemma 29), the LHS is smaller than $(nT_0)^{-c}$ for some c. We can also check $\boldsymbol{P}(\|\boldsymbol{Z}\| \geq R) \leq (nT_0)^{-c}$ for some c, thus we prove the lemma. □

Feature Covariance. We can directly estimate the covariance of features by

$$\hat{\boldsymbol{\Sigma}}_F = \frac{1}{N}\sum_{j=1}^{n_1}\sum_{i=1}^{T} \boldsymbol{x}_{i,j}\boldsymbol{x}_{i,j}^\top, \tag{D.20}$$

The mean of this estimator is $\boldsymbol{\Sigma}_F$ and we can estimate the top r eigenvector of $\boldsymbol{\Sigma}_F$ with $\tilde{\mathcal{O}}(r)$ samples.

As we have defined in Phase 1, features $\boldsymbol{x}_{i,j}$ are generated from $\mathcal{N}(0, \boldsymbol{\Sigma}_F)$. We aim to estimate the covariance $\boldsymbol{\Sigma}_F$. Although there are different kinds of algorithms, such as maximum likelihood estimator Anderson et al. (1970), to be consistent with the algorithms in the latter sections, we study the sample covariance matrix defined by (D.20).

Lemma 28. *Suppose* $\boldsymbol{x}_i$*,* $i = 1, ..., N$ *are generated independently from* $\mathcal{N}(0, \boldsymbol{\Sigma}_F)$*. We estimate* (D.20)*, then when* $N \gtrsim r_F$*, with probability* $1 - \mathcal{O}((N\mathbf{tr}(\boldsymbol{\Sigma}_F))^{-C})$*,*

$$\|\hat{\boldsymbol{\Sigma}}_F - \boldsymbol{\Sigma}_F\| \lesssim \sqrt{\frac{\|\boldsymbol{\Sigma}_F\|\mathbf{tr}(\boldsymbol{\Sigma}_F)}{N}}.$$

Denote the span of top s_F *eigenvectors of* $\boldsymbol{\Sigma}_F$ *as* $\boldsymbol{W}$ *and the span of top* s_F *eigenvectors of* $\hat{\boldsymbol{\Sigma}}_F$ *as* $\hat{\boldsymbol{W}}$*. Let* $\delta_\lambda = \lambda_{s_F}(\boldsymbol{\Sigma}_F) - \lambda_{s_F+1}(\boldsymbol{\Sigma}_F)$*. Then if* $N \gtrsim \frac{\|\boldsymbol{\Sigma}_F\|\mathbf{tr}(\boldsymbol{\Sigma}_F)}{\delta_\lambda^2}$*, we have*

$$\sin(\angle \boldsymbol{W}, \hat{\boldsymbol{W}}) \lesssim \sqrt{\frac{\|\boldsymbol{\Sigma}_F\|\mathbf{tr}(\boldsymbol{\Sigma}_F)}{N\delta_\lambda^2}}$$

Example 5. *When* $\boldsymbol{\Sigma}_F = diag(\boldsymbol{I}_{s_F}, 0)$*, we have* $\sin(\angle \boldsymbol{W}, \hat{\boldsymbol{W}}) \lesssim \sqrt{\frac{s_F}{N}}$*.*

Lemma 28 gives the quality of the estimation of the covariance of features $\boldsymbol{x}$. When the condition number of the matrix $\boldsymbol{\Sigma}_F$ is close to 1, we need $N \gtrsim d$ to get an estimation with error $\mathcal{O}(1)$. However, when the matrix $\boldsymbol{\Sigma}_F$ is close to rank r_F, the amount of samples to achieve the same error is smaller, and we can use $N \gtrsim r_F$ samples to get $\mathcal{O}(1)$ estimation error.

We will use Bernstein type concentration results to bound its error, and a similar technique will be used for $\hat{\boldsymbol{M}}$ in the next sections.

Proof. First we observe that, the features $\boldsymbol{x}_{i,j}$ among different tasks are generated i.i.d. from $\mathcal{N}(0, \boldsymbol{\Sigma}_F)$. So we can rewrite (D.20) as

$$\hat{\boldsymbol{\Sigma}}_F = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top \tag{D.21}$$

where $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_F)$. The error of $\hat{\boldsymbol{\Sigma}}_F$ depends on N regardless of T and n_1 respectively.

First, we know by concentration inequality

$$\boldsymbol{P}(\|\boldsymbol{x}\boldsymbol{x}^\top\| - \mathbf{tr}(\boldsymbol{\Sigma}_F) \geq t) = \boldsymbol{P}(\|\boldsymbol{x}\|^2 - \mathbf{tr}(\boldsymbol{\Sigma}_F) \geq t) \leq \exp(-c\min\{\frac{t^2}{\mathbf{tr}(\boldsymbol{\Sigma}_F^2)}, \frac{t}{\|\boldsymbol{\Sigma}_F\|}\}). \tag{D.22}$$

We will use the fact $\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_F^2)} \leq \mathbf{tr}(\boldsymbol{\Sigma}_F)$. Define $K = C_0 \log(N\mathbf{tr}(\boldsymbol{\Sigma}_F))\mathbf{tr}(\boldsymbol{\Sigma}_F)$, $\boldsymbol{Z} = \boldsymbol{x}\boldsymbol{x}^\top$, $\boldsymbol{Z}' = \boldsymbol{Z} \cdot \mathbf{1}\{\|\boldsymbol{Z}\| \leq K\}$ where $\mathbf{1}$ means indicator function ($\mathbf{1}(\text{True}) = 1, \mathbf{1}(\text{False}) = 0$), for some positive number C_0. Then

$$\begin{aligned}
\|\boldsymbol{E}(\boldsymbol{Z} - \boldsymbol{Z}')\| &\leq \int_{t=K}^{\infty} (\exp(-c\frac{t^2}{\mathbf{tr}^2(\boldsymbol{\Sigma}_F)}) + \exp(-c\frac{t}{\|\boldsymbol{\Sigma}_F\|}))dt \\
&\leq \int_{t=K}^{\infty} (\exp(-c\frac{t}{\mathbf{tr}(\boldsymbol{\Sigma}_F)}) + \exp(-c\frac{t}{\|\boldsymbol{\Sigma}_F\|}))dt \\
&\leq 2\frac{\mathbf{tr}(\boldsymbol{\Sigma}_F)}{c}\exp(-c\frac{K}{\mathbf{tr}(\boldsymbol{\Sigma}_F)}) \\
&\leq \frac{\sqrt{K\mathbf{tr}^2(\boldsymbol{\Sigma}_F)}}{c}\exp(-\frac{cK}{\mathbf{tr}(\boldsymbol{\Sigma}_F)}) \\
&\lesssim (N\mathbf{tr}(\boldsymbol{\Sigma}_F))^{-C}
\end{aligned}$$

where $C \geq C_0 - 3/2$. Then we compute $(\boldsymbol{x}\boldsymbol{x}^\top)^2 = \|\boldsymbol{x}\|^2\boldsymbol{x}\boldsymbol{x}^\top$. Let $\boldsymbol{\Sigma}_F$ be diagonal (the proof is invariant from the basis. In other words, if $\boldsymbol{\Sigma}_F$ is not diagonal, then we can make the eigenvectors of $\boldsymbol{\Sigma}_F$ as basis and the proof applies). Then

$$\boldsymbol{E}(\|\boldsymbol{x}\|^2\boldsymbol{x}\boldsymbol{x}^\top)_{ij} = \begin{cases} \boldsymbol{\Sigma}_{Fii}(\mathbf{tr}(\boldsymbol{\Sigma}_F) + 2\boldsymbol{\Sigma}_{Fii}), & i = j, \\ 0, & i \neq j. \end{cases} \tag{D.23}$$

So $\|\boldsymbol{E}(\|\boldsymbol{x}\|^2\boldsymbol{x}\boldsymbol{x}^\top)\| \leq \|\boldsymbol{\Sigma}_F\|(\mathbf{tr}(\boldsymbol{\Sigma}_F) + 2\|\boldsymbol{\Sigma}_F\|) \approx \|\boldsymbol{\Sigma}_F\|\mathbf{tr}(\boldsymbol{\Sigma}_F)$. $\approx$ means $\gtrsim$ and $\lesssim$.

Using Lemma 27, with (D.22) and the inequality above, we get that with probability

$1 - \mathcal{O}((N\mathbf{tr}(\mathbf{\Sigma}_F))^{-C})$,

$$\|\hat{\mathbf{\Sigma}}_F - \mathbf{\Sigma}_F\| \lesssim \log(N\mathbf{tr}(\mathbf{\Sigma}_F)) \left(\frac{\log(N\mathbf{tr}(\mathbf{\Sigma}_F))\mathbf{tr}(\mathbf{\Sigma}_F)}{N} + \sqrt{\frac{\|\mathbf{\Sigma}_F\|\mathbf{tr}(\mathbf{\Sigma}_F)}{N}} \right). \tag{D.24}$$

If the number above is smaller than $\lambda_r - \lambda_{r+1}$, we have that

$$N \gtrsim \frac{\|\mathbf{\Sigma}_F\|\mathbf{tr}(\mathbf{\Sigma}_F)}{(\lambda_r - \lambda_{r+1})^2} \tag{D.25}$$

which is $\mathcal{O}(r)$ if condition number is 1.

The bound of the angle of top R eigenvector subspace is a direct application of the following lemma.

Lemma 29. *Davis & Kahan (1970) Let $\boldsymbol{A}$ be a square matrix. Let $\hat{\boldsymbol{W}}$, $\boldsymbol{W}$ denote the span of top r singular vectors of $\hat{A}$ and $\boldsymbol{A}$. Suppose $\|\hat{\boldsymbol{A}} - \boldsymbol{A}\| \leq \Delta$, and $\sigma_r(\boldsymbol{A}) - \sigma_{r+1}(\boldsymbol{A}) \geq \Delta$, then*

$$\sin(\angle \boldsymbol{W}, \hat{\boldsymbol{W}}) \leq \frac{\Delta}{\sigma_r(\boldsymbol{A}) - \sigma_{r+1}(\boldsymbol{A}) - \Delta}.$$

So that the error of principle subspace recovery of feature covariance is upper bounded by $\frac{\|\hat{\mathbf{\Sigma}}_F - \mathbf{\Sigma}_F\|}{\sigma_r(\mathbf{\Sigma}_F) - \sigma_{r+1}(\mathbf{\Sigma}_F) - \|\hat{\mathbf{\Sigma}}_F - \mathbf{\Sigma}_F\|}$, where $\|\hat{\mathbf{\Sigma}}_F - \mathbf{\Sigma}_F\|$ is calculated in (D.24). □

D.3.2 Method of moment

This section contains three parts. We first bound the norm of task vectors. Then we analyze the second result of Thm. 10, where n_1 is lower bounded by effective rank. Last we prove the first result of Thm. 10 which is a generalization of Tripuraneni et al. (2020).

Property of task vectors

We first study the property of the tasks $\beta_1, ..., \beta_T$. We know that, for any $\beta \sim \mathcal{N}(0, \mathbf{\Sigma}_T)$,

$$\boldsymbol{P}(\|\beta\|^2 - \mathbf{tr}(\mathbf{\Sigma}_T) \geq t) \leq \exp(-c\min\{\frac{t^2}{\mathbf{tr}(\mathbf{\Sigma}_T^2)}, \frac{t}{\|\mathbf{\Sigma}_T\|}\}).$$

So that with probability at least $1 - \delta$, we have

$$\begin{aligned} \|\beta_i\|^2 &\lesssim \mathbf{tr}(\mathbf{\Sigma}_T) + \sqrt{(\log(1/\delta) + \log(T))\mathbf{tr}(\mathbf{\Sigma}_T^2)} + (\log(1/\delta) + \log(T))\|\mathbf{\Sigma}_T\| \\ &\lesssim \mathbf{tr}(\mathbf{\Sigma}_T) + \log(T/\delta)\sqrt{\mathbf{tr}(\mathbf{\Sigma}_T^2)} \lesssim \mathbf{tr}(\mathbf{\Sigma}_T)\log(T/\delta), \ \forall i = 1, ..., T. \end{aligned} \tag{D.26}$$

With similar technique we know that with probability at least $1 - \delta$,

$$\|\mathbf{\Sigma}_F\beta_i\|^2 \lesssim \mathbf{tr}(\mathbf{\Sigma}_F\mathbf{\Sigma}_T\mathbf{\Sigma}_F) + \log(T/\delta)\sqrt{\mathbf{tr}((\mathbf{\Sigma}_F\mathbf{\Sigma}_T\mathbf{\Sigma}_F)^2)}, \ \forall i = 1, ..., T. \tag{D.27}$$

$$\|\mathbf{\Sigma}_F^{1/2}\beta_i\|^2 \lesssim \mathbf{tr}(\mathbf{\Sigma}_F^{1/2}\mathbf{\Sigma}_T\mathbf{\Sigma}_F^{1/2}) + \log(T/\delta)\sqrt{\mathbf{tr}((\mathbf{\Sigma}_F^{1/2}\mathbf{\Sigma}_T\mathbf{\Sigma}_F^{1/2})^2)}, \ \forall i = 1, ..., T. \tag{D.28}$$

We will use $\delta = T^{-c}$ for some constant c so that $\log(T/\delta) = (c+1)\log(T) \approx \log(T)$. Later, we will use the norm bounds of above quantities which happen with probability at least $1 - T^{-c}$.

Estimating with fewer samples when each task contains enough samples

In this part we will prove Lemma 31, which is the second case of Theorem 10. First we will give a description of standard normal features, then prove the general version.

Lemma 30. *(Standard normal feature, noiseless) Let data be generated as in Phase 1, let $\mathcal{S} = \max\{\|\mathbf{\Sigma}_F\|, \|\mathbf{\Sigma}_T\|\}$ in this theorem and the following section*[2]*, $r = \mathbf{tr}(\mathbf{\Sigma}_T\mathbf{\Sigma}_F)$, $r_F = \mathbf{tr}(\mathbf{\Sigma}_F)$, $r_T = \mathbf{tr}(\mathbf{\Sigma}_T)$. Suppose $\sigma = 0$, $\mathbf{\Sigma}_F = I$, and suppose the rank of $\mathbf{\Sigma}_T$ is s_T. Define $\hat{\beta}_i = n_1^{-1}\sum_{j=1}^{n_1} y_{i,j}\boldsymbol{x}_{i,j}$, $\boldsymbol{B} = [\beta_1, ..., \beta_T]$, and $\hat{\boldsymbol{B}} = [\hat{\beta}_1, ..., \hat{\beta}_T]$. Let $n_1 > c_1 r_T \lambda_{s_T}^{-1}(\mathbf{\Sigma}_T)$,*

[2] in the main body we assumed $\mathcal{S} = 1$ for simplicity.

with probability $1 - \mathcal{O}(T^{-C})$*, where* C *is constant,*

$$\sigma_{\max}(\hat{\boldsymbol{B}} - \boldsymbol{B}) \lesssim \sqrt{\frac{Tr_T}{n_1}}.$$

Denote the span of top s_T *singular column vectors of* $\hat{\boldsymbol{B}}$ *and* $\boldsymbol{\Sigma}_T$ *as* $\hat{\boldsymbol{W}}, \boldsymbol{W}$*, then*

$$\sin(\angle\hat{\boldsymbol{W}}, \boldsymbol{W}) \lesssim \sqrt{\frac{r_T}{n_1\lambda_{s_T}(\boldsymbol{\Sigma}_T)}}.$$

For example, if $\boldsymbol{\Sigma}_T = \mathrm{diag}(I_{s_T}, 0)$, then $\sin(\angle\hat{\boldsymbol{W}}, \boldsymbol{W}) \lesssim \sqrt{s_T/n_1}$.

Proof. We first estimate β_i with

$$\hat{\beta}_i = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{i,j}\boldsymbol{x}_{i,j}.$$

Then we fix β_i and compute the covariance of $y_{i,j}\boldsymbol{x}_{i,j}$ (its mean is β_i).

$$\mathrm{Cov}(y_{i,j}\boldsymbol{x}_{i,j} - \beta_i) = \boldsymbol{E}(\boldsymbol{x}_{i,j}\boldsymbol{x}_{i,j}^\top\beta_i\beta_i^\top\boldsymbol{x}_{i,j}\boldsymbol{x}_{i,j}^\top) - \beta_i\beta_i^\top \precsim \|\beta_i\|^2 I.$$

The first term is similar to (D.23), where the bound can is in (Tripuraneni et al., 2020, Lemma 5). The vector $\hat{\beta}_i$ is the average of $y_{i,j}\boldsymbol{x}_{i,j}$ over all j. With concentration we know that

$$\mathrm{Cov}(\hat{\beta}_i - \beta_i) \precsim \frac{\|\beta_i\|^2}{n_1} I. \tag{D.29}$$

Let $\boldsymbol{B} = [\beta_1, ..., \beta_T]$, and $\hat{\boldsymbol{B}} = [\hat{\beta}_1, ..., \hat{\beta}_T]$. Then we know the covariance of each column of $\hat{\boldsymbol{B}} - \boldsymbol{B}$ is bounded by (D.29). Thus with a constant c and probability $1 - \exp(-cT^2)$,

$$\sigma^2_{\max}(\hat{\boldsymbol{B}} - \boldsymbol{B}) \lesssim \frac{T\|\beta_i\|^2}{n_1}. \tag{D.30}$$

We have proved in (D.26) that $\|\beta_i\|^2 \leq \log(T)\mathbf{tr}(\boldsymbol{\Sigma}_T)$ with probability $1 - T^{-c}$. The columns of $\boldsymbol{B}$ is generated from $\mathcal{N}(0, \boldsymbol{\Sigma}_T)$, so that

$$\sigma_{\max}(\hat{\boldsymbol{B}} - \boldsymbol{B}) \lesssim \sqrt{\frac{T \log(T) \mathbf{tr}(\boldsymbol{\Sigma}_T)}{n_1}}.$$

Now we study $\boldsymbol{B}$. We know that $\boldsymbol{E}(\boldsymbol{B}\boldsymbol{B}^\top) = \boldsymbol{E}(\sum_{i=1}^T \beta_i \beta_i^\top) = T\boldsymbol{\Sigma}_T$. $\boldsymbol{B}$ is a matrix with independent columns. Thus let $n_1 > c_1 \mathbf{tr}(\boldsymbol{\Sigma}_T)\lambda_{s_T}^{-1}(\boldsymbol{\Sigma}_T)$, $T > \max\{c_2 d, \frac{\|\boldsymbol{\Sigma}_T\|\mathbf{tr}(\boldsymbol{\Sigma}_T)}{\lambda_{s_T}^2(\boldsymbol{\Sigma}_T)}\}$, then with Lemma 28, for Gaussian matrix with independent columns Vershynin (2010), with probability at least $1 - \mathcal{O}(T^{-c_3} + (T\mathbf{tr}(\boldsymbol{\Sigma}_T))^{-c_4} + \exp(-c_5 T^2)) = 1 - \mathcal{O}(T^{-C})$, where c_i are constants,

$$\sigma_{s_T}(\boldsymbol{B}) \geq \sqrt{T\lambda_{s_T}(\boldsymbol{\Sigma}_T) - \mathcal{O}(\sqrt{T}\|\boldsymbol{\Sigma}_T\|\mathbf{tr}(\boldsymbol{\Sigma}_T))}.$$

Denote the span of top s_T singular vectors of $\hat{\boldsymbol{B}}$ and $\boldsymbol{\Sigma}_T$ as $\hat{\boldsymbol{W}}, \boldsymbol{W}$, with Lemma 29,

$$\sin(\angle \hat{\boldsymbol{W}}, \boldsymbol{W}) \leq \sqrt{\frac{\log(T)\mathbf{tr}(\boldsymbol{\Sigma}_T)}{n_1 \lambda_{s_T}(\boldsymbol{\Sigma}_T)}}.$$

□

Next, we will propose a theorem with general feature covariance and noisy data, which is a generalization of Lemma 30.

Lemma 31. *Let data be generated as in Phase 1. Suppose* $\hat{\boldsymbol{b}}_i = n_1^{-1}\sum_{j=1}^{n_1} y_{i,j}\boldsymbol{x}_{i,j}$*,* $\boldsymbol{B} = \boldsymbol{\Sigma}_F[\beta_1, ..., \beta_T]$*, and* $\hat{\boldsymbol{B}} = [\hat{\boldsymbol{b}}_1, ..., \hat{\boldsymbol{b}}_T]$*. Let* $\delta_\lambda = \lambda_{s_T}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) - \lambda_{s_T+1}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F))$*, suppose* $\boldsymbol{\Sigma}_F$ *is approximately rank* s_F*,*

$$n_1 \gtrsim (\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|,$$
$$T \gtrsim \max\{s_F, \frac{d\lambda_{s_F+1}(\boldsymbol{\Sigma}_F)}{\|\boldsymbol{\Sigma}_F\|}\},$$

then with probability $1 - \mathcal{O}(T^{-C})$*, where* C *is constant,*

$$\sigma_{\max}(\hat{\boldsymbol{B}} - \boldsymbol{B}) \lesssim \sqrt{\frac{T(\mathbf{tr}(\boldsymbol{\Sigma}_T \boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1}}.$$

Denote the span of top s_T *singular vectors of* $\hat{\boldsymbol{B}}$ *and* $\boldsymbol{\Sigma}_F \boldsymbol{\Sigma}_T \boldsymbol{\Sigma}_F$ *as* $\hat{\boldsymbol{W}}, \boldsymbol{W}$*, if further we assume* $T \gtrsim \frac{\|\boldsymbol{\Sigma}_F \boldsymbol{\Sigma}_T \boldsymbol{\Sigma}_F\| \mathbf{tr}(\boldsymbol{\Sigma}_F \boldsymbol{\Sigma}_T \boldsymbol{\Sigma}_F)}{\delta_\lambda^2}$*, then*

$$\sin(\angle \hat{\boldsymbol{W}}, \boldsymbol{W}) \lesssim \sqrt{\frac{(\mathbf{tr}(\boldsymbol{\Sigma}_T \boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1 \delta_\lambda^2}}.$$

Example 6. *Suppose* $\boldsymbol{\Sigma}_F = diag(I_{s_F}, \iota I_{d-s_F})$*, and* $\boldsymbol{\Sigma}_T = diag(I_{s_T}, 0)$*,* $\sigma = 0$*. Suppose* $\iota d < s_F$*. Then with* $T \gtrsim s_F$*,* $n_1 \gtrsim s_T$ *so that* $N \gtrsim s_F s_T$*,*

$$\sin(\angle \hat{\boldsymbol{W}}, \boldsymbol{W}) \lesssim \sqrt{s_T / n}.$$

Proof. We let $\boldsymbol{x}_{i,j} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_F)$. For the ith task, let

$$\hat{\boldsymbol{b}}_i = \frac{1}{n_1} \sum_{j=1}^{n_1} y_{i,j} \boldsymbol{x}_{i,j}.$$

We fix β_i and compute

$$\boldsymbol{E}(y_{i,j} \boldsymbol{x}_{i,j}) \precsim \boldsymbol{E}(\boldsymbol{x}_{i,j} \boldsymbol{x}_{i,j}^\top \beta_i) = \boldsymbol{\Sigma}_F \beta_i, \tag{D.31}$$

and

$$\mathrm{Cov}(y_{i,j} \boldsymbol{x}_{i,j} - \boldsymbol{\Sigma}_F \beta_i) \precsim (\beta_i^\top \boldsymbol{\Sigma}_F \beta_i) \boldsymbol{\Sigma}_F + \sigma^2 \boldsymbol{\Sigma}_F. \tag{D.32}$$

To get the bound above, we can adopt the technique in (Tripuraneni et al., 2020, Lemma 5) such that, write $\boldsymbol{x}_{i,j} = \boldsymbol{\Sigma}_F^{1/2} \boldsymbol{z}$, and reduce to $\boldsymbol{E}((\boldsymbol{z}^\top \boldsymbol{\Sigma}_F^{1/2} \beta_i)^2 \boldsymbol{\Sigma}_F^{1/2} \boldsymbol{z} \boldsymbol{z}^\top \boldsymbol{\Sigma}_F^{1/2})$. The proof of

(Tripuraneni et al., 2020, Lemma 5) gives the explicit bound of $\|\boldsymbol{E}((\boldsymbol{z}^\top \boldsymbol{\alpha})^2 \boldsymbol{z}\boldsymbol{z}^\top)\|$ for any $\boldsymbol{\alpha}$ that equals above. The vector $\hat{\boldsymbol{b}}_i$ is the average of $y_{i,j}\boldsymbol{x}_{i,j}$ over all $j = 1, ..., n_1$. With concentration we know that

$$\mathrm{Cov}(\hat{\boldsymbol{b}}_i - \boldsymbol{\Sigma}_F \beta_i) \precsim \frac{\beta_i^\top \boldsymbol{\Sigma}_F \beta_i + \sigma^2}{n_1} \boldsymbol{\Sigma}_F. \tag{D.33}$$

Suppose $\boldsymbol{B} = \boldsymbol{\Sigma}_F[\beta_1, ..., \beta_T]$, and $\hat{\boldsymbol{B}} = [\boldsymbol{b}_1, ..., \boldsymbol{b}_T]$. $\hat{\boldsymbol{B}} - \boldsymbol{B}$ is a matrix with independent columns. Suppose $\boldsymbol{X}$ is approximately rank s_F, Let $\boldsymbol{V}_{s_F} \in \mathbb{R}^{d\times d}$ be the projection onto the top-R sigular vector space of $\boldsymbol{\Sigma}_F$ and $\boldsymbol{V}_{s_F^\perp} \in \mathbb{R}^{d\times d}$ be the projection onto the $s_F + 1$ to dth sigular vector space of $\boldsymbol{\Sigma}_F$. With T columns and $T \geq s_F$, we know that

$$\sigma_{\max}(\boldsymbol{V}_{s_F}(\hat{\boldsymbol{B}} - \boldsymbol{B})) \lesssim \frac{T(\max_i \beta_i^\top \boldsymbol{\Sigma}_F \beta_i + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1}$$
$$\sigma_{\max}(\boldsymbol{V}_{s_F^\perp}(\hat{\boldsymbol{B}} - \boldsymbol{B})) \lesssim \frac{\max\{T, d\}(\max_i \beta_i^\top \boldsymbol{\Sigma}_F \beta_i + \sigma^2)\lambda_{s_T+1}(\boldsymbol{\Sigma}_F)}{n_1}$$

With similar argument as before, with probability $1 - \exp(-cT^2)$ for constant c,

$$\sigma^2_{\max}(\hat{\boldsymbol{B}} - \boldsymbol{B}) \lesssim \frac{\max\{T\|\boldsymbol{\Sigma}_F\|, d\lambda_{s_F+1}(\boldsymbol{\Sigma}_F)\}(\max_i \beta_i^\top \boldsymbol{\Sigma}_F \beta_i + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1}. \tag{D.34}$$

We know in (D.28) that $\|\boldsymbol{\Sigma}_F^{1/2}\beta_i\|^2 \leq \mathcal{O}(\log(T)\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F))$ with probability $1 - T^{-c}$ for constant c. So that

$$\sigma_{\max}(\hat{\boldsymbol{B}} - \boldsymbol{B}) \lesssim \sqrt{\frac{\max\{T\|\boldsymbol{\Sigma}_F\|, d\lambda_{s_F+1}(\boldsymbol{\Sigma}_F)\}(\log(T)\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1}}. \tag{D.35}$$

Now we study $\boldsymbol{B}$. $\boldsymbol{E}(\boldsymbol{B}\boldsymbol{B}^\top) = \boldsymbol{E}(\boldsymbol{\Sigma}_F(\sum_{i=1}^T \beta_i\beta_i^\top)\boldsymbol{\Sigma}_F) = T\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F$.

Thus let

$$n_1 > C_1(\log(T)\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|.$$

Now apply the concentration of Gaussian matrix with independent columns Vershynin (2010).

With probability $1 - \mathcal{O}(T^{-C_1} + (T\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F))^{-C_2} + \exp(-C_3T^2))$, where C_i are constants (the probability can be simplified as $1 - \mathcal{O}(T^{-C})$),

$$\sigma_{s_T}(\boldsymbol{B}) \geq \sqrt{T(\lambda_{s_T}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) - \lambda_{s_T+1}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F)) - \mathcal{O}(\sqrt{T\|\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F\|\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F)})}.$$

Denote the span of top s_T singular vectors of $\hat{\boldsymbol{B}}$ and $\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F$ as $\hat{\boldsymbol{W}}, \boldsymbol{W}$, let

$$T \gtrsim \max\{s_F, \frac{d\lambda_{s_F+1}(\boldsymbol{\Sigma}_F)}{\|\boldsymbol{\Sigma}_F\|}, \frac{\|\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F\|\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F)}{(\lambda_{s_T}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) - \lambda_{s_T+1}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F))^2}\} \tag{D.36}$$

we plug in (D.35) and Lemma 29,

$$\begin{aligned}\sin(\angle\hat{\boldsymbol{W}}, \boldsymbol{W}) &\lesssim \sqrt{(\frac{d\lambda_{s_F+1}(\boldsymbol{\Sigma}_F)}{T\|\boldsymbol{\Sigma}_F\|} + 1) \cdot \frac{(\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1(\lambda_{s_T}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) - \lambda_{s_T+1}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F))}} \\ &\approx \sqrt{\frac{(\mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) + \sigma^2)\|\boldsymbol{\Sigma}_F\|}{n_1(\lambda_{s_T}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F) - \lambda_{s_T+1}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F))}}.\end{aligned}$$

□

Method of moments with arbitrary n_1

In this subsection we will analyze $\hat{\boldsymbol{B}}$ with any n_1, and propose the error of MoM estimator.

Suppose there are at least two samples per task, we can separate the samples into two halves, and compute the following estimator.

Lemma 32. *Let data be generated as in Phase 1, and let n_1 be a even number. Define $\hat{\boldsymbol{b}}_{i,1} = 2n_1^{-1}\sum_{j=1}^{n_1/2} y_{i,j}\boldsymbol{x}_{i,j}$, $\hat{\boldsymbol{b}}_{i,2} = 2n_1^{-1}\sum_{j=n_1/2+1}^{n_1} y_{i,j}\boldsymbol{x}_{i,j}$. Define*

$$\begin{aligned}\hat{\boldsymbol{M}} &= n_1^{-1}\sum_{i=1}^{T}(\boldsymbol{b}_{i,1}\boldsymbol{b}_{i,2}^\top + \boldsymbol{b}_{i,2}\boldsymbol{b}_{i,1}^\top), \\ \boldsymbol{M} &= \boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F.\end{aligned}$$

Then there is a constant $c > 10$, *with probability* $1 - N^{-c}$,

$$\|\hat{\boldsymbol{M}} - \boldsymbol{M}\| \lesssim (r + \sigma^2)\sqrt{\frac{r_F}{N}} + \sqrt{\frac{r_T}{T}}.$$

Proof. For simplicity of notation, we will define a random vector $\boldsymbol{x}$ with zero mean and covariance $\boldsymbol{\Sigma}_F$, a random vector β with zero mean and covariance $\boldsymbol{\Sigma}_T$, a random variable ε with zero mean and covariance σ, and they are subGaussian[3]. Let $y = \boldsymbol{x}^\top \beta + \varepsilon$. We first estimate the mean of $\hat{\boldsymbol{M}}$.

Note that if we fix β, $\hat{\boldsymbol{b}}_{i,1}, \hat{\boldsymbol{b}}_{i,2}$ are i.i.d., so

$$\begin{aligned}
\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1}) &= \boldsymbol{E}_{\boldsymbol{x},\varepsilon}(y\boldsymbol{x}) = \boldsymbol{E}_{\boldsymbol{x},\varepsilon}((\boldsymbol{x}^\top \beta + \varepsilon)\boldsymbol{x}) = \boldsymbol{\Sigma}_F \beta, \\
\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{M}}) &= \frac{1}{2}(\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1})\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,2})^\top + \boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,2})\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1})^\top) \\
&= \boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1})\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1})^\top = \frac{1}{T}\boldsymbol{\Sigma}_F(\sum_{i=1}^{T} \beta_i \beta_i^\top)\boldsymbol{\Sigma}_F.
\end{aligned}$$

We take expectation over β_i and get $\boldsymbol{M}$. We define the right hand side as $\bar{\boldsymbol{M}}$ for the proof below.

Next, we will bound $\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|$.

(Tripuraneni et al., 2020, Lemma 3) proposes that, with probability $1 - \delta$,

$$\begin{aligned}
\|\boldsymbol{x}_{i,j}\|^2 &\lesssim \log(1/\delta)\mathbf{tr}(\boldsymbol{\Sigma}_F), \\
(\boldsymbol{x}_{i,j}^\top \beta_i)^2 &\lesssim \log(1/\delta)\mathbf{tr}(\boldsymbol{\Sigma}_F \boldsymbol{\Sigma}_T), \\
\varepsilon_{ij}^2 &\lesssim \log(1/\delta)\sigma^2.
\end{aligned}$$

If we enumerate $i = 1, ..., T$ and $j = 1, ..., n_1$, there are in total $Tn_1 = N$ terms. So we set

[3] We remove the subscripts when there is no confusion.

$\delta = N^{-c+1}$ for a constant $c > 1$, then with probability $1 - N^{-c}$, for all i, j we have

$$\|y_{i,j}\boldsymbol{x}_{i,j}\| = \|(\boldsymbol{x}_{i,j}\beta_i + \varepsilon_{ij})\boldsymbol{x}_{i,j}\| \lesssim \log^{3/2}(N)\sqrt{(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\mathbf{tr}(\boldsymbol{\Sigma}_F)}.$$

Define $\boldsymbol{\delta}_{i,l} = \hat{\boldsymbol{b}}_{i,l} - \boldsymbol{\Sigma}_F\beta_i$ for $l = 1, 2$ (we will use $l = 1$ below, the result for $l = 2$ is the same). Note that $\boldsymbol{\delta}_i$ is zero mean. With (Kong et al., 2020b, Prop. 5.1) we have with probability $1 - N^{-c}$,

$$\|\boldsymbol{\delta}_{i,1}\| \lesssim n_1^{-1/2}\log^{5/2}(N)\sqrt{(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\mathbf{tr}(\boldsymbol{\Sigma}_F)} \tag{D.37}$$

Define

$$\begin{aligned}
\boldsymbol{Z}_i &= \hat{\boldsymbol{b}}_{i,1}\hat{\boldsymbol{b}}_{i,2}^\top - \boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1}\hat{\boldsymbol{b}}_{i,2}^\top) \\
&= (\boldsymbol{\Sigma}_F\beta_i + \boldsymbol{\delta}_{i,1})(\boldsymbol{\Sigma}_F\beta_i + \boldsymbol{\delta}_{i,2})^\top - \boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\hat{\boldsymbol{b}}_{i,1}\hat{\boldsymbol{b}}_{i,2}^\top) \\
&= \boldsymbol{\delta}_{i,1}(\boldsymbol{\Sigma}_F\beta_i)^\top + \boldsymbol{\Sigma}_F\beta_i\boldsymbol{\delta}_{i,2}^\top + \boldsymbol{\delta}_{i,1}\boldsymbol{\delta}_{i,2}^\top - \boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\boldsymbol{\delta}_{i,1}\boldsymbol{\delta}_{i,2}^\top).
\end{aligned}$$

Then

$$\begin{aligned}
\|\boldsymbol{E}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\| \leq{}& \|\boldsymbol{E}(\boldsymbol{\Sigma}_F\beta_i\boldsymbol{\delta}_{i,2}^\top + \boldsymbol{\delta}_{i,1}(\boldsymbol{\Sigma}_F\beta_i)^\top)(\boldsymbol{\Sigma}_F\beta_i\boldsymbol{\delta}_{i,2}^\top + \boldsymbol{\delta}_{i,1}(\boldsymbol{\Sigma}_F\beta_i)^\top)^\top\| \\
&+ \|\boldsymbol{E}\boldsymbol{\delta}_{i,1}\boldsymbol{\delta}_{i,2}^\top\boldsymbol{\delta}_{i,2}\boldsymbol{\delta}_{i,1}^\top\|.
\end{aligned} \tag{D.38}$$

Then we can use (D.37) and (D.27) to bound the first term by

$$n_1^{-1}\log^6(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma)\mathbf{tr}(\boldsymbol{\Sigma}_F)\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)\|\boldsymbol{\Sigma}_F\|^2.$$

And

$$\begin{aligned}\boldsymbol{E}_{\boldsymbol{x},\varepsilon}\boldsymbol{\delta}_{i,1}\boldsymbol{\delta}_{i,2}^\top\boldsymbol{\delta}_{i,2}\boldsymbol{\delta}_{i,1}^\top &= (\boldsymbol{E}_{\boldsymbol{x}}\boldsymbol{\delta}_{i,2}^\top\boldsymbol{\delta}_{i,2})\|\boldsymbol{E}_{\boldsymbol{x}}\boldsymbol{\delta}_{i,1}\boldsymbol{\delta}_{i,1}^\top\| \\ &\lesssim n_1^{-2}(\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\boldsymbol{x}^\top\beta+\varepsilon)^2\boldsymbol{x}^\top\boldsymbol{x})\|\boldsymbol{E}_{\boldsymbol{x},\varepsilon}(\boldsymbol{x}^\top\beta+\varepsilon)^2\boldsymbol{x}\boldsymbol{x}^\top\| \\ &\lesssim n_1^{-2}(\mathbf{tr}^2(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^4)\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|.\end{aligned}$$

The second line is due to the fact that $\boldsymbol{\delta}_{i,l}$ is the difference of $(\boldsymbol{x}^\top\beta+\varepsilon)\boldsymbol{x}$ and its mean, and covariance is upper bounded by variance (not subtracting the mean). The n_1^{-2} factor comes from the average over n_1 terms. The reasoning of the last line is same as (D.32). Now we can go back to (D.38) and get

$$\|\boldsymbol{E}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\| \lesssim n_1^{-1}\log^6(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)^2\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|^2.$$

Next we need to bound the norm of $\boldsymbol{Z}_i$. We use (D.37) and (D.27), with probability $1-N^{-c}$,

$$\begin{aligned}\|\boldsymbol{Z}_i\| \le{}& n_1^{-1/2}\log^3(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_F)}\|\boldsymbol{\Sigma}_F\| \\ &+ n_1^{-1}\log^5(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\mathbf{tr}(\boldsymbol{\Sigma}_F).\end{aligned}$$

Define the upper bound for $\|\boldsymbol{E}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\|, \|\boldsymbol{Z}_i\|$ as Z_1, Z_2 (the right hand side of two above inequalities). Now we apply Bernstein type inequality (Lemma 27), with probability $1 - N^{-c}$,

$$\begin{aligned}
&\|\hat{\boldsymbol{M}} - \bar{\boldsymbol{M}}\| \\
&= \|T^{-1}\sum_{i=1}^{T} \boldsymbol{Z}_i - \boldsymbol{E}_{\boldsymbol{x}}\boldsymbol{Z}_i\| \\
&\lesssim \log(TZ_2)\Big(T^{-1/2}\log(N)Z_1^{1/2} + T^{-1}Z_2\log(TZ_2)\Big) \\
&\lesssim \log(TZ_2)\Big(\sqrt{\frac{\log^6(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T) + \mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)^2\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|^2}{n_1 T}} \\
&\quad + \frac{\log^3(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T) + \mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_F)}\|\boldsymbol{\Sigma}_F\|}{n_1^{1/2}T} \\
&\quad + \frac{\log^5(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\mathbf{tr}(\boldsymbol{\Sigma}_F)}{T}\Big) \\
&= \log(TZ_2)\cdot\Big(\log^3(N)\|\boldsymbol{\Sigma}_F\|(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T) + \mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\sqrt{\frac{\mathbf{tr}(\boldsymbol{\Sigma}_F)}{N}} \\
&\quad + \frac{\log^5(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T) + \mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_F)}\|\boldsymbol{\Sigma}_F\|}{N^{1/2}T^{1/2}}\Big).
\end{aligned}$$

The term

$$\|\boldsymbol{\Sigma}_F\|(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T) + \mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\sqrt{\frac{\mathbf{tr}(\boldsymbol{\Sigma}_F)}{N}}$$

is the dominant term as shown in the theorem. □

The following method of moment estimator is used in Tripuraneni et al. (2020), where $n_1 \geq 1$. In other words, if there is one sample per task, one can use the following estimator.

Lemma 33. *Let data be generated as in Phase 1. Define* $\hat{\boldsymbol{b}}_i = n_1^{-1}\sum_{j=1}^{n_1} y_{i,j}\boldsymbol{x}_{i,j}$, $\boldsymbol{B} =$

$\boldsymbol{\Sigma}_F[\beta_1, ..., \beta_T]$, and $\hat{\boldsymbol{B}} = [\hat{\boldsymbol{b}}_1, ..., \hat{\boldsymbol{b}}_T]$. Define

$$\begin{aligned}
\hat{\mathbf{G}} &= \hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top = T^{-1}\sum_{i=1}^{T} \hat{\boldsymbol{b}}_i \hat{\boldsymbol{b}}_i^\top, \\
\mathbf{G} &= \boldsymbol{E}(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top) = \boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F + n_1^{-1}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F + \mathbf{tr}(\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_F)\boldsymbol{\Sigma}_F + \sigma^2\boldsymbol{\Sigma}_F), \\
\bar{\boldsymbol{\Sigma}}_T &= \sum_{i=1}^{T} \beta_i\beta_i^\top, \\
\bar{\mathbf{G}} &= \boldsymbol{\Sigma}_F\bar{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_F + n_1^{-1}(\boldsymbol{\Sigma}_F\bar{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_F + \mathbf{tr}(\bar{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_F)\boldsymbol{\Sigma}_F + \sigma^2\boldsymbol{\Sigma}_F)
\end{aligned}$$

With probability $1 - N^c$,

$$\|\hat{\mathbf{G}} - \bar{\mathbf{G}}\| \lesssim \|\boldsymbol{\Sigma}_F\|(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T) + \mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T) + \sigma^2)\sqrt{\frac{\mathbf{tr}(\boldsymbol{\Sigma}_F)}{N}}.$$

Proof. First, we compute the expectation of $\hat{\mathbf{G}}$.

$$\begin{aligned}
\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\hat{\mathbf{G}} &= \boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}T^{-1}(\sum_{i=1}^{T} \hat{\boldsymbol{b}}_i \hat{\boldsymbol{b}}_i^\top), \\
\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\hat{\boldsymbol{b}}_i \hat{\boldsymbol{b}}_i^\top &= \boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\left(n_1^{-1}\sum_{j=1}^{n_1}(\beta_i^\top \boldsymbol{x}_{i,j} + \varepsilon_{ij})\boldsymbol{x}_{i,j}\right)\left(n_1^{-1}\sum_{j=1}^{n_1}(\beta_i^\top \boldsymbol{x}_{i,j} + \varepsilon_{ij})\boldsymbol{x}_{i,j}\right)^\top \\
&= n_1^{-1}\sigma^2\boldsymbol{\Sigma}_F + \boldsymbol{E}_{\boldsymbol{x}}(n_1^{-1}\sum_{j=1}^{n_1}\boldsymbol{x}_{i,j}\boldsymbol{x}_{i,j}^\top\beta_i)(n_1^{-1}\sum_{j=1}^{n_1}\boldsymbol{x}_{i,j}\boldsymbol{x}_{i,j}^\top\beta_i)^\top. \qquad \text{(D.39)}
\end{aligned}$$

Now we will study the second term. (D.31) states that $\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}(\hat{\boldsymbol{b}}_i) = \boldsymbol{\Sigma}_F\beta_i$. And $\hat{\boldsymbol{b}}_i$ is an average of n_1 terms, we use the expression of the covariance of sample means to get

$$\begin{aligned}
\mathbf{Cov}(\hat{\boldsymbol{b}}_i) &= n_1^{-1}\mathbf{Cov}(\boldsymbol{x}\boldsymbol{x}^\top\beta_i), \qquad \text{(D.40)} \\
\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\hat{\boldsymbol{b}}_i \hat{\boldsymbol{b}}_i^\top &= \boldsymbol{E}_{\boldsymbol{x}}(n_1^{-1}\sum_{j=1}^{n_1}\boldsymbol{x}_{i,j}\boldsymbol{x}_{i,j}^\top\beta_i)(n_1^{-1}\sum_{j=1}^{n_1}\boldsymbol{x}_{i,j}\boldsymbol{x}_{i,j}^\top\beta_i)^\top \\
&= \boldsymbol{\Sigma}_F\beta_i\beta_i^\top\boldsymbol{\Sigma}_F + n_1^{-1}\mathbf{Cov}(\boldsymbol{x}\boldsymbol{x}^\top\beta_i) \qquad \text{(D.41)}
\end{aligned}$$

Now we study $\mathbf{Cov}(\boldsymbol{x}\boldsymbol{x}^\top\beta_i)$.

$$\begin{aligned}\mathbf{Cov}(\boldsymbol{x}\boldsymbol{x}^\top\beta_i) &= \boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{x}\boldsymbol{x}^\top\beta_i - \boldsymbol{\Sigma}_F\beta_i)(\boldsymbol{x}\boldsymbol{x}^\top\beta_i - \boldsymbol{\Sigma}_F\beta_i)^\top \\ &= \boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{x}\boldsymbol{x}^\top\beta_i)(\boldsymbol{x}\boldsymbol{x}^\top\beta_i)^\top - \boldsymbol{\Sigma}_F\beta_i\beta_i^\top\boldsymbol{\Sigma}_F\end{aligned}$$

Let $\boldsymbol{x} = \sqrt{\boldsymbol{\Sigma}_F}\boldsymbol{z}$ so that $\boldsymbol{z} \sim \mathcal{N}(0, I)$. Let two indices $k, l \in [d]$. When $k \neq l$,

$$\begin{aligned}\boldsymbol{E}_{\boldsymbol{x}}[(\boldsymbol{x}\boldsymbol{x}^\top\beta_i)(\boldsymbol{x}\boldsymbol{x}^\top\beta_i)^\top]_{kl} &= \boldsymbol{E}_{\boldsymbol{z}}(\sum_{j=1}^d \beta_{i,j}\sigma_j\boldsymbol{z}_j)^2\sigma_k\boldsymbol{z}_k\sigma_l\boldsymbol{z}_l \\ &= 2\sigma_k^2\sigma_l^2\beta_{i,k}\beta_{i,l}\end{aligned}$$

And

$$\begin{aligned}\boldsymbol{E}_{\boldsymbol{x}}[(\boldsymbol{x}\boldsymbol{x}^\top\beta_i)(\boldsymbol{x}\boldsymbol{x}^\top\beta_i)^\top]_{kk} &= \boldsymbol{E}_{\boldsymbol{z}}(\sum_{j=1}^d \beta_{i,j}\sigma_j\boldsymbol{z}_j)^2\sigma_k^2\boldsymbol{z}_k^2 \\ &= \mathbf{tr}(\beta_i^\top\boldsymbol{\Sigma}_F\beta_i)\sigma_k^2 + 2\sigma_k^4\beta_{i,k}^2.\end{aligned}$$

So that

$$\begin{aligned}\boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{x}\boldsymbol{x}^\top\beta_i)(\boldsymbol{x}\boldsymbol{x}^\top\beta_i)^\top &= 2\boldsymbol{\Sigma}_F\beta_i\beta_i^\top\boldsymbol{\Sigma}_F + \mathbf{tr}(\beta_i^\top\boldsymbol{\Sigma}_F\beta_i), \\ \mathbf{Cov}(\boldsymbol{x}\boldsymbol{x}^\top\beta_i) &= \boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{x}\boldsymbol{x}^\top\beta_i)(\boldsymbol{x}\boldsymbol{x}^\top\beta_i)^\top - \boldsymbol{\Sigma}_F\beta_i\beta_i^\top\boldsymbol{\Sigma}_F \\ &= \boldsymbol{\Sigma}_F\beta_i\beta_i^\top\boldsymbol{\Sigma}_F + \mathbf{tr}(\beta_i^\top\boldsymbol{\Sigma}_F\beta_i)\boldsymbol{\Sigma}_F.\end{aligned}$$

We plug it back into (D.41) and (D.39) and get

$$\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top = \boldsymbol{\Sigma}_F\beta_i\beta_i^\top\boldsymbol{\Sigma}_F + n_1^{-1}(\boldsymbol{\Sigma}_F\beta_i\beta_i^\top\boldsymbol{\Sigma}_F + \mathbf{tr}(\beta_i^\top\boldsymbol{\Sigma}_F\beta_i)\boldsymbol{\Sigma}_F + \sigma^2\boldsymbol{\Sigma}_F).$$

Define $\bar{\mathbf{\Sigma}}_T = \frac{1}{T}\sum_{j=1}^T \beta_j\beta_j^\top$. So that

$$\begin{aligned}
\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\hat{\mathbf{G}} &= \boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}T^{-1}(\sum_{i=1}^T \hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top) \\
&= \mathbf{\Sigma}_F\bar{\mathbf{\Sigma}}_T\mathbf{\Sigma}_F + n_1^{-1}(\mathbf{\Sigma}_F\bar{\mathbf{\Sigma}}_T\mathbf{\Sigma}_F + \mathbf{tr}(\bar{\mathbf{\Sigma}}_T\mathbf{\Sigma}_F)\mathbf{\Sigma}_F + \sigma^2\mathbf{\Sigma}_F) := \bar{\mathbf{G}}. \\
\boldsymbol{E}_\beta\hat{\mathbf{G}} &= \mathbf{G}.
\end{aligned}$$

We fix all β_i and study $\boldsymbol{E}_{\boldsymbol{x},y,\varepsilon}\hat{\mathbf{G}}$. Now we need to show how fast $\hat{\mathbf{G}}$ converges to $\bar{\mathbf{G}}$.

Define

$$\begin{aligned}
\boldsymbol{Z}_i &= \hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top - \boldsymbol{E}_{\boldsymbol{x}}(\hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top) \\
&= (\mathbf{\Sigma}_F\beta_i + \boldsymbol{\delta}_i)(\mathbf{\Sigma}_F\beta_i + \boldsymbol{\delta}_i)^\top - \boldsymbol{E}_{\boldsymbol{x}}(\mathbf{\Sigma}_F\beta_i + \boldsymbol{\delta}_i)(\mathbf{\Sigma}_F\beta_i + \boldsymbol{\delta}_i)^\top \\
&= \mathbf{\Sigma}_F\beta_i\boldsymbol{\delta}_i^\top + \boldsymbol{\delta}_i(\mathbf{\Sigma}_F\beta_i)^\top + \boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top - \boldsymbol{E}_{\boldsymbol{x}}(\mathbf{\Sigma}_F\beta_i\boldsymbol{\delta}_i^\top + \boldsymbol{\delta}_i(\mathbf{\Sigma}_F\beta_i)^\top + \boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top).
\end{aligned}$$

Then

$$\|\boldsymbol{E}\boldsymbol{Z}_i^2\| \le \|\boldsymbol{E}(\mathbf{\Sigma}_F\beta_i\boldsymbol{\delta}_i^\top + \boldsymbol{\delta}_i(\mathbf{\Sigma}_F\beta_i)^\top)^2\| + \|\boldsymbol{E}\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\|.$$

Then we can use (D.37) and (D.27) to bound the first term

$$\|\boldsymbol{E}\boldsymbol{Z}_i^2\| \lesssim n_1^{-1}\log^6(N)(\mathbf{tr}(\mathbf{\Sigma}_F\mathbf{\Sigma}_T) + \sigma)\mathbf{tr}(\mathbf{\Sigma}_F)\mathbf{tr}(\mathbf{\Sigma}_F^2\mathbf{\Sigma}_T)\|\mathbf{\Sigma}_F\|^2 + \|\boldsymbol{E}\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\| \tag{D.42}$$

So we need to bound $\|\boldsymbol{E}\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\|$. Note that $\boldsymbol{\delta}_i$ is the average of $\boldsymbol{x}_{i,j}(\boldsymbol{x}_{i,j}^\top\beta_i + \varepsilon_{ij})$ with respect to index $j = 1, ..., n_1$. So we just let $\boldsymbol{x} \sim \mathcal{N}(0, \mathbf{\Sigma}_F)$ and study $\boldsymbol{x}(\boldsymbol{x}^\top\beta_i + \varepsilon_{ij})$. Denote it by $\boldsymbol{u}_i$.

$$\begin{aligned}\|\boldsymbol{E}_{\boldsymbol{x}}\boldsymbol{u}_i\boldsymbol{u}_i^\top\boldsymbol{u}_i\boldsymbol{u}_i^\top\| &= \|\boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{x}^\top\beta_i+\varepsilon_{ij})^4\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{x}\boldsymbol{x}^\top\| \\ &\lesssim \|\boldsymbol{E}_{\boldsymbol{x}}((\boldsymbol{x}^\top\beta_i)^4+\sigma^4)\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{x}\boldsymbol{x}^\top\| \\ &\lesssim (\mathbf{tr}^2(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^4)\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|.\end{aligned}$$

So that

$$\|\boldsymbol{E}\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\| \lesssim n_1^{-2}(\mathbf{tr}^2(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^4)\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|.$$

Now we can go back to (D.42) and get

$$\|\boldsymbol{E}\boldsymbol{Z}_i^2\| \lesssim n_1^{-1}\log^6(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)^2\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|^2.$$

Next we need to bound the norm of $\boldsymbol{Z}_i$. We use (D.37) and (D.27), with probability $1-N^{-c}$,

$$\begin{aligned}\|\boldsymbol{Z}_i\| \le\, & n_1^{-1/2}\log^3(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_F)}\|\boldsymbol{\Sigma}_F\| \\ &+ n_1^{-1}\log^5(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\mathbf{tr}(\boldsymbol{\Sigma}_F).\end{aligned}$$

Define the upper bound for $\|\boldsymbol{E}\boldsymbol{Z}_i^2\|, \|\boldsymbol{Z}_i\|$ as Z_1, Z_2 (the right hand side of two above inequalities). With Bernstein type inequality (Lemma 27),with probability $1 - N^{-c}$,

$$
\begin{aligned}
&\|\hat{\mathbf{G}} - \bar{\mathbf{G}}\| \\
&= \|T^{-1}\sum_{i=1}^{T} \boldsymbol{Z}_i - \boldsymbol{E}_{\boldsymbol{x}}\boldsymbol{Z}_i\| \\
&\lesssim \log(TZ_2)\Big(T^{-1/2}\log(N)Z_1^{1/2} + T^{-1}Z_2\log(TZ_2)\Big) \\
&\lesssim \log(TZ_2)\Big(\sqrt{\frac{\log^6(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)^2\mathbf{tr}(\boldsymbol{\Sigma}_F)\|\boldsymbol{\Sigma}_F\|^2}{n_1T}} \\
&\quad + \frac{\log^3(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_F)}\|\boldsymbol{\Sigma}_F\|}{n_1^{1/2}T} \\
&\quad + \frac{\log^5(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\mathbf{tr}(\boldsymbol{\Sigma}_F)}{T}\Big) \\
&= \log(TZ_2)\cdot\Big(\log^3(N)\|\boldsymbol{\Sigma}_F\|(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\sqrt{\frac{\mathbf{tr}(\boldsymbol{\Sigma}_F)}{N}} \\
&\quad + \frac{\log^5(N)(\mathbf{tr}(\boldsymbol{\Sigma}_F^2\boldsymbol{\Sigma}_T)+\mathbf{tr}(\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_T)+\sigma^2)\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_F)}\|\boldsymbol{\Sigma}_F\|}{N^{1/2}T^{1/2}}\Big).
\end{aligned}
$$

□

D.4 Proof of robustness of optimal representation

Theorem 11. *Let $\boldsymbol{\Lambda}_{\underline{\theta}}(R)$, $\boldsymbol{\Lambda}^*_{\underline{\theta}}(R)$ be as defined above, and $r_F = \mathbf{tr}(\boldsymbol{\Sigma}_F)$, $r_T = \mathbf{tr}(\boldsymbol{\Sigma}_T), r = \mathbf{tr}(\tilde{\boldsymbol{\Sigma}}_T)$. The risk of meta-learning algorithm satisfies*[4]

$$
risk(\boldsymbol{\Lambda}_{\underline{\theta}}(R), \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F) - risk(\boldsymbol{\Lambda}^*_{\underline{\theta}}(R), \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F) \lesssim \frac{n_2^2}{d(R-n_2)(2n_2 - R\underline{\theta})\underline{\theta}}\left[(r+\sigma^2)\sqrt{\frac{r_F}{N}} + \sqrt{\frac{r_T}{T}}\right].
$$

Proof. In the proof below, we use $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^*$ to replace $\boldsymbol{\Lambda}_{\underline{\theta}}(R), \boldsymbol{\Lambda}^*_{\underline{\theta}}(R)$ for simplicity. We first

[4] The bracketed expression applies first conclusion of Theorem 11. One can plug in the second as well.

decompose the risk as

$$\begin{aligned}
&\operatorname{risk}(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F) - \operatorname{risk}(\boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F) \\
= &\underbrace{\operatorname{risk}(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F) - \operatorname{risk}(\boldsymbol{\Lambda}^*, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F)}_{\leq 0} \\
&+ [\operatorname{risk}(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F) - \operatorname{risk}(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F)] + [\operatorname{risk}(\boldsymbol{\Lambda}^*, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F) - \operatorname{risk}(\boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F)].
\end{aligned}$$

We know $\operatorname{risk}(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F) - \operatorname{risk}(\boldsymbol{\Lambda}^*, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F) \leq 0$ due to the optimality of $\boldsymbol{\Lambda}$ with task covariance $\hat{\boldsymbol{\Sigma}}_T$. Now we will bound $\operatorname{risk}(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F) - \operatorname{risk}(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F)$ for arbitrary $\boldsymbol{\Lambda}$, and it automatically works for $\operatorname{risk}(\boldsymbol{\Lambda}^*, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F) - \operatorname{risk}(\boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F)$. Note that in (5.4) we know that

$$\operatorname{risk}(\boldsymbol{\Lambda}', \boldsymbol{\Sigma}'_T) = f(\boldsymbol{\theta}; \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F) := \sum_{i=1}^{R} \frac{n_2(1-\boldsymbol{\theta}_i)^2}{R(n_2 - \|\boldsymbol{\theta}\|^2)} \tilde{\boldsymbol{\Sigma}}^R_{T,i} + \frac{n_2}{n_2 - \|\boldsymbol{\theta}\|^2}\sigma^2. \tag{D.43}$$

This function is linear in $\boldsymbol{\Sigma}_T$ thus we know that

$$|\operatorname{risk}(\boldsymbol{\Lambda}^*, \hat{\boldsymbol{\Sigma}}_T, \boldsymbol{\Sigma}_F) - \operatorname{risk}(\boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_F)| \leq \frac{n_2}{d(n_2 - \|\boldsymbol{\theta}\|^2)}\mathcal{E}. \tag{D.44}$$

Now we need to bound $\|\boldsymbol{\theta}\|^2$. With the constraint $\underline{\theta} \leq \boldsymbol{\theta} < 1 - \frac{R-n_2}{n_2}\underline{\theta}$ and $\sum \boldsymbol{\theta}_i = n_2$, we know that the maximum of $\|\boldsymbol{\theta}\|^2$ happens when $(R - n_2)$ among $\boldsymbol{\theta}_i$ are $\underline{\theta}$ and the others are $1 - \frac{R-n_2}{n_2}\underline{\theta}$. With this we have

$$\begin{aligned}
\|\boldsymbol{\theta}\|^2 &\leq (R-n_2)\underline{\theta}^2 + n_2(1 - \frac{R-n_2}{n_2}\underline{\theta})^2 \\
&= (R-n_2)\underline{\theta}^2 + n_2 - 2(R-n_2)\underline{\theta} + \frac{(R-n_2)^2}{n_2}\underline{\theta}^2 \\
&= n_2 - 2(R-n_2)\underline{\theta} + \frac{(R-n_2)R}{n_2}\underline{\theta}^2
\end{aligned}$$

Thus

$$n_2 - \|\boldsymbol{\theta}\|^2 \geq (R - n_2)\underline{\theta}(2n_2 - R\underline{\theta}).$$

Plugging it into (D.44) and (D.43) leads to the theorem.

□