HiLD 2026: 4th Workshop on High-dimensional Learning Dynamics 1–38

Training Dynamics of Softmax Self-Attention: Fast Global Convergence via Preconditioning

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

We study the training dynamics of gradient descent in a softmax self-attention layer trained to perform linear regression and propose a first-order optimization algorithm which converges to the globally optimal self-attention parameters at a geometric rate. Our analysis proceeds in two steps. First, we show that in the infinite-data limit the regression problem solved by the self-attention layer is equivalent to a nonconvex matrix factorization problem. Second, we exploit this connection to design a novel "structure-aware" variant of gradient descent which efficiently optimizes the original finite-data regression objective. Our optimization algorithm features several innovations over vanilla gradient descent, including a reparameterization which ensures that the empirical gradient remains close to the population gradient throughout training, a data-dependent preconditioner and data-dependent regularizer which help avoid spurious stationary points, and a spectral initialization of parameters which lie near the manifold of global minima with high probability. We prove that the excess risk of the model trained by our algorithm decreases exponentially fast in the number of gradient descent iterations, up to an additional error term that decreases as $1/n$ in the size of the training set.

1. Introduction

The self-attention mechanism is a neural architecture originally proposed by Bahdanau et al. (2014) for machine translation. It was subsequently adopted by Vaswani et al. (2017) to form the basis of the Transformer architecture, which underlies many recent advances in natural language processing (OpenAI, 2023) and computer vision (Dosovitskiy, 2020). It has proven to be remarkably versatile, and can additionally be trained to mimic various algorithms from statistics, optimization, and machine learning (Garg et al., 2022). Despite its numerous empirical successes, our theoretical understanding of the self-attention mechanism remains poor. Many of the prior works on the theoretical behavior of self-attention are conditional; they prove that *if* the self-attention parameters could be optimized to their globally optimal values then the resulting model would exhibit strong performance on various downstream tasks, but they do not establish *when* such optimization is possible or *how* it should be performed, e.g., (Bai et al., 2024; Li et al., 2023). Analysis of the training dynamics of gradient descent in a softmax self-attention layer is exceedingly challenging due to the pairwise nonlinear interactions appearing in the softmax function.

A recent line of theoretical works (e.g., (Zhang et al., 2024; Ahn et al., 2024, 2023; Chen et al., 2024)) seek to understand the optimization dynamics of self-attention in the setting of random linear regression[1] which was empirically investigated by Garg et al. (2022) and

1. The term 'random linear regression' is a misnomer, since we are studying a setting where a nonlinear model is used to fit data which is generated by a planted linear model (i.e., nonlinear regression). We adopt this terminology to remain consistent with preexisting literature.

© .

Von Oswald et al. (2023). In this model, the covariates are drawn from a distribution, and the response variables are a noisy linear function of the covariates. The performance of a predictor is measured using the square loss. A natural question is whether a self-attention mechanism can be trained to accurately predict the label corresponding to a given covariate; even in this simple setting, this is not at all obvious due to the nonconvexity of the loss in the model parameters. While several of these papers derive various global convergence guarantees, they suffer from two drawbacks. First, most of these works only study a simplified, linearized variant of self-attention instead of the original softmax attention mechanism of Bahdanau et al. (2014). Second, all of these works only study the optimization dynamics in an asymptotic limit where either the learning algorithm has access to infinitely many samples, or has an unlimited budget of gradient iterations to converge to optimality; none of these works quantify how model performance depends on the number of samples or the compute budget. In this paper we address both of these challenges.

1.1. Main Contributions

We consider a setting where the number of self-attention parameters is fixed, and bound the generalization error as a function of the size of the training set n and the number of gradient descent iterations m. Our contributions can be summarized as follows.

In Section 3, we study the population loss, i.e., the asymptotic limit of the training loss as n approaches infinity. We show that this loss has a simple closed-form description and show that it is equivalent to a certain weighted matrix factorization loss. Using ideas from the matrix factorization literature, we propose a novel regularizer of the population loss and show that the regularized population loss has infinitely many global minima which together form a smooth connected manifold. While this regularized loss is globally nonconvex, we prove that it exhibits one-point strong convexity and one-point smoothness near the manifold of global minima in a certain geometry in which the inner product between two points is weighted by the covariance of the data distribution.

In Section 4 we leverage the geometric results of Section 3 to design a "structure-aware" gradient descent algorithm which finds a set of model parameters with near-optimal population loss. Our algorithm features several innovations over standard optimization algorithms such as SGD and Adam. First, our algorithm initializes the parameters of the self-attention layer using spectral information from the training data; we show that these parameters lie near the manifold of global optima with high probability. Second, our algorithm incorporates a data-dependent regularizer and a data-dependent preconditioner which mirror the form of the regularizer and preconditioner described in Theorem 2. Our algorithm can be viewed as evolving each of the self-attention parameters in the geometry most natural to that parameter. We next present our main result: a scaling law which describes how the generalization error decreases as a function of n and m. We decompose the excess risk of the model trained by our optimization algorithm into two pieces: a statistical error term, which measures how well the random, finite-data empirical loss approximates the population loss, and an optimization error term, which measures the distance between the parameters found by our model and the globally optimal parameters. We show that the statistical error decreases at a fast $1/n$ rate, up to logarithmic factors, and the optimization error decays exponentially in m. To the best of our knowledge, this is the first result which

establishes fast (i.e., geometric rate) global convergence of a first-order method on a softmax self-attention training objective in any setting, and also the first result which establishes a sharp $1/n$ statistical rate in any setting.

In the course of proving our main results, we establish novel concentration inequalities which bound how often the softmax-weighted mean and softmax-weighted covariance deviate far from their respective means; these results may be of independent interest.

2. Model

We study regression using the square loss, where the covariates are d-dimensional and the response variables are p-dimensional. Specifically, we consider a setting where we are given n samples $\{x_i, y_i\}_{i=1}^n$, where each x_i is drawn independently from $\mathcal{N}(0, \Sigma)$ and each response y_i has the form $y_i = Mx_i + z_i$ for some fixed weight matrix $M \in \mathbb{R}^{p\times d}$. The noise variables $\{z_i\}_{i=1}^n$ are drawn i.i.d. and independently from the covariates from $\mathcal{N}(0, \Omega)$. Our goal is to learn a prediction rule which, when given a fresh covariate $x \sim \mathcal{N}(0, \Sigma)$, generates a prediction $\hat{y}$ which is close to Mx. We consider the family of regression functions consisting of single-layer single-head softmax self-attention functions; such functions are parameterized by $\theta = (A, B)$, where $A \in \mathbb{R}^{p\times d}$ and $B \in \mathbb{R}^{d\times d}$. We think of θ as the vertical concatenation A and B, so that $\theta \in \mathbb{R}^{(p+d)\times d}$. Given a fresh covariate $x \in \mathbb{R}^d$, such functions predict a corresponding label $\hat{y}$ given by

$$\hat{y} = A\left(\frac{\sum_{j=1}^n \exp(x^\top Bx_j)x_j}{\sum_{j=1}^n \exp(x^\top Bx_j)}\right).$$

The prediction $\hat{y}$ is the image of a convex combination of the covariates $\{x_j\}_{j=1}^n$ under the linear map A, where the weights of this convex combination are determined by the nonlinear softmax function parameterized by B. We note that in the language of the original self-attention paper (Vaswani et al., 2017), the parameter A is called the value matrix, while B is the product of the key and query matrices.

We define the in-sample empirical loss

$$\hat{L}(\theta) = \frac{1}{2n}\sum_{i=1}^n \left\| A\left(\frac{\sum_{j=1}^n \exp(x_i^\top Bx_j)x_j}{\sum_{j=1}^n \exp(x_i^\top Bx_j)}\right) - y_i\right\|_2^2. \tag{1}$$

We also define the population loss

$$L(\theta) = \frac{1}{2}\mathbb{E}_{x_1,z_1}\left\| A\left(\frac{\mathbb{E}_{x_2}[\exp(x_1^\top Bx_2)x_2]}{\mathbb{E}_{x_2}[\exp(x_1^\top Bx_2)]}\right) - (Mx_1 + z_1)\right\|_2^2, \tag{2}$$

where x_1 and x_2 are sampled independently from $\mathcal{N}(0, \Sigma)$ and z_1 is sampled from $\mathcal{N}(0, \Omega)$. The population loss has the following intuitive interpretation. When n is large, we expect that each of the summations appearing in the numerator and the denominator of the predictor $\hat{y}$ should approach their respective expectations. Averaging over the individual losses, we obtain the population loss.

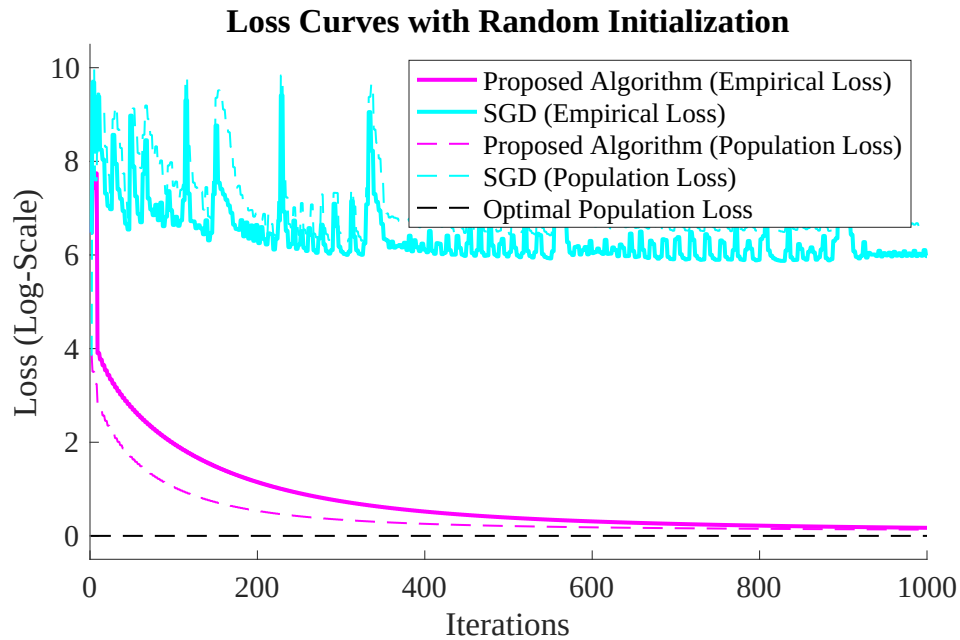

Figure 1: We plot the training losses and test losses incurred by our algorithm and SGD, where both algorithms are initialized at the same random point. Notice that SGD has an extremely spiky loss curve due to the unstable behavior of the softmax function. The loss curve associated to our algorithm is much smoother due to reparameterization.

3. Structure of the population loss

We introduce the reparameterization

$$A = \hat{\gamma}^{-1}\|\tilde{B}\|_F^A, \qquad B = \hat{\gamma}\|\tilde{B}\|_F^{-1}\tilde{B},$$

and now think of the loss as being parameterized by the parameters $\theta = (\tilde{A}, \tilde{B})$. We characterize the population loss $L(\theta)$ in closed form, and show that a regularized variant $Q(\theta) = L(\theta) + R(\theta)$ of the population loss obeys certain convexity and smoothness properties near its minima. A crucial aspect of our result is that these properties hold only in a certain geometry in which the inner product and norm are reweighted by the covariance Σ. As a consequence, our convexity and smoothness results are not stated in terms of the raw gradient $\nabla Q(\theta)$, but rather in terms of the preconditioned gradient $P^{-1}\nabla Q$, where P^{-1} is a preconditioner depending on Σ. A complete characterization of the population loss appears in Section A.2.

4. Main Result

We propose a first-order algorithm (formally described in the display Algorithm 1) and show that the population loss approaches its optimal value exponentially fast in the number of

Algorithm 1: Preconditioned Gradient Descent for Self-Attention

Input: Data $\{(x_i, y_i)\}_{i=1}^n$, step size $\eta > 0$, iteration budget m, failure probability δ

Output: Self-Attention parameters (A, B)

$$\hat{\Sigma} \leftarrow n^{-1} \sum_{i=1}^n x_i x_i^\top, \qquad \hat{M} \leftarrow n^{-1} \sum_{i=1}^n y_i x_i^\top \hat{\Sigma}^{-1},$$

$$\hat{U}\hat{\Gamma}\hat{V}^\top \leftarrow \mathrm{SVD}(\hat{M}\hat{\Sigma}^{1/2}),$$

$$\tilde{A}_0 \leftarrow \hat{U}\hat{\Gamma}^{1/2}\hat{\Sigma}^{-1/2}, \qquad \tilde{B}_0 \leftarrow \hat{\Sigma}^{-1/2}\hat{V}\hat{\Gamma}^{1/2}\hat{\Sigma}^{-1/2},$$

$$\hat{\nu} \leftarrow \max\left(1, \sqrt{\mathrm{Tr}(\hat{\Sigma})} + 2\sqrt{2\|\hat{\Sigma}\|_{\mathrm{op}} \log n}\right),$$

$$p \leftarrow \max(2, \log(1/\delta)),$$

$$\hat{\gamma} \leftarrow \frac{1}{2}\hat{\nu}^{-1} p^{-1/2} \|\hat{\Sigma}\|_{\mathrm{op}}^{-1/2},$$

$$A(\tilde{A}, \tilde{B}) \leftarrow \hat{\gamma}^{-1}\|\tilde{B}\|_F \tilde{A}, \qquad B(\tilde{B}) \leftarrow \hat{\gamma}\|\tilde{B}\|_F^{-1}\tilde{B},$$

$$\pi_{ij}(\tilde{A}, \tilde{B}) \leftarrow \frac{\exp\left(x_i^\top B(\tilde{B}) x_j\right)}{\sum_{\ell=1}^n \exp\left(x_i^\top B(\tilde{B}) x_\ell\right)},$$

$$\tilde{L}(\tilde{A}, \tilde{B}) \leftarrow \frac{1}{2n} \sum_{i=1}^n \left\| A(\tilde{A}, \tilde{B}) \sum_{j=1}^n \pi_{ij}(\tilde{A}, \tilde{B}) x_j - y_i \right\|_2^2,$$

$$\tilde{R}(\tilde{A}, \tilde{B}) \leftarrow \frac{1}{8} \left\| \hat{\Sigma}^{1/2}\left(\tilde{A}^\top \tilde{A} - \tilde{B}^\top \hat{\Sigma} \tilde{B}\right)\hat{\Sigma}^{1/2} \right\|_F^2,$$

$$\tilde{Q}(\tilde{A}, \tilde{B}) \leftarrow \tilde{L}(\tilde{A}, \tilde{B}) + \tilde{R}(\tilde{A}, \tilde{B}).$$

for $t = 1, \ldots, m$ **do**

$\quad \tilde{A}_t \leftarrow \tilde{A}_{t-1} - \eta \nabla_{\tilde{A}} \tilde{Q}(\tilde{A}_{t-1}, \tilde{B}_{t-1})$

$\quad \tilde{B}_t \leftarrow \tilde{B}_{t-1} - \eta \hat{\Sigma}^{-1} \nabla_{\tilde{B}} \tilde{Q}(\tilde{A}_{t-1}, \tilde{B}_{t-1})$

end

return $\left(A(\tilde{A}_m, \tilde{B}_m),\ B(\tilde{B}_m)\right)$

gradient descent iterations, up to an additional error term that decreases as $1/n$ in the size of the training set.

Theorem 1 *Suppose the step size η is set sufficiently small. Let θ_m denote the parameters generated by Algorithm 1 after m iterations. With probability $1 - \delta$, the following inequality holds for all sufficiently large n:*

$$L(\theta_m) - L^\star \leq K_1 \frac{\log^5(n) \log^2(mn/\delta)}{n} + K_2 \exp(-K_3\, m),$$

where $K_1, K_2, K_3 \geq 0$ are constants depending on M and Σ and independent of m, n, δ.

We defer the proof to the Appendix.

Appendix A. Appendix

A.1. Optimization

A.2. Proof of Theorem 1

Proof We assume that the "good events" described in Lemma 24 occur. These events imply that Algorithm 1 selects an initial set of parameters which is close to the manifold of global minima, and that the empirical quantities $\hat{\Sigma}, \hat{\gamma}$, etc. are close to their population counterparts. Lemma 24 shows that these events simultaneously occur with probability $1 - \delta/2$, provided that n is sufficiently large. Let $Q(\theta)$ and $\mathcal{S}$ be as described in Theorem 2. Define the empirical regularized loss

$$\hat{Q}(\theta) = \hat{L}(\theta) + \hat{R}(\theta),$$

where we define

$$\hat{L}(\tilde{A}, \tilde{B}) = \frac{1}{2n} \sum_{i=1}^{n} \left\| A \frac{\sum_{j=1}^{n} \exp(x_i^\top B x_j) x_j}{\sum_{j=1}^{n} \exp(x_i^\top B x_j)} - y_i \right\|_2^2,$$
$$\hat{R}(\tilde{A}, \tilde{B}) = \frac{1}{8} \left\| \hat{\Sigma}^{1/2} (\tilde{A}^\top \tilde{A} - \tilde{B}^\top \hat{\Sigma} \tilde{B}) \hat{\Sigma}^{1/2} \right\|_F^2$$

and set

$$A = \hat{\gamma}^{-1} \|\tilde{B}\|_F \tilde{A}, \qquad B = \hat{\gamma} \|\tilde{B}\|_F^{-1} \tilde{B}.$$

Define

$$\hat{Z}_t = \hat{P}^{-1} \nabla \hat{Q}(\theta_t).$$

Notice that the gradient update used by Algorithm 1 is exactly

$$\theta_{t+1} = \theta_t - \eta \hat{Z}_t.$$

Let $\theta_t^\star$ denote a projection in P-norm of the point θ_t onto $\mathcal{S}$. Define the residual

$$\Delta_t = \theta_t - \theta_t^\star$$

and the potential

$$\phi_t = \|\Delta_t\|_P^2 .$$

Notice that on the events described in Lemma 24, $\phi_0 \leq \varepsilon_0$. We will show that in each iteration, ϕ_t shrinks by a multiplicative factor, up to a small additive error. This will show that the sequence of iterates $\{\theta_t\}_{t=1}^m$ converges (approximately) to $\mathcal{S}$, which implies that $L(\theta_m)$ is near the optimal loss $L^\star$. It is clear that

$$\phi_{t+1} \leq \|\theta_{t+1} - \theta_t^\star\|_P^2 ,$$

because the point $\theta_t^\star$ can only be farther away from θ_{t+1} than $\theta_{t+1}^\star$. Applying Lemma 5, we see that

$$\begin{aligned} \phi_{t+1} &\leq \left\| \theta_t - \eta \hat{Z}_t - \theta_t^\star \right\|_P^2 \\ &= \|\Delta_t\|_P^2 + \eta^2 \|\hat{Z}_t\|_P^2 - 2\eta \langle \hat{Z}_t, \Delta_t \rangle_P \\ &= \mu(\eta) \|\Delta_t\|_P^2 + \nu \|\xi_t\|_P^2 + \nu \|\zeta_t\|_P^2, \end{aligned}$$

where we defined the random variables ξ_t, ζ_t as in Lemma 5 and set

$$\mu(\eta) = (1 - \alpha\eta + 3\beta\eta^2)$$

and

$$\nu = 2\alpha^{-1}\eta + 3\eta^2.$$

Optimizing over η, we see that $\mu(\eta)$ is minimized when $\eta = \frac{\alpha}{6\beta}$, in which case $\mu(\eta) = 1 - \frac{\alpha^2}{12\beta}$. We emphasize that $0 < \mu < 1$. Theorem 7 shows that the variables $\{\xi_t, \zeta_t\}_{t=0}^{m-1}$ are uniformly bounded with probability $1 - \delta/2$:

$$\sup_{t\in[m]} \max\left\{\|\xi_{t-1}\|_P^2,\ \|\zeta_{t-1}\|_P^2\right\} \leq \frac{K\log^5(n)\log^2(mn/\delta)}{n}$$

for some constant $K \geq 0$. A standard inductive argument shows that

$$\phi_m \leq \frac{2\nu K\log^5(n)\log^2(mn/\delta)}{(1-\mu)n} + \varepsilon_0\mu^m.$$

Applying Lemma 6, we see that this implies that

$$Q(\theta_m) - Q^\star \leq \frac{K_1\log^5(n)\log^2(mn/\delta)}{n} + K_2\mu^m$$

for some $K_1, K_2 \geq 0$, where we set $Q^\star = \inf_\theta Q(\theta)$. Observing that $Q(\theta_m) \geq L(\theta_m)$ and $Q^\star = L^\star$ finishes the proof.

■

A.3. Structure of the Population Loss

Theorem 2 *The population loss $L(\theta)$ and the regularized population loss $Q(\theta)$ have the following properties:*

1. *The population loss can be written as*

$$L(\theta) = L^\star + \frac{1}{2}\left\|\tilde{A}\Sigma\tilde{B}^\top\Sigma^{1/2} - M\Sigma^{1/2}\right\|_F^2,$$

where $L^\star = \frac{1}{2}\operatorname{Tr}(\Omega)$ is the irreducible loss.

2. *Define the regularized population loss*

$$Q(\theta) = L(\theta) + R(\theta),$$

where we set

$$R(\theta) = \frac{1}{8}\left\|\Sigma^{1/2}(\tilde{A}^\top\tilde{A} - \tilde{B}^\top\Sigma\tilde{B})\Sigma^{1/2}\right\|_F^2.$$

Let $U\Gamma V^\top$ be a singular value decomposition of $M\Sigma^{1/2}$, where $U \in \mathbb{R}^{p\times d}$ and $V \in \mathbb{R}^{d\times d}$ satisfy $U^\top U = V^\top V = I_d$ and $\Gamma \in \mathbb{R}^{d\times d}$ is diagonal. Let $\mathcal{S}$ be the smooth manifold consisting of points of the form

$$\begin{bmatrix}\tilde{A}\\ \tilde{B}\end{bmatrix} = \begin{bmatrix}U\Gamma^{1/2}J^\top\Sigma^{-1/2}\\ \Sigma^{-1/2}V\Gamma^{1/2}J^\top\Sigma^{-1/2}\end{bmatrix}$$

for some $J \in \mathbb{O}_d$. Each point $\theta \in \mathcal{S}$ is a global minimum of $Q(\cdot)$, and in particular satisfies $L(\theta) = L^\star$ and $R(\theta) = 0$.

3. *Define the $(p+d)\times(p+d)$ block-diagonal matrix $P = \mathrm{diag}(I_p, \Sigma)$. Define the P-weighted inner product*
$$\langle \theta_1, \theta_2 \rangle_P = \mathrm{Tr}(\theta_1^\top P \theta_2)$$
and the associated P-norm
$$\|\theta\|_P = \sqrt{\langle \theta, \theta \rangle_P}.$$
The regularized population loss $Q(\theta)$ exhibits the following "one-point strong convexity" and "one-point smoothness" properties. Let $\theta^\star$ denote a projection in the P-norm of θ onto $\mathcal{S}$. Let
$$\varepsilon_0 = \min\left\{1,\ \frac{K_0}{12K_1\|\Sigma\|_{\mathrm{op}}},\ \sqrt{\frac{K_0}{8\|\Sigma\|_{\mathrm{op}}^2}}\right\}.$$
For all θ which are ε_0-close to $\mathcal{S}$ in the P-norm, the following bounds hold:
$$\alpha\|\theta - \theta^\star\|_P^2 \le \left\langle P^{-1}\nabla Q(\theta), \theta - \theta^\star \right\rangle_P, \qquad \textit{(one-point strong convexity)} \tag{3}$$
$$\|P^{-1}\nabla Q(\theta)\|_P^2 \le \beta\|\theta - \theta^\star\|_P^2, \qquad \textit{(one-point strong smoothness)} \tag{4}$$
where we define
$$\alpha = \frac{K_0}{4}, \qquad \beta = \left(\frac{7}{2} + \frac{7}{4}\kappa^2(\Sigma)\right)K_1^2 + \frac{21}{4}\|\Sigma\|_{\mathrm{op}}^2 K_1 + \frac{7}{4}\|\Sigma\|_{\mathrm{op}}^4,$$
and K_0, K_1 are defined as in Lemma 4.

Proof To prove the first part of Theorem 2, we first recall that the population loss is
$$L(\theta) = \frac{1}{2}\mathbb{E}_{x_1,z_1}\left\|A\left(\frac{\mathbb{E}_{x_2}[\exp(x_1^\top B x_2)x_2]}{\mathbb{E}_{x_2}[\exp(x_1^\top B x_2)]}\right) - (Mx_1 + z_1)\right\|_2^2,$$
where
$$A = \hat{\gamma}^{-1}\|\tilde{B}\|_F\tilde{A}, \qquad B = \hat{\gamma}\|\tilde{B}\|_F^{-1}\tilde{B}.$$
Each of the expectations over x_2 can be computed using a standard completion-of-squares argument:
$$\mathbb{E}_{x_2}[\exp(x_1^\top B x_2)x_2] = \exp\left(\frac{1}{2}x_1^\top B\Sigma B^\top x_1\right)\Sigma B^\top x_1, \quad \mathbb{E}_{x_2}[\exp(x_1^\top B x_2)] = \exp\left(\frac{1}{2}x_1^\top B\Sigma B^\top x_1\right).$$
Canceling terms, we see that $L(\theta)$ can be written as
$$L(\theta) = \frac{1}{2}\mathbb{E}_{x_1,z_1}\left\|A\Sigma B^\top x_1 - (Mx_1 + z_1)\right\|_2^2.$$
Integrating with respect to x_1 and z_1 and using the fact that $\mathbb{E}[x_1 z_1^\top] = 0$ because x_1 and z_1 are independent, we see that
$$L(\theta) = \frac{1}{2}\mathrm{Tr}(\Omega) + \frac{1}{2}\left\|A\Sigma B^\top\Sigma^{1/2} - M\Sigma^{1/2}\right\|_F^2.$$

Recall that we assume the reparameterization

$$A = \hat{\gamma}^{-1}\|\tilde{B}\|_F \tilde{A}, \qquad B = \hat{\gamma}\|\tilde{B}\|_F^{-1}\tilde{B}.$$

It follows that

$$L(\theta) = \frac{1}{2}\operatorname{Tr}(\Omega) + \frac{1}{2}\left\|\tilde{A}\Sigma\tilde{B}^\top\Sigma^{1/2} - M\Sigma^{1/2}\right\|_F^2.$$

We now turn to the second part of Theorem 2. Recall that $Q(\theta) = L(\theta) + R(\theta)$, where $L(\theta) \geq L^\star$ and $R(\theta) \geq 0$. It follows that any point $\theta = (A, B)$ such that $L(\theta) = L^\star$ and $R(\theta) = 0$ is a global minimizer of $Q(\theta)$. The condition $L(\theta) = L^\star$ implies that

$$\tilde{A}\Sigma\tilde{B}^\top\Sigma^{1/2} = M\Sigma^{1/2},$$

while the condition $R(\theta) = 0$ implies that

$$\tilde{A}^\top\tilde{A} = \tilde{B}^\top\Sigma\tilde{B}.$$

It is easy to check that all pairs $(\tilde{A}, \tilde{B})$ of the form

$$\begin{bmatrix}\tilde{A}\\ \tilde{B}\end{bmatrix} = \begin{bmatrix}U\Gamma^{1/2}J^\top\Sigma^{-1/2}\\ \Sigma^{-1/2}V\Gamma^{1/2}J^\top\Sigma^{-1/2}\end{bmatrix}$$

satisfy both equations. We now prove the third part of Theorem 2. The proof depends crucially on Lemma 3. The significance of this lemma is that is shows that the first-order condition satisfied by $\theta^\star$ implies a symmetry condition which allows us to establish strong convexity of $Q(\theta)$. This symmetry condition is the key reason why we choose the P-norm to measure the distance between θ and $\mathcal{S}$.

We first establish one-point strong convexity before proving one-point smoothness. Our proof is inspired by the proof of Theorem 3.3 in (Tu et al., 2016), but is modified to account for the fact that $L(\theta)$ not symmetric in $\tilde{A}$ and $\tilde{B}$ due to the extra factor of $\Sigma^{1/2}$ attached to $\tilde{B}$. An algebraic calculation shows that $Q(\theta)$ can be rewritten in the form

$$Q(\theta) = L^\star + \frac{1}{8}\left\|P^{1/2}(\theta\Sigma\theta^\top - 2\operatorname{Sym}(M))P^{1/2}\right\|_F^2 - \frac{1}{2}\|M\Sigma^{1/2}\|_F^2,$$

where we define

$$\operatorname{Sym}(M) = \begin{bmatrix}0 & M\\ M^\top & 0\end{bmatrix}.$$

Define

$$\tilde{\theta}^\star = \begin{bmatrix}I_p & 0\\ 0 & -I_d\end{bmatrix}\theta^\star.$$

Notice that

$$2\operatorname{Sym}(M) = \theta^\star\Sigma\theta^{\star\top} - \tilde{\theta}^\star\Sigma\tilde{\theta}^{\star\top} \tag{5}$$

and

$$\theta^{\star\top}P\tilde{\theta}^\star = 0. \tag{6}$$

Set $\Delta = \theta - \theta^\star$. Applying (5), we see that

$$\begin{aligned}
\langle P^{-1}\nabla Q(\theta), \Delta\rangle_P &= \langle \nabla Q(\theta), \Delta\rangle \\
&= \frac{1}{2}\left\langle P(\theta\Sigma\theta^\top - 2\mathrm{Sym}(M))P\theta\Sigma, \Delta\right\rangle \\
&= \frac{1}{2}\left\langle P(\theta\Sigma\theta^\top - \theta^\star\Sigma\theta^{\star\top})P\theta\Sigma, \Delta\right\rangle + \frac{1}{2}\left\langle P\tilde{\theta}^\star\Sigma\tilde{\theta}^{\star\top}P\theta\Sigma, \Delta\right\rangle. \qquad (7)
\end{aligned}$$

We lower-bound each term of (7) separately. It is convenient to lower-bound the second term of (7) first. Notice that

$$\frac{1}{2}\left\langle P\tilde{\theta}^\star\Sigma\tilde{\theta}^{\star\top}P\theta\Sigma, \Delta\right\rangle = \frac{1}{2}\mathrm{Tr}(\theta^\top P\tilde{\theta}^\star\Sigma\tilde{\theta}^{\star\top}P\theta\Sigma) - \frac{1}{2}\mathrm{Tr}(\theta^{\star\top}P\tilde{\theta}^\star\Sigma\tilde{\theta}^{\star\top}P\theta\Sigma).$$

The first term is non-negative because it is the trace of the product of two psd matrices. We see that the second term is equal to zero in light of (6). This proves that the second term of (7) is non-negative.

We now show that the first term of (7) is bounded below by a constant multiple of $\|\Delta\|_F^2$, provided that θ is sufficiently close to $\mathcal{S}$. We observe that

$$\begin{aligned}
\frac{1}{2}\left\langle P(\theta\Sigma\theta^\top - \theta^\star\Sigma\theta^{\star\top})P\theta\Sigma, \Delta\right\rangle &= \frac{1}{2}\mathrm{Tr}\left(\Delta^\top P(\theta\Sigma\theta^\top - \theta^\star\Sigma\theta^{\star\top})P\theta\Sigma\right) \\
&= \frac{1}{2}\mathrm{Tr}\left(\Delta^\top P(\theta^\star\Sigma\Delta^\top + \Delta\Sigma\theta^{\star\top} + \Delta\Sigma\Delta^\top)P(\theta^\star + \Delta)\Sigma\right) \\
&= S + T,
\end{aligned}$$

where we set S be the sum of terms which are quadratic in Δ and set T be the sum of all remaining terms:

$$\begin{aligned}
S &= \frac{1}{2}\mathrm{Tr}\left(\Delta^\top P\theta^\star\Sigma\Delta^\top P\theta^\star\Sigma + \Delta^\top P\Delta\Sigma\theta^{\star\top}P\theta^\star\Sigma\right), \\
T &= \frac{1}{2}\mathrm{Tr}\left(\Delta^\top P\Delta\Sigma\Delta^\top P\theta^\star\Sigma + \Delta^\top P\Delta\Sigma\Delta^\top P\Delta\Sigma + \Delta^\top P\theta^\star\Sigma\Delta^\top P\Delta\Sigma + \Delta^\top P\Delta\Sigma\theta^{\star\top}P\Delta\Sigma\right).
\end{aligned}$$

We lower-bound S and T individually. We see that the first term of S is equal to $\frac{1}{2}\|\Delta^\top P\theta^\star\Sigma\|_F^2$ using the symmetry condition described in Lemma 3, while the second term is equal to $\frac{1}{2}\|P^{1/2}\theta^\star\Sigma\Delta^\top P^{1/2}\|_F^2$. Applying elementary properties of the Frobenius norm, the fact that $\|X\|_P = \|P^{1/2}X\|_F$ for all matrices X, and Lemma 4, we see that

$$\begin{aligned}
S &\geq \frac{1}{2}\sigma_{\min}^2(P^{1/2}\theta^\star\Sigma)\|\Delta\|_P^2 \\
&\geq \frac{1}{2}K_0\|\Delta\|_P^2.
\end{aligned}$$

Applying the Cauchy-Schwarz inequality, Lemma 3, and elementary properties of the Frobenius norm, we see that

$$T \geq -\frac{3}{2}K_1\|\Sigma\|_{\mathrm{op}}\|\Delta\|_P^3 - \|\Sigma\|_{\mathrm{op}}^2\|\Delta\|_P^4.$$

Putting the pieces together, we see that

$$\langle P^{-1}\nabla Q(\theta), \Delta\rangle_P \geq \frac{1}{4}K_0\|\Delta\|_P^2$$

provided that

$$\|\Delta\|_P \leq \min\left\{\frac{K_0}{12K_1\|\Sigma\|_{\mathrm{op}}}, \sqrt{\frac{K_0}{8\|\Sigma\|_{\mathrm{op}}^2}}\right\}.$$

We now establish strong smoothness of $Q(\theta)$ near $\mathcal{S}$. We see that

$$\begin{aligned}
\|P^{-1}\nabla Q(\theta)\|_P^2 &= \frac{1}{4}\|(\theta\Sigma\theta^\top - 2\mathrm{Sym}(M))P\theta\Sigma\|_P^2 \\
&= \frac{1}{4}\|((\theta^\star+\Delta)\Sigma(\theta^\star+\Delta)^\top - 2\mathrm{Sym}(M))P(\theta^\star+\Delta)\Sigma\|_P^2 \\
&\leq \sum_{i=1}^{7} T_i,
\end{aligned}$$

where we define

$$\begin{aligned}
T_1 &= \frac{7}{4}\left\|(\theta^\star\Sigma\theta^{\star\top} - 2\mathrm{Sym}(M))P\,\Delta\,\Sigma\right\|_P^2, \\
T_2 &= \frac{7}{4}\left\|\theta^\star\Sigma\Delta^\top P\,\theta^\star\,\Sigma\right\|_P^2, \\
T_3 &= \frac{7}{4}\left\|\Delta\Sigma\theta^{\star\top} P\,\theta^\star\,\Sigma\right\|_P^2, \\
T_4 &= \frac{7}{4}\left\|\theta^\star\Sigma\Delta^\top P\,\Delta\,\Sigma\right\|_P^2, \\
T_5 &= \frac{7}{4}\left\|\Delta\Sigma\theta^{\star\top} P\,\Delta\,\Sigma\right\|_P^2, \\
T_6 &= \frac{7}{4}\left\|\Delta\Sigma\Delta^\top P\,\theta^\star\,\Sigma\right\|_P^2, \\
T_7 &= \frac{7}{4}\left\|\Delta\Sigma\Delta^\top P\,\Delta\,\Sigma\right\|_P^2,
\end{aligned}$$

and use the easily-verified algebraic fact that

$$(\theta^\star\Sigma\theta^{\star\top} - 2\mathrm{Sym}(M))P\,\theta^\star\Sigma = 0.$$

Recall that for every matrix X, $\|X\|_P = \|P^{1/2}X\|_F$. We recall that

$$\theta^\star\Sigma\theta^{\star\top} - 2\mathrm{Sym}(M) = \tilde{\theta}^\star\Sigma\tilde{\theta}^{\star\top}$$

and observe that $\|P^{1/2}\tilde{\theta}^\star\Sigma\|_{\mathrm{op}} = \|P^{1/2}\theta^\star\Sigma\|_{\mathrm{op}}$. It is clear that $\|\Delta\|_P^4, \|\Delta\|_P^6 \leq \|\Delta\|_P^2$ for all Δ such that $\|\Delta\|_P^2 \leq 1$. Applying Lemma 4 and elementary properties of the Frobenius norm, we see that

$$\|P^{-1}\nabla Q(\theta)\|_P^2 \leq \left(\left(\frac{7}{2} + \frac{7}{4}\kappa^2(\Sigma)\right)K_1^2 + \frac{21}{4}\|\Sigma\|_{\mathrm{op}}^2 K_1 + \frac{7}{4}\|\Sigma\|_{\mathrm{op}}^4\right)\|\Delta\|_P^2$$

for all Δ such that $\|\Delta\|_P \leq 1$. ■

A.4. Lemmas used in the proof of Theorem 2

Lemma 3 *Fix any $\theta = (\tilde{A}, \tilde{B})$ where $\tilde{A} \in \mathbb{R}^{p\times d}$ and $\tilde{B} \in \mathbb{R}^{d\times d}$ and let $\theta^\star$ be the projection of θ onto $\mathcal{S}$ in the P-norm. The matrix $\Delta^\top P\theta^\star\Sigma$ is symmetric.*

Proof Recall from the second part of Theorem 2 that each point on $\mathcal{S}$ has the form

$$\begin{bmatrix} U\Gamma^{1/2}J^\top\Sigma^{-1/2} \\ \Sigma^{-1/2}V\Gamma^{1/2}J^\top\Sigma^{-1/2} \end{bmatrix}$$

for some orthogonal $J \in \mathbb{R}^{d\times d}$, where $U\Gamma V^\top$ is the singular value decomposition of $M\Sigma^{1/2}$. It follows that

$$\theta^\star = S\Gamma^{1/2}J^{\star\top}\Sigma^{-1/2},$$

where we set

$$S = \begin{bmatrix} U \\ \Sigma^{-1/2}V \end{bmatrix}, \qquad J^\star = \operatorname*{argmin}_{J\in\mathbb{O}_d} F(J)$$

and

$$F(J) = \frac{1}{2}\left\|\theta - S\Gamma^{1/2}J^\top\Sigma^{-1/2}\right\|_P^2.$$

We relax the constraint that J is orthogonal and define the Lagrangian

$$G(J,\Lambda) = F(J) + \frac{1}{2}\operatorname{Tr}(\Lambda(J^\top J - I_d)).$$

We may assume that $\Lambda \in \mathbb{R}^{d\times d}$ is symmetric without loss of generality, since the matrix $J^\top J - I_d$ is clearly symmetric and the trace of the product of a symmetric matrix and a skew-symmetric matrix is zero. Any orthogonal minimizer $J^\star$ of $F(J)$ satisfies the first-order condition $\nabla_J G(J,\Lambda) = 0$ for some symmetric Λ. We see that this condition is precisely

$$-\Sigma^{-1/2}\theta^\top PS\Gamma^{1/2} + \Sigma^{-1}J^\star\Gamma^{1/2}(S^\top PS)\Gamma^{1/2} + J^\star\Lambda = 0.$$

Observing that $S^\top PS = 2I_d$ and rearranging, we obtain

$$\Sigma^{-1/2}\theta^\top PS\Gamma^{1/2} = 2\Sigma^{-1}J^\star\Gamma + J^\star\Lambda. \tag{8}$$

Recall that $\Delta = \theta - \theta^\star$ and $\theta^\star = S\Gamma^{1/2}J^{\star\top}\Sigma^{-1/2}$. It follows that

$$\begin{aligned} \Sigma^{-1/2}\theta^\top PS\Gamma^{1/2} &= \Sigma^{-1/2}\Delta^\top PS\Gamma^{1/2} + \Sigma^{-1/2}\theta^{\star\top}PS\Gamma^{1/2} \\ &= \Sigma^{-1/2}\Delta^\top PS\Gamma^{1/2} + 2\Sigma^{-1}J^\star\Gamma, \end{aligned}$$

where we once again used the identity $S^\top PS = 2I_d$. In light of (8), this equation implies that

$$\Sigma^{-1/2}\Delta^\top PS\Gamma^{1/2} = J^\star\Lambda. \tag{9}$$

We use this identity to prove that $\Delta^\top P\theta^\star\Sigma$ is symmetric. Plugging in our expression for $\theta^\star$ once more, we see that

$$\begin{aligned} \Delta^\top P\theta^\star\Sigma &= \Delta^\top PS\Gamma^{1/2}J^{\star\top}\Sigma^{1/2} \\ &= \Sigma^{1/2}J^\star\Lambda J^{\star\top}\Sigma^{1/2}, \end{aligned}$$

where we applied (9) in the second step. This matrix is symmetric because Λ is symmetric. ■

Lemma 4 *Let $\theta^\star = (\tilde{A}^\star, \tilde{B}^\star)$ be any point in $\mathcal{S}$. The following inequalities hold:*

$$K_0 \leq \sigma_{\min}^2(P^{1/2}\theta^\star\Sigma), \qquad \sigma_1^2(P^{1/2}\theta^\star\Sigma) \leq K_1,$$

where we set

$$K_0 = 2\sigma_{\min}(M\Sigma^{1/2})\sigma_{\min}(\Sigma) \qquad K_1 = 2\sigma_1(M\Sigma^{1/2})\sigma_1(\Sigma).$$

In addition, the following identity holds:

$$\Sigma^{1/2}\tilde{A}^{\star\top}\tilde{A}^\star\Sigma^{1/2} = \Sigma^{1/2}\tilde{B}^{\star\top}\Sigma\tilde{B}^\star\Sigma^{1/2} = J\Gamma J^\top.$$

Proof Recall from the second part of Theorem 2 that each point $\theta^\star = (\tilde{A}^\star, \tilde{B}^\star)$ on $\mathcal{S}$ has the form

$$\begin{bmatrix}\tilde{A}^\star \\ \tilde{B}^\star\end{bmatrix} = \begin{bmatrix} U\Gamma^{1/2}J^\top\Sigma^{-1/2} \\ \Sigma^{-1/2}V\Gamma^{1/2}J^\top\Sigma^{-1/2}\end{bmatrix}$$

for some orthogonal $J \in \mathbb{R}^{d\times d}$, where $U\Gamma V^\top$ is the singular value decomposition of $M\Sigma^{1/2}$. We see that

$$P^{1/2}\theta^\star\Sigma = \begin{bmatrix}U \\ V\end{bmatrix}\Gamma^{1/2}J^\top\Sigma^{1/2}.$$

Using the fact that

$$U^\top U + V^\top V = 2I_d$$

and the fact that J is orthogonal, and applying elementary properties of singular values, we obtain the bound

$$2\sigma_d(M\Sigma^{1/2})\sigma_d(\Sigma) \leq \sigma_d^2(P^{1/2}\theta^\star\Sigma).$$

The upper bound

$$\sigma_1^2(P^{1/2}\theta^\star\Sigma) \leq 2\sigma_1(M\Sigma^{1/2})\sigma_1(\Sigma)$$

follows from a similar calculation. We observe via direct calculation that

$$\Sigma^{1/2}\tilde{A}^{\star\top}\tilde{A}^\star\Sigma^{1/2} = \Sigma^{1/2}\tilde{B}^{\star\top}\Sigma\tilde{B}^\star\Sigma^{1/2} = J\Gamma J^\top.$$

■

A.5. Approximate strong convexity and strong smoothness of the empirical loss

The following lemma shows that the strong convexity and smoothness established in Theorem 2 for the population loss holds approximately on the empirical loss:

Lemma 5 *Let α, β be as described in Theorem 2. Using the notation defined in the proof of Theorem 1, the following inequalities hold for each $t \in [m]$:*

$$\langle \hat{Z}_t, \Delta_t\rangle_P \geq \frac{\alpha}{2}\|\Delta_t\|_P^2 - \alpha^{-1}\|\xi_t\|_P^2 - \alpha^{-1}\|\zeta_t\|_P^2 \tag{10}$$

$$\|\hat{Z}_t\|_P^2 \leq 3\beta\|\Delta_t\|_P^2 + 3\|\xi_t\|_P^2 + 3\|\zeta_t\|_P^2, \tag{11}$$

where we define the random variables

$$\xi_t = \hat{Z}_t - \hat{P}^{-1}\nabla Q^{\text{emp}}(\theta_t), \qquad \zeta_t = \hat{P}^{-1}\nabla Q^{\text{emp}}(\theta_t) - P^{-1}\nabla Q(\theta_t).$$

Proof Let Q^{emp} be defined as

$$Q^{\text{emp}}(\tilde{A}, \tilde{B}) = \frac{1}{2}\|\tilde{A}\Sigma\tilde{B}^\top\hat{\Sigma}^{1/2} - M\hat{\Sigma}^{1/2}\|_F^2 + \frac{1}{8}\|\hat{\Sigma}^{1/2}(\tilde{A}^\top\tilde{A} - \tilde{B}^\top\hat{\Sigma}\tilde{B})\hat{\Sigma}^{1/2}\|_F^2.$$

We emphasize that the covariance sandwiched between A and B is the true covariance, not the sample covariance. Define the error residuals

$$\xi_t = \hat{Z}_t - \hat{P}^{-1}\nabla Q^{\text{emp}}(\theta_t), \qquad \zeta_t = \hat{P}^{-1}\nabla Q^{\text{emp}}(\theta_t) - P^{-1}\nabla Q(\theta_t).$$

We first establish the approximate one-point strong convexity property (10). We observe that

$$\begin{aligned}
\langle \hat{Z}_t, \Delta_t\rangle_P &= \left\langle \xi_t + \zeta_t + P^{-1}\nabla Q(\theta_t),\, \Delta_t\right\rangle_P \\
&= \langle \xi_t + \zeta_t,\, \Delta_t\rangle_P + \left\langle P^{-1}\nabla Q(\theta_t),\, \Delta_t\right\rangle_P \\
&\geq \langle \xi_t + \zeta_t,\, \Delta_t\rangle_P + \alpha\|\Delta\|_P^2,
\end{aligned}$$

where we applied the one-point strong convexity property established in Theorem 2. We now lower-bound

$$\langle \xi_t + \zeta_t,\, \Delta_t\rangle_P\,.$$

Applying the Cauchy-Schwarz inequality, we see that

$$\langle \xi_t, \Delta_t\rangle_P \geq -\|\xi_t\|_P\|\Delta_t\|_P.$$

We recall Young's inequality: when $u, v \geq 0$ then $uv \leq \frac{1}{2\nu}u^2 + \frac{\nu}{2}v^2$ for any $\nu > 0$. Applying this inequality with $\nu = \frac{\alpha}{2}$, we see that

$$\langle \xi_t, \Delta_t\rangle \geq -\alpha^{-1}\|\xi_t\|_P^2 - \frac{\alpha}{4}\|\Delta_t\|_P^2.$$

An identical calculation shows that

$$\langle \zeta_t, \Delta_t\rangle \geq -\alpha^{-1}\|\zeta_t\|_P^2 - \frac{\alpha}{4}\|\Delta_t\|_P^2.$$

This proves the approximate one-point strong convexity property.

We now establish the approximate one-point smoothness property (11). We observe that

$$\begin{aligned}
\|\hat{Z}_t\|_P^2 &= \left\|P^{-1}\nabla Q(\theta_t) + \xi_t + \zeta_t\right\|_P^2 \\
&\leq 3\left\|P^{-1}\nabla Q(\theta_t)\right\|_P^2 + 3\|\xi_t\|_P^2 + 3\|\zeta_t\|_P^2 \\
&\leq 3\beta\,\|\Delta_t\|_P^2 + 3\|\xi_t\|_P^2 + 3\|\zeta_t\|_P^2,
\end{aligned}$$

where we applied the one-point strong smoothness property established in Theorem 2. ■

Lemma 6 (Descent Lemma) *Let β, ε_0 and $\mathcal{S}$ be defined as in Theorem 2. Suppose that θ is ε_0-close to $\mathcal{S}$ in the P-norm. Let $\theta^\star$ denote the projection of θ onto $\mathcal{S}$ in the P-norm and let $\Delta = \theta - \theta^\star$. The following inequality holds:*

$$Q(\theta) - Q^\star \leq \frac{\sqrt{\beta}}{2}\|\Delta\|_P^2.$$

Proof For all $t \in [0,1]$, let $s(t) = \theta^\star + t(\theta - \theta^\star)$. Notice that

$$\begin{aligned}
Q(\theta) - Q^\star &= Q(\theta) - Q(\theta^\star) \\
&= Q(s(1)) - Q(s(0)) \\
&= \int_0^1 \frac{d}{dt} Q(s(t))\, dt \\
&= \int_0^1 \langle \nabla Q(s(t)), \theta - \theta^\star \rangle\, dt \\
&= \int_0^1 \langle P^{-1} \nabla Q(s(t)), \theta - \theta^\star \rangle_P\, dt \\
&\leq \int_0^1 \|P^{-1} \nabla Q(s(t))\|_P \|\theta - \theta^\star\|_P\, dt \\
&\leq \int_0^1 \sqrt{\beta} t \|\theta - \theta^\star\|_P^2\, dt \\
&= \frac{\sqrt{\beta}}{2} \|\Delta\|_P^2,
\end{aligned}$$

where we applied the Cauchy-Schwarz inequality and the one-point smoothness property described in Theorem 2. ■

Appendix B. Concentration

Recall that our analysis is based on controlling the pairwise distances between the gradients of the regularized loss functions $\hat{Q}(\theta), Q^{\text{emp}}(\theta)$, and $Q(\theta)$. In particular, we wish to argue that gradient descent on $\hat{Q}$ behaves almost identically to gradient descent on $Q(\theta)$ when n is large. These functions are defined as follows:

$$\begin{aligned}
\hat{Q}(\theta) &= \hat{L}(\theta) + \hat{R}(\theta), \\
Q^{\text{emp}} &= L^{\text{emp}}(\theta) + \hat{R}(\theta), \\
Q(\theta) &= L(\theta) + R(\theta),
\end{aligned}$$

where we define

$$\begin{aligned}
L(\theta) &= \frac{1}{2} \|\tilde{A} \Sigma \tilde{B}^\top \Sigma^{1/2} - M \Sigma^{1/2}\|_F^2, \\
L^{\text{emp}}(\theta) &= \frac{1}{2} \|\tilde{A} \Sigma \tilde{B}^\top \hat{\Sigma}^{1/2} - M \hat{\Sigma}^{1/2}\|_F^2, \\
\hat{L}(\theta) &= \frac{1}{2n} \sum_{i=1}^n \left\| A \left(\frac{\sum_{j=1}^n \exp(x_i^\top B x_j) x_j}{\sum_{j=1}^n \exp(x_i^\top B x_j)} \right) - y_i \right\|_2^2, \\
R(\theta) &= \frac{1}{8} \|\Sigma^{1/2} (\tilde{A}^\top \tilde{A} - \tilde{B}^\top \Sigma \tilde{B}) \Sigma^{1/2}\|_F^2, \\
\hat{R}(\theta) &= \frac{1}{8} \|\hat{\Sigma}^{1/2} (\tilde{A}^\top \tilde{A} - \tilde{B}^\top \hat{\Sigma} \tilde{B}) \hat{\Sigma}^{1/2}\|_F^2,
\end{aligned}$$

and recall that

$$A = \hat{\gamma}^{-1}\|\tilde{B}\|_F \tilde{A}, \qquad B = \hat{\gamma}\|\tilde{B}\|_F^{-1}\tilde{B}.$$

We emphasize that only a single copy of Σ is replaced by $\hat{\Sigma}$ in the definition of $L^{\text{emp}}(\theta)$.

The proof of Theorem 1 requires that the variables are uniformly bounded in at all points of the gradient descent trajectory. Specifically, we prove:

Theorem 7 (Uniform concentration along gradient descent trajectory) *With probability at least $1-\delta/2$, the following inequality holds:*

$$\sup_{t\in[m]} \max\left\{\|\xi_{t-1}\|_P^2,\ \|\zeta_{t-1}\|_P^2,\right\} \leq \frac{K\log^5(n)\log(mn/\delta)}{n}$$

for some constant $K \geq 0$.

The rest of this section is used to prove Theorem 7.

B.1. Proof of Theorem 7

We first show that

$$\sup_{t\in[m]} \|\xi_{t-1}\|_P^2 \leq \frac{K\log^5(n)\log^2(mn/\delta)}{n}$$

with probability at least $1-\delta/2$ (Lemma 11), and then show that

$$\sup_{t\in[m]} \|\zeta_{t-1}\|_P^2 \leq \frac{K\log^5(n)\log(mn/\delta)}{n}$$

with probability at least $1-\delta/2$ (Lemma 12); the union bound implies the claim.

B.1.1. Uniform control of $\|\xi_t\|_P^2$

Recall that

$$L^{\text{emp}}(\tilde{A},\tilde{B}) = \frac{1}{2}\left\|(\tilde{A}\Sigma\tilde{B}^\top - M)\hat{\Sigma}^{1/2}\right\|_F^2.$$

Set

$$R = \max\left(1, \sqrt{\operatorname{Tr}(\Sigma)} + 2\sqrt{2\|\Sigma\|_{\text{op}}\log n}\right), \qquad q = \max\left(2, \log\frac{48mn}{\delta}\right).$$

Lemma 8 *Fix $\theta = (\tilde{A},\tilde{B}) \in \bar{\mathcal{S}}$, and let*

$$A = \hat{\gamma}^{-1}\|\tilde{B}\|_F \tilde{A}, \qquad B = \hat{\gamma}\|\tilde{B}\|_F^{-1}\tilde{B}.$$

Set

$$G_A = \nabla_A \hat{L}(A,B) - \nabla_A L^{\text{emp}}(A,B), \qquad G_B = \nabla_B \hat{L}(A,B) - \nabla_B L^{\text{emp}}(A,B).$$

On the event that the eigenvalues of $\hat{\Sigma}$ are bounded above and below by positive constants,

$$\|\xi(\theta)\|_P^2 \leq K\left(\hat{\gamma}^{-2}\|G_A\|_F^2 + \hat{\gamma}^2\|G_B\|_F^2\right).$$

Proof Since L^{emp} is invariant under the rescaling from $(\tilde{A}, \tilde{B})$ to (A, B), the chain rule gives

$$\nabla_{\tilde{A}} \tilde{L}(\tilde{A}, \tilde{B}) - \nabla_{\tilde{A}} L^{\mathrm{emp}}(\tilde{A}, \tilde{B}) = \frac{\|\tilde{B}\|_F}{\hat{\gamma}} G_A$$

and

$$\nabla_{\tilde{B}} \tilde{L}(\tilde{A}, \tilde{B}) - \nabla_{\tilde{B}} L^{\mathrm{emp}}(\tilde{A}, \tilde{B}) = \frac{\hat{\gamma}}{\|\tilde{B}\|_F} G_B + \frac{\mathrm{Tr}(G_A^\top A) - \mathrm{Tr}(G_B^\top B)}{\|\tilde{B}\|_F^2} \tilde{B}.$$

Shrinking ε_0 if necessary, compactness of $\mathcal{S}$ implies that

$$K^{-1} \le \|\tilde{B}\|_F \le K, \qquad \|\tilde{A}\|_F \le K$$

uniformly over $\theta \in \bar{\mathcal{S}}$. Hence

$$\|A\|_F \le K\hat{\gamma}^{-1}, \qquad \|B\|_F = \hat{\gamma}.$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned} \left|\mathrm{Tr}(G_A^\top A) - \mathrm{Tr}(G_B^\top B)\right|^2 &\le 2\|G_A\|_F^2 \|A\|_F^2 + 2\|G_B\|_F^2 \|B\|_F^2 \\ &\le K\left(\hat{\gamma}^{-2}\|G_A\|_F^2 + \hat{\gamma}^2 \|G_B\|_F^2\right). \end{aligned}$$

It follows that

$$\left\|\nabla \tilde{L}(\tilde{A}, \tilde{B}) - \nabla L^{\mathrm{emp}}(\tilde{A}, \tilde{B})\right\|_F^2 \le K\left(\hat{\gamma}^{-2}\|G_A\|_F^2 + \hat{\gamma}^2 \|G_B\|_F^2\right).$$

Finally,

$$\xi(\theta) = \hat{P}^{-1}\left(\nabla \tilde{L}(\tilde{A}, \tilde{B}) - \nabla L^{\mathrm{emp}}(\tilde{A}, \tilde{B})\right),$$

and the assumed bounds on the eigenvalues of $\hat{\Sigma}$ imply

$$\|\hat{P}^{-1} Z\|_P^2 \le K\|Z\|_F^2.$$

The claim follows. ■

Lemma 9 *Fix $\theta = (\tilde{A}, \tilde{B}) \in \bar{\mathcal{S}}$, and let*

$$A = \hat{\gamma}^{-1}\|\tilde{B}\|_F \tilde{A}, \qquad B = \hat{\gamma}\|\tilde{B}\|_F^{-1} \tilde{B}.$$

Set

$$\bar{\mu}_i = \mathbb{E}_{-i}[\mu_i], \qquad \bar{\Sigma}_i = \mathbb{E}_{-i}[\Sigma_i],$$

and define

$$G_{1,A} = \frac{1}{n}\sum_{i=1}^n A\bar{\mu}_i \bar{\mu}_i^\top - \frac{1}{n}\sum_{i=1}^n M x_i \bar{\mu}_i^\top$$

and

$$G_{1,B} = \frac{1}{n}\sum_{i=1}^{n} x_i \bar{\mu}_i^\top A^\top A \bar{\Sigma}_i - \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top M^\top A \bar{\Sigma}_i.$$

On the event that $\hat{\gamma} \le \gamma$ *and* $\hat{\gamma}^{-2} \le KR^2 q$*, with probability at least* $1 - \delta/(8m)$*,*

$$\begin{aligned}\hat{\gamma}^{-2} &\left\| \nabla_A \hat{L}(A,B) - G_{1,A} \right\|_F^2 \\ &+ \hat{\gamma}^2 \left\| \nabla_B \hat{L}(A,B) - G_{1,B} \right\|_F^2 \le \frac{KR^{10}q^2}{n}\end{aligned}$$

for sufficiently large n*.*

Proof Recall that

$$\nabla_A \hat{L}(A,B) = \frac{1}{n}\sum_{i=1}^{n}(A\mu_i - Mx_i)\mu_i^\top - \frac{1}{n}\sum_{i=1}^{n} z_i \mu_i^\top$$

and

$$\begin{aligned}\nabla_B \hat{L}(A,B) = \frac{1}{n}\sum_{i=1}^{n} x_i \mu_i^\top A^\top A \Sigma_i - \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top M^\top A \Sigma_i \\ - \frac{1}{n}\sum_{i=1}^{n} x_i z_i^\top A \Sigma_i.\end{aligned}$$

Set

$$N_A = -\frac{1}{n}\sum_{i=1}^{n} z_i \mu_i^\top, \qquad N_B = -\frac{1}{n}\sum_{i=1}^{n} x_i z_i^\top A \Sigma_i.$$

Applying Jensen's inequality,

$$\left\| \nabla_A \hat{L}(A,B) - G_{1,A} \right\|_F^2 \le 2(D_1 + D_2) + 2\|N_A\|_F^2$$

and

$$\left\| \nabla_B \hat{L}(A,B) - G_{1,B} \right\|_F^2 \le 2(D_3 + D_4) + 2\|N_B\|_F^2,$$

where

$$\begin{aligned}D_1 &= \frac{2}{n}\sum_{i=1}^{n} \left\| A\mu_i\mu_i^\top - A\bar{\mu}_i\bar{\mu}_i^\top \right\|_F^2, \\ D_2 &= \frac{2}{n}\sum_{i=1}^{n} \left\| Mx_i\mu_i^\top - Mx_i\bar{\mu}_i^\top \right\|_F^2, \\ D_3 &= \frac{2}{n}\sum_{i=1}^{n} \left\| x_i\mu_i^\top A^\top A\Sigma_i - x_i\bar{\mu}_i^\top A^\top A\bar{\Sigma}_i \right\|_F^2, \\ D_4 &= \frac{2}{n}\sum_{i=1}^{n} \left\| x_i x_i^\top M^\top A(\Sigma_i - \bar{\Sigma}_i) \right\|_F^2.\end{aligned}$$

By the conditional concentration bounds for μ_i and Σ_i, the conditional bias bounds, and the Gaussian norm bound, with probability at least $1-\delta/(16m)$,

$$\|x_i\|_2 \le R, \qquad \|\mu_i - \bar{\mu}_i\|_2^2 \le \frac{KR^4 q}{n}, \qquad \|\Sigma_i - \bar{\Sigma}_i\|_F^2 \le \frac{KR^6 q}{n}$$

for every $i \in [n]$, and

$$\|\bar{\mu}_i - \Sigma B^\top x_i\|_2^2 \le \frac{KR^2}{n^2}, \qquad \|\bar{\Sigma}_i - \Sigma\|_F^2 \le \frac{KR^8}{n^2}.$$

Since $\|B\|_F = \hat{\gamma}$, it follows that

$$\|\bar{\mu}_i\|_2^2 \le K\left(\hat{\gamma}^2 R^2 + \frac{R^2}{n^2}\right), \qquad \|\bar{\Sigma}_i\|_F^2 \le K.$$

Moreover, since $\theta \in \bar{\mathcal{S}}$,

$$\|A\|_F \le K\hat{\gamma}^{-1}.$$

We first bound D_1. Since

$$\begin{aligned}\mu_i\mu_i^\top - \bar{\mu}_i\bar{\mu}_i^\top &= (\mu_i - \bar{\mu}_i)(\mu_i - \bar{\mu}_i)^\top \\ &\quad + (\mu_i - \bar{\mu}_i)\bar{\mu}_i^\top + \bar{\mu}_i(\mu_i - \bar{\mu}_i)^\top,\end{aligned}$$

we obtain

$$D_1 \le K\hat{\gamma}^{-2}\left[\frac{R^8 q^2}{n^2} + \frac{R^4 q}{n}\left(\hat{\gamma}^2 R^2 + \frac{R^2}{n^2}\right)\right].$$

Similarly,

$$D_2 \le \frac{KR^6 q}{n}.$$

For D_3, we use

$$\begin{aligned}\mu_i^\top A^\top A\Sigma_i - \bar{\mu}_i^\top A^\top A\bar{\Sigma}_i &= (\mu_i - \bar{\mu}_i)^\top A^\top A(\Sigma_i - \bar{\Sigma}_i) \\ &\quad + (\mu_i - \bar{\mu}_i)^\top A^\top A\bar{\Sigma}_i \\ &\quad + \bar{\mu}_i^\top A^\top A(\Sigma_i - \bar{\Sigma}_i).\end{aligned}$$

It follows that

$$\begin{aligned}D_3 \le K\hat{\gamma}^{-4}R^2\Bigg[&\frac{R^{10}q^2}{n^2} + \frac{R^4 q}{n} \\ &+ \left(\hat{\gamma}^2 R^2 + \frac{R^2}{n^2}\right)\frac{R^6 q}{n}\Bigg].\end{aligned}$$

Finally,

$$D_4 \le \frac{K\hat{\gamma}^{-2}R^{10}q}{n}.$$

Using $\hat{\gamma}^{-2} \le KR^2 q$, the preceding four displays imply, for sufficiently large n,

$$\hat{\gamma}^{-2}(D_1 + D_2) + \hat{\gamma}^2(D_3 + D_4) \le \frac{KR^{10}q^2}{n}.$$

It remains to bound the noise terms. Conditional on the covariates, N_A and N_B are Gaussian random matrices. On the event $\max_i \|x_i\|_2 \le R$,

$$\|\mu_i\|_2 \le R, \qquad \|\Sigma_i\|_{\mathrm{op}} \le 4R^2.$$

Consequently, Gaussian concentration gives, with probability at least $1 - \delta/(16m)$,

$$\|N_A\|_F^2 \le \frac{KR^2 q}{n}, \qquad \|N_B\|_F^2 \le \frac{K\hat{\gamma}^{-2} R^6 q}{n}.$$

Therefore,

$$\hat{\gamma}^{-2}\|N_A\|_F^2 + \hat{\gamma}^2\|N_B\|_F^2 \le \frac{KR^{10}q^2}{n}.$$

Combining the preceding bounds proves the claim. ■

Lemma 10 *Fix $\theta = (\tilde{A}, \tilde{B}) \in \bar{\mathcal{S}}$, and let A, B, $G_{1,A}$, and $G_{1,B}$ be defined as above. On the event that $\hat{\gamma} \le \gamma$ and $\hat{\gamma}^{-2} \le KR^2 q$, with probability at least $1 - \delta/(8m)$,*

$$\begin{aligned}\hat{\gamma}^{-2} &\|G_{1,A} - \nabla_A L^{\mathrm{emp}}(A,B)\|_F^2 \\ &+ \hat{\gamma}^2 \|G_{1,B} - \nabla_B L^{\mathrm{emp}}(A,B)\|_F^2 \le \frac{KR^{10}q^2}{n}\end{aligned}$$

for sufficiently large n.

Proof Recall that

$$\nabla_A L^{\mathrm{emp}}(A,B) = (A\Sigma B^\top - M)\hat{\Sigma} B \Sigma$$

and

$$\nabla_B L^{\mathrm{emp}}(A,B) = \hat{\Sigma}(A\Sigma B^\top - M)^\top A\Sigma.$$

Applying Jensen's inequality,

$$\|G_{1,A} - \nabla_A L^{\mathrm{emp}}(A,B)\|_F^2 \le E_1 + E_2$$

and

$$\|G_{1,B} - \nabla_B L^{\mathrm{emp}}(A,B)\|_F^2 \le E_3 + E_4,$$

where

$$\begin{aligned}E_1 &= \frac{2}{n}\|A\|_F^2 \sum_{i=1}^n \left\|\bar{\mu}_i\bar{\mu}_i^\top - \Sigma B^\top x_i x_i^\top B\Sigma\right\|_F^2, \\ E_2 &= \frac{2}{n}\|M\|_F^2 \sum_{i=1}^n \|x_i\|_2^2 \left\|\bar{\mu}_i - \Sigma B^\top x_i\right\|_2^2, \\ E_3 &= \frac{2}{n}\sum_{i=1}^n \|x_i\|_2^2 \left\|\bar{\mu}_i^\top A^\top A\bar{\Sigma}_i - x_i^\top B\Sigma A^\top A\Sigma\right\|_2^2, \\ E_4 &= \frac{2}{n}\|M\|_F^2\|A\|_F^2 \sum_{i=1}^n \|x_i\|_2^4 \left\|\bar{\Sigma}_i - \Sigma\right\|_F^2.\end{aligned}$$

On the event $\max_i \|x_i\|_2 \le R$, the conditional bias bounds give

$$\|\bar{\mu}_i - \Sigma B^\top x_i\|_2^2 \le \frac{KR^2}{n^2}, \qquad \|\bar{\Sigma}_i - \Sigma\|_F^2 \le \frac{KR^8}{n^2}.$$

For E_1, we write

$$\begin{aligned}\bar{\mu}_i\bar{\mu}_i^\top - \Sigma B^\top x_i x_i^\top B\Sigma = {} & (\bar{\mu}_i - \Sigma B^\top x_i)(\bar{\mu}_i - \Sigma B^\top x_i)^\top \\ & + (\bar{\mu}_i - \Sigma B^\top x_i)x_i^\top B\Sigma \\ & + \Sigma B^\top x_i(\bar{\mu}_i - \Sigma B^\top x_i)^\top.\end{aligned}$$

Using $\|A\|_F \le K\hat{\gamma}^{-1}$ and $\|B\|_F = \hat{\gamma}$ gives

$$E_1 \le K\hat{\gamma}^{-2}\left(\frac{R^4}{n^4} + \frac{\hat{\gamma}^2 R^4}{n^2}\right).$$

Also,

$$E_2 \le \frac{KR^4}{n^2}.$$

For E_3, we write

$$\begin{aligned}&\bar{\mu}_i^\top A^\top A\bar{\Sigma}_i - x_i^\top B\Sigma A^\top A\Sigma \\ &\quad = (\bar{\mu}_i - \Sigma B^\top x_i)^\top A^\top A(\bar{\Sigma}_i - \Sigma) \\ &\qquad + (\bar{\mu}_i - \Sigma B^\top x_i)^\top A^\top A\Sigma \\ &\qquad + x_i^\top B\Sigma A^\top A(\bar{\Sigma}_i - \Sigma).\end{aligned}$$

Hence

$$E_3 \le K\hat{\gamma}^{-4}R^2\left(\frac{R^{10}}{n^4} + \frac{R^2}{n^2} + \frac{\hat{\gamma}^2 R^{10}}{n^2}\right).$$

Finally,

$$E_4 \le \frac{K\hat{\gamma}^{-2}R^{12}}{n^2}.$$

Using $\hat{\gamma}^{-2} \le KR^2 q$, the preceding bounds imply, for sufficiently large n,

$$\hat{\gamma}^{-2}(E_1 + E_2) + \hat{\gamma}^2(E_3 + E_4) \le \frac{KR^{10}q^2}{n}.$$

The claim follows. ■

Lemma 11 (Uniform control of the finite-softmax approximation error) *Suppose $\theta_0, \dots, \theta_{m-1} \in \bar{\mathcal{S}}$. There exists a constant $K > 0$, depending only on the problem parameters, such that with probability at least $1 - \delta/2$,*

$$\sup_{t\in[m]} \|\xi_{t-1}\|_P^2 \le \frac{KR^{10}q^2}{n}.$$

In particular,

$$\sup_{t\in[m]} \|\xi_{t-1}\|_P^2 \le \frac{K\log^5(n)\log^2(48mn/\delta)}{n}.$$

Proof By covariance concentration and the definition of $\hat{\gamma}$, with probability at least $1-\delta/4$,

$$\lambda_{\min}(\hat{\Sigma}) \geq K^{-1}, \qquad \|\hat{\Sigma}\|_{\mathrm{op}} \leq K, \qquad \hat{\gamma} \leq \gamma.$$

Moreover,

$$\hat{\gamma}^{-2} = \frac{4p\hat{\nu}^2}{\|\hat{\Sigma}\|_{\mathrm{op}}} \leq KR^2 q,$$

where p is defined in Algorithm 1 and $p \leq q$.

Fix $t \in [m]$. Lemmas 9 and 10, together with

$$\|X+Y\|_F^2 \leq 2\|X\|_F^2 + 2\|Y\|_F^2,$$

give

$$\hat{\gamma}^{-2}\|G_A\|_F^2 + \hat{\gamma}^2\|G_B\|_F^2 \leq \frac{KR^{10}q^2}{n}$$

except on an event of probability at most $\delta/(4m)$. Lemma 8 therefore gives

$$\|\xi_{t-1}\|_P^2 \leq \frac{KR^{10}q^2}{n}.$$

Taking a union bound over $t \in [m]$ proves

$$\sup_{t\in[m]} \|\xi_{t-1}\|_P^2 \leq \frac{KR^{10}q^2}{n}$$

with probability at least $1-\delta/2$. Finally, $R^{10} \leq K\log^5 n$ for sufficiently large n, which proves the second claim. ■

B.1.2. Uniform control of $\|\zeta_t\|_P^2$

Lemma 12 *Suppose $\bar{\mathcal{S}}$ is the ε_0-neighborhood of $\mathcal{S}$, where ε_0 is defined as in Theorem 2. There exists a constant $K > 0$ such that with probability at least $1-\delta/2$,*

$$\sup_{\theta\in\bar{\mathcal{S}}} \|\zeta(\theta)\|_P^2 \leq \frac{K\log(2/\delta)}{n}.$$

Consequently, if $\theta_0, \ldots, \theta_{m-1} \in \bar{\mathcal{S}}$, then

$$\sup_{t\in[m]} \|\zeta_{t-1}\|_P^2 \leq \frac{K\log(2/\delta)}{n}.$$

Proof Fix $\theta = (\tilde{A}, \tilde{B}) \in \bar{\mathcal{S}}$ and let

$$\Delta = \hat{\Sigma} - \Sigma.$$

By the definition of L^{emp},

$$\hat{P}^{-1}\nabla L^{\mathrm{emp}}(\theta) = \begin{bmatrix} (\tilde{A}\Sigma\tilde{B}^\top - M)\hat{\Sigma}\tilde{B}\Sigma \\ (\tilde{A}\Sigma\tilde{B}^\top - M)^\top\tilde{A}\Sigma \end{bmatrix},$$

whereas

$$P^{-1}\nabla L(\theta) = \begin{bmatrix} (\tilde{A}\Sigma\tilde{B}^\top - M)\Sigma\tilde{B}\Sigma \\ (\tilde{A}\Sigma\tilde{B}^\top - M)^\top\tilde{A}\Sigma \end{bmatrix}.$$

Hence,

$$\hat{P}^{-1}\nabla L^{\mathrm{emp}}(\theta) - P^{-1}\nabla L(\theta) = \begin{bmatrix} (\tilde{A}\Sigma\tilde{B}^\top - M)\Delta\tilde{B}\Sigma \\ 0 \end{bmatrix}.$$

Similarly,

$$\hat{P}^{-1}\nabla \hat{R}(\theta) - P^{-1}\nabla R(\theta) = \frac{1}{2}\begin{bmatrix} \tilde{A}H \\ -\tilde{B}H \end{bmatrix},$$

where

$$H = \hat{\Sigma}\left(\tilde{A}^\top\tilde{A} - \tilde{B}^\top\hat{\Sigma}\tilde{B}\right)\hat{\Sigma} - \Sigma\left(\tilde{A}^\top\tilde{A} - \tilde{B}^\top\Sigma\tilde{B}\right)\Sigma.$$

Using $\hat{\Sigma} = \Sigma + \Delta$, we can rewrite

$$H = \Delta\left(\tilde{A}^\top\tilde{A} - \tilde{B}^\top\Sigma\tilde{B}\right)\Sigma + \hat{\Sigma}\left(\tilde{A}^\top\tilde{A} - \tilde{B}^\top\Sigma\tilde{B}\right)\Delta - \hat{\Sigma}\tilde{B}^\top\Delta\tilde{B}\hat{\Sigma}.$$

Therefore,

$$\zeta(\theta) = \begin{bmatrix} (\tilde{A}\Sigma\tilde{B}^\top - M)\Delta\tilde{B}\Sigma + \frac{1}{2}\tilde{A}H \\ -\frac{1}{2}\tilde{B}H \end{bmatrix}.$$

Since $\bar{\mathcal{S}}$ is bounded,

$$\sup_{\theta\in\bar{\mathcal{S}}}\left(\|\tilde{A}\|_F + \|\tilde{B}\|_F + \|\tilde{A}\Sigma\tilde{B}^\top - M\|_F + \|\tilde{A}^\top\tilde{A} - \tilde{B}^\top\Sigma\tilde{B}\|_F\right) \le K.$$

On the event $\|\Delta\|_{\mathrm{op}} \le 1$, we also have $\|\hat{\Sigma}\|_{\mathrm{op}} \le K$. The preceding expression for H therefore gives, uniformly over $\theta \in \bar{\mathcal{S}}$,

$$\|H\|_F \le K\|\Delta\|_F.$$

It follows from the expression for $\zeta(\theta)$ and equivalence of the P-norm and Frobenius norm that

$$\sup_{\theta\in\bar{\mathcal{S}}}\|\zeta(\theta)\|_P^2 \le K\|\hat{\Sigma} - \Sigma\|_F^2.$$

Finally, Gaussian covariance concentration implies that with probability at least $1 - \delta/2$,

$$\|\hat{\Sigma} - \Sigma\|_F^2 \le \frac{K\log(2/\delta)}{n}.$$

Combining the preceding two displays yields

$$\sup_{\theta\in\bar{\mathcal{S}}}\|\zeta(\theta)\|_P^2 \le \frac{K\log(2/\delta)}{n}.$$

The claimed bound on $\sup_{t\in[m]}\|\zeta_{t-1}\|_P^2$ follows immediately whenever $\theta_0, \ldots, \theta_{m-1} \in \bar{\mathcal{S}}$. ■

Appendix C. Lemmas used in the proof of Theorem 7

C.1. Conditional moments and conditional variance of softmax weights

We use the following classical inequality to establish concentration of the softmax-weighted mean and softmax-weighted covariance.

Theorem 13 (Pisier's inequality) *Let f be a smooth, symmetric function of $x_1, \dots x_n$, where each x_i is sampled i.i.d. from $\mathcal{N}(0, \Sigma)$. The following moment bound holds:*

$$\mathbb{E}\big\|f(x_1, \dots x_n) - \mathbb{E}[f(x_1, \dots x_n)]\big\|_F^p \leq C_p (n-1)^{p/2} \|\Sigma\|_{\text{op}}^{p/2} \mathbb{E}\big\|\nabla_{x_1} f(x_1, \dots x_n)\big\|_F^p$$

Lemma 14 (Some useful gradients) *The following identities hold:*

$$\nabla_{x_2} \pi_{12} = \pi_{12}(1 - \pi_{12}) B^\top x_1,$$

$$\nabla_{x_2} \pi_{11} = -\pi_{11}\pi_{12} B^\top x_1,$$

$$\nabla_{x_2} \mu_1 = \pi_{12}(x_2 - \mu_1) x_1^\top B + \pi_{12} I,$$

$$\nabla_{x_2} \Sigma_1 = \pi_{12}\Big(\Big((x_2 - \mu_1)(x_2 - \mu_1)^\top - \Sigma_1\Big) \otimes (B^\top x_1)\Big) \; + \; \pi_{12}\Big(I \otimes (x_2 - \mu_1) \; + \; (x_2 - \mu_1) \otimes I\Big).$$

Proof These identities are easily verified via direct calculation. ■

Lemma 15 (Conditional moments of softmax weights) *Fix any $p \geq 1$. The following inequalities hold:*

$$\begin{aligned}
\mathbb{E}_{-1}[\pi_{12}^p] &\leq \frac{1}{(n-1)^p} \exp\left(\frac{p^2}{2} x_1^\top B \Sigma B^\top x_1\right) \\
\mathbb{E}_{-1}[\pi_{11}^p] &\leq \frac{1}{(n-1)^p} \exp\left(p\, x_1^\top B x_1\right) \exp\left(\frac{p^2}{2(n-1)} x_1^\top B \Sigma B^\top x_1\right).
\end{aligned}$$

Proof To prove the first inequality, we observe that

$$\begin{aligned}
\pi_{12} &= \frac{\exp(x_1^\top B x_2)}{\sum_{j=1}^n \exp(x_1^\top B x_j)} \\
&\leq \frac{\exp(x_1^\top B x_2)}{\sum_{j=2}^n \exp(x_1^\top B x_j)} \\
&\leq \frac{1}{n-1} \exp\left(\frac{n-2}{n-1} x_1^\top B x_2\right) \exp\left(-\frac{1}{n-1} x_1^\top B \sum_{j=3}^n x_j\right),
\end{aligned}$$

where we applied the AM-GM inequality in the last step. It follows that

$$\begin{aligned}
\mathbb{E}_{-1}[\pi_{12}^p] &= \frac{1}{(n-1)^p} \mathbb{E}_{-1}\left[\exp\left(\frac{p(n-2)}{n-1} x_1^\top B x_2\right) \exp\left(-\frac{p}{n-1} x_1^\top B \sum_{j=3}^n x_j\right)\right] \\
&= \frac{1}{(n-1)^p} \mathbb{E}_2\left[\exp\left(\frac{p(n-2)}{n-1} x_1^\top B x_2\right)\right] \mathbb{E}_{-1,-2}\left[\exp\left(-\frac{p}{n-1} x_1^\top B \sum_{j=3}^n x_j\right)\right],
\end{aligned}$$

where we used the fact that x_2 is independent of $\{x_j\}_{j=3}^n$. We observe that

$$\mathbb{E}_2\left[\exp\left(\frac{p(n-2)}{n-1}x_1^\top Bx_2\right)\right] = \exp\left(\frac{p^2(n-2)^2}{2(n-1)^2}x_1^\top B\Sigma B^\top x_1\right)$$

and

$$\mathbb{E}_{-1,-2}\left[\exp\left(-\frac{p}{n-1}x_1^\top B\sum_{j=3}^n x_j\right)\right] = \exp\left(\frac{p^2(n-2)}{2(n-1)^2}x_1^\top B\Sigma B^\top x_1\right).$$

Putting the pieces together, we see that

$$\mathbb{E}_{-1}[\pi_{12}^p] \leq \frac{1}{(n-1)^p}\exp\left(\frac{p^2}{2}x_1^\top B\Sigma B^\top x_1\right),$$

where we used the numerical fact that

$$\frac{(n-2)^2}{(n-1)^2} + \frac{n-2}{(n-1)^2} \leq 1.$$

We now prove the second inequality using a similar calculation. We observe that

$$\begin{aligned}
\pi_{11} &= \frac{\exp(x_1^\top Bx_1)}{\sum_{j=1}^n \exp(x_1^\top Bx_j)} \\
&\leq \frac{\exp(x_1^\top Bx_1)}{\sum_{j=2}^n \exp(x_1^\top Bx_j)} \\
&\leq \frac{1}{n-1}\exp\left(x_1^\top Bx_1\right)\exp\left(-\frac{1}{n-1}x_1^\top B\sum_{j=2}^n x_j\right),
\end{aligned}$$

where we applied the AM-GM inequality in the last step. It follows that

$$\mathbb{E}_{-1}[\pi_{11}^p] = \frac{1}{(n-1)^p}\exp\left(p\,x_1^\top Bx_1\right)\mathbb{E}_{-1}\left[\exp\left(-\frac{p}{n-1}x_1^\top B\sum_{j=2}^n x_j\right)\right].$$

We observe that

$$\mathbb{E}_{-1}\left[\exp\left(-\frac{p}{n-1}x_1^\top B\sum_{j=2}^n x_j\right)\right] = \exp\left(\frac{p^2}{2(n-1)}x_1^\top B\Sigma B^\top x_1\right),$$

which completes the proof. ■

Lemma 16 *Set*

$$R = \max\left(1,\ \sqrt{\mathrm{Tr}(\Sigma)} + 2\sqrt{2\|\Sigma\|_{\mathrm{op}}\log n}\right).$$

Fix $x_1 \in \mathbb{R}^d$ such that $\|x_1\|_2 \leq R$. The following inequality holds:

$$\left\|\mathbb{E}_{-1}[\mu_1] - \Sigma B^\top x_1\right\|_2^2 \leq \frac{KR^2}{n^2},$$

where $K \geq 0$ is a constant, for sufficiently large n.

Proof We see that

$$\begin{aligned}\left\| \mathbb{E}_{-1}[\mu_1] - \Sigma B^\top x_1 \right\|_F^2 &= \left\| \mathbb{E}_{-1}\left[\sum_{j=1}^n \pi_{1j} x_j\right] - \Sigma B^\top x_1 \right\|_2^2 \\ &= \left\| \mathbb{E}_{-1}\left[\sum_{j=1}^n \pi_{1j}(x_j - \Sigma B^\top x_1)\right] \right\|_2^2,\end{aligned}$$

where we used the fact that the softmax weights sum to unity. Using the fact that the covariates are i.i.d., we see that this expression is at most

$$2\left\| \mathbb{E}_{-1}[\pi_{11}](I - \Sigma B^\top)x_1 \right\|_2^2 + 2\left\|(n-1)\,\mathbb{E}_{-1}\left[\pi_{12}(x_2 - \Sigma B^\top x_1)\right]\right\|_2^2. \tag{12}$$

We bound each of the two terms of (12) separately. Applying homogeneity of the Frobenius norm and Lemma 15, we see that the first term is at most

$$\frac{2}{(n-1)^2}\exp\left(\frac{2}{n-1}x_1^\top B\Sigma B^\top x_1\right)\exp(2x_1^\top B x_1)\|I - \Sigma B^\top\|_{\mathrm{op}}^2\|x_1\|_2^2.$$

Applying the assumptions that $\|x_1\|_2 \leq R$ and the assumption that $\|B\|_{\mathrm{op}} \leq \gamma$, we see that this term is bounded above by KR^2n^{-2}. We now bound the second term of (12). Applying Gaussian integration by parts, we observe that

$$\mathbb{E}_{-1}\left[\pi_{12}(x_2 - \Sigma B^\top x_1)\right] = -\mathbb{E}_{-1}[\pi_{12}^2]\Sigma B^\top x_1.$$

It follows that the second term of (12) is at most

$$2(n-1)^2\left(\mathbb{E}_{-1}[\pi_{12}^2]\right)^2\left\|\Sigma B^\top x_1\right\|_F^2.$$

Applying Lemma 15, we see that this is at most

$$\frac{2}{(n-1)^2}\exp(4x_1^\top B\Sigma B^\top x_1)\|\Sigma B^\top\|_{\mathrm{op}}^2\,\|x_1\|_2^2.$$

Applying our assumptions on x_1 and B once more, we see that this term is at most KR^2n^{-2}. ■

Lemma 17 *Set*

$$R = \max\left(1,\ \sqrt{\mathrm{Tr}(\Sigma)} + 2\sqrt{2\|\Sigma\|_{\mathrm{op}}\log n}\right).$$

Fix $x_1 \in \mathbb{R}^d$ such that $\|x_1\|_2 \leq R$. The following inequality holds:

$$\left\|\mathbb{E}_{-1}[\Sigma_1] - \Sigma\right\|_F^2 \leq \frac{KR^8}{n^2},$$

where $K \geq 0$ is a constant.

Proof Recall that

$$\Sigma_1 = \sum_{j=1}^{n} \pi_{1j} x_j x_j^\top - \mu_1 \mu_1^\top.$$

It follows that

$$\|\mathbb{E}_{-1}[\Sigma_1] - \Sigma\|_F^2 \leq 2 \left\| \mathbb{E}_{-1}\left[\sum_{j=1}^{n} \pi_{1j} x_j x_j^\top\right] - (\Sigma + \Sigma B^\top x_1 x_1^\top B \Sigma) \right\|_F^2 + 2 \left\| \mathbb{E}_{-1}[\mu_1 \mu_1^\top] - \Sigma B^\top x_1 x_1^\top B \Sigma \right\|_F^2. \tag{13}$$

We bound each of these two terms separately.

We first bound the first term of (13). Using the fact that the covariates are i.i.d., we see that

$$\mathbb{E}_{-1}\left[\sum_{j=1}^{n} \pi_{1j} x_j x_j^\top\right] = \mathbb{E}_{-1}[\pi_{11}] x_1 x_1^\top + (n-1)\mathbb{E}_{-1}[\pi_{12} x_2 x_2^\top].$$

By Gaussian integration by parts,

$$\mathbb{E}_{-1}[\pi_{12} x_2 x_2^\top] = \mathbb{E}_{-1}[\pi_{12}]\Sigma + \mathbb{E}_{-1}[\pi_{12}(1-\pi_{12})(1-2\pi_{12})]\, \Sigma B^\top x_1 x_1^\top B \Sigma, \tag{14}$$

where we used the identities

$$\nabla_{x_2} \pi_{12} = \pi_{12}(1-\pi_{12}) B^\top x_1, \qquad \nabla_{x_2}^2 \pi_{12} = \pi_{12}(1-\pi_{12})(1-2\pi_{12}) B^\top x_1 x_1^\top B.$$

The softmax weights sum to one, which implies the elementary identity

$$\mathbb{E}_{-1}[\pi_{11}] + (n-1)\mathbb{E}_{-1}[\pi_{12}] = 1. \tag{15}$$

Applying (14) and (15), we see that the expression

$$\mathbb{E}_{-1}\left[\sum_{j=1}^{n} \pi_{1j} x_j x_j^\top\right] - (\Sigma + \Sigma B^\top x_1 x_1^\top B \Sigma)$$

can be rewritten as

$$\mathbb{E}_{-1}[\pi_{11}](x_1 x_1^\top - \Sigma) + ((n-1)\mathbb{E}_{-1}[\pi_{12}(1-\pi_{12})(1-2\pi_{12})] - 1)\, \Sigma B^\top x_1 x_1^\top B \Sigma.$$

We observe that

$$\pi_{12}(1-\pi_{12})(1-2\pi_{12}) = \pi_{12} - 3\pi_{12}^2 + 2\pi_{12}^3,$$

which implies that

$$(n-1)\mathbb{E}_{-1}[\pi_{12}(1-\pi_{12})(1-2\pi_{12})] - 1 = -\mathbb{E}_{-1}[\pi_{11}] - 3(n-1)\mathbb{E}_{-1}[\pi_{12}^2] + 2(n-1)\mathbb{E}_{-1}[\pi_{12}^3],$$

where we applied the identity (15). Applying Lemma 15 and the scaling assumptions on B, we see that for sufficiently large n, the following inequalities hold:

$$\mathbb{E}_{-1}[\pi_{11}] \leq \frac{K}{n}, \qquad \mathbb{E}_{-1}[\pi_{12}^2] \leq \frac{K}{n^2}, \qquad \mathbb{E}_{-1}[\pi_{12}^3] \leq \frac{K}{n^3}.$$

It follows that

$$|(n-1)\mathbb{E}_{-1}[\pi_{12}(1-\pi_{12})(1-2\pi_{12})]-1| \leq \frac{K}{n}.$$

Putting the pieces together, we see that

$$\left\|\mathbb{E}_{-1}\left[\sum_{j=1}^{n}\pi_{1j}x_jx_j^\top\right]-\Sigma-\Sigma B^\top x_1x_1^\top B\Sigma\right\|_F^2 \leq \frac{K}{n^2}\left(\|x_1x_1^\top-\Sigma\|_F^2+\|\Sigma B^\top x_1x_1^\top B\Sigma\|_F^2\right).$$

Using the assumption that $\|x_1\|_2 \leq R$ and the fact that $R \geq 1$, we obtain the bound

$$\left\|\mathbb{E}_{-1}\left[\sum_{j=1}^{n}\pi_{1j}x_jx_j^\top\right]-\Sigma-\Sigma B^\top x_1x_1^\top B\Sigma\right\|_F^2 \leq \frac{KR^4}{n^2}.$$

We now bound the second term of (13). We observe that

$$\mathbb{E}_{-1}[\mu_1\mu_1^\top]-\Sigma B^\top x_1x_1^\top B\Sigma$$

can be rewritten as the sum of two terms, namely

$$\mathbb{E}_{-1}\left[(\mu_1-\mathbb{E}_{-1}[\mu_1])(\mu_1-\mathbb{E}_{-1}[\mu_1])^\top\right]+(\mathbb{E}_{-1}[\mu_1]-\Sigma B^\top x_1)(\mathbb{E}_{-1}[\mu_1]-\Sigma B^\top x_1)^\top$$

and

$$(\mathbb{E}_{-1}[\mu_1]-\Sigma B^\top x_1)(\Sigma B^\top x_1)^\top+\Sigma B^\top x_1(\mathbb{E}_{-1}[\mu_1]-\Sigma B^\top x_1)^\top.$$

Applying Jensen's inequality, we see that

$$\left\|\mathbb{E}_{-1}[\mu_1\mu_1^\top]-\Sigma B^\top x_1x_1^\top B\Sigma\right\|_F^2 \leq 4\mathbb{E}_{-1}\left[\|\mu_1-\mathbb{E}_{-1}[\mu_1]\|_2^4\right]+8\|\Sigma B^\top x_1\|_2^2\,\|\mathbb{E}_{-1}[\mu_1]-\Sigma B^\top x_1\|_2^2+4\|\mathbb{E}_{-1}[\mu_1]-\Sigma B^\top$$

Applying Lemma 18 with $p=4$, along with Lemma 22 and the fact that $R \geq 1$, we see that

$$\mathbb{E}_{-1}\left[\|\mu_1-\mathbb{E}_{-1}[\mu_1]\|_2^4\right] \leq \frac{KR^8}{n^2}.$$

Lemma 16 immediately implies that

$$\|\mathbb{E}_{-1}[\mu_1]-\Sigma B^\top x_1\|_2^2 \leq \frac{KR^2}{n^2}.$$

Furthermore, it is clear that

$$\|\Sigma B^\top x_1\|_2^2 \leq KR^2.$$

These inequalities together imply the bound

$$\left\|\mathbb{E}_{-1}[\mu_1\mu_1^\top]-\Sigma B^\top x_1x_1^\top B\Sigma\right\|_F^2 \leq \frac{KR^8}{n^2}$$

for sufficiently large n. ■

Appendix D. Concentration of the softmax mean and softmax covariance

Lemma 18 *The following inequality holds for all $p \geq 2$:*

$$\mathbb{E}_{-1}\big\|\mu_1 - \mathbb{E}_{-1}[\mu_1]\big\|_F^p \leq C_p \tau_1^{2p^2} 2^{p-1}(n-1)^{-p/2}\|\Sigma\|_{\mathrm{op}}^{p/2}\Big(2^p\|B\|_{\mathrm{op}}^p\big(\mathbb{E}_{-1}[M_n^{4p}]\big)^{1/2} + d^{p/2}\Big),$$

where $\tau_1 = \exp\left(\frac{1}{2}x_1^\top B\Sigma B^\top x_1\right)$.

Proof

Pisier's inequality yields the bound

$$\mathbb{E}_{-1}\big\|\mu_1 - \mathbb{E}_{-1}[\mu_1]\big\|_F^p \leq C_p(n-1)^{p/2}\|\Sigma\|_{\mathrm{op}}^{p/2}\mathbb{E}_{-1}\big\|\nabla_{x_2}\mu_1\big\|_F^p, \tag{16}$$

where we used the fact that the covariates $\{x_j\}_{j=2}^n$ are i.i.d. We recall from Lemma 14 that

$$\nabla_{x_2}\mu_1 = \pi_{12}(x_2 - \mu_1)x_1^\top B + \pi_{12}I.$$

We see that

$$\big\|\nabla_{x_2}\mu_1\big\|_F^p \leq 2^{p-1}\pi_{12}^p\Big(\|x_2 - \mu_1\|_F^p\|B\|_{\mathrm{op}}^p\|x_1\|_2^p + d^{p/2}\Big). \tag{17}$$

Set $M_n = \sup_{i\in[n]}\|x_i\|_2$. It is clear that $\|x_1\|_2 \leq M_n$. Applying the triangle inequality and using the fact that μ_1 is a convex combination of the covariates, we see that $\|x_2 - \mu_1\|_2 \leq 2M_n$. Plugging these bounds into (17), we obtain the bound

$$\big\|\nabla_{x_2}\mu_1\big\|_F^p \leq 2^{p-1}\pi_{12}^p\Big(2^p\|B\|_{\mathrm{op}}^p M_n^{2p} + d^{p/2}\Big). \tag{18}$$

Plugging this bound into (16) and applying the Cauchy-Schwarz inequality, we obtain the bound

$$\mathbb{E}_{-1}\big\|\mu_1 - \mathbb{E}_{-1}[\mu_1]\big\|_F^p \leq C_p 2^{p-1}(n-1)^{p/2}\|\Sigma\|_{\mathrm{op}}^{p/2}\big(\mathbb{E}_{-1}[\pi_{12}^{2p}]\big)^{1/2}\Big(2^p\|B\|_{\mathrm{op}}^p\big(\mathbb{E}_{-1}[M_n^{4p}]\big)^{1/2} + d^{p/2}\Big).$$

Applying Lemma 15, we obtain the claim. ■

Lemma 19 *The following inequality holds for all $p \geq 2$:*

$$\mathbb{E}_{-1}\big\|\Sigma_1 - \mathbb{E}_{-1}[\Sigma_1]\big\|_F^p \leq C_p \tau_1^{2p^2} 4^{p-1}(n-1)^{-p/2}\|\Sigma\|_{\mathrm{op}}^{p/2}\Big(2^{2p+1}\|B\|_{\mathrm{op}}^p\big(\mathbb{E}_{-1}M_n^{6p}\big)^{1/2} + 2^{p+1}d^{p/2}\big(\mathbb{E}_{-1}M_n^{2p}\big)^{1/2}\Big),$$

where $\tau_1 = \exp\left(\frac{1}{2}x_1^\top B\Sigma B^\top x_1\right)$.

Proof Pisier's inequality yields the bound

$$\mathbb{E}_{-1}\big\|\Sigma_1 - \mathbb{E}_{-1}[\Sigma_1]\big\|_F^p \leq C_p(n-1)^{p/2}\|\Sigma\|_{\mathrm{op}}^{p/2}\mathbb{E}_{-1}\big\|\nabla_{x_2}\Sigma_1\big\|_F^p, \tag{19}$$

where we used the fact that the covariates $\{x_j\}_{j=2}^n$ are i.i.d. We recall from Lemma 14 that

$$\nabla_{x_2}\Sigma_1 = \pi_{12}\Big(\Big(\big(x_2 - \mu_1\big)\big(x_2 - \mu_1\big)^\top - \Sigma_1\Big)\otimes(B^\top x_1)\Big) + \pi_{12}\Big(I\otimes(x_2 - \mu_1) + (x_2 - \mu_1)\otimes I\Big).$$

We see that

$$\left\|\nabla_{x_2}\Sigma_1\right\|_F^p \leq 4^{p-1}\pi_{12}^p\Big(\|B\|_{\mathrm{op}}^p\|x_1\|_2^p\,\|x_2-\mu_1\|_F^{2p}+\|B\|_{\mathrm{op}}^p\|x_1\|_2^p\,\|\Sigma_1\|_F^p+2d^{p/2}\|x_2-\mu_1\|_2^p\Big). \quad (20)$$

Set $M_n = \sup_{i\in[n]}\|x_i\|_2$. It is clear that $\|x_1\|_2 \leq M_n$. Applying the triangle inequality and using the fact that μ_1 is a convex combination of the covariates, we see that $\|x_2-\mu_1\|_2 \leq 2M_n$. Similarly, convexity of the function $\|\cdot\|_F^p$ implies that

$$\|\Sigma_1\|_F^p \leq \sup_{i\in[n]}\left\|(x_i-\mu_1)(x_i-\mu_1)^\top\right\|_F^p$$

which in turn implies the bound $\|\Sigma_1\|_F^p \leq 4^p M_n^{2p}$. Plugging these bounds into (20), we obtain the bound

$$\left\|\nabla_{x_2}\Sigma_1\right\|_F^p \leq C_p 4^{p-1}\pi_{12}^p\Big(2^{2p+1}\|B\|_{\mathrm{op}}^p M_n^{3p}+2^{p+1}d^{p/2}M_n^p\Big).$$

Plugging this bound into (19) and applying the Cauchy-Schwarz inequality, we obtain the bound

$$\mathbb{E}_{-1}\left\|\Sigma_1-\mathbb{E}_{-1}[\Sigma_1]\right\|_F^p \leq C_p 4^{p-1}(n-1)^{p/2}\|\Sigma\|_{\mathrm{op}}^{p/2}\left(\mathbb{E}_{-1}\pi_{12}^{2p}\right)^{1/2}\Big(2^{2p+1}\|B\|_{\mathrm{op}}^p\left(\mathbb{E}_{-1}M_n^{6p}\right)^{1/2}+2^{p+1}d^{p/2}\left(\mathbb{E}_{-1}M_n^{2p}\right)^{1/2}\Big).$$

Applying Lemma 15, we obtain the claim. ■

Theorem 20 *Set*

$$R = \max\left(1,\ \sqrt{\mathrm{Tr}(\Sigma)}+2\sqrt{2\|\Sigma\|_{\mathrm{op}}\log n}\right), \qquad p = \max(2,\log(1/\delta)).$$

Let γ be defined as in Lemma 24 and assume that $\|B\|_F \leq \gamma$. Fix $\delta \in (0,1)$ and $x_1 \in \mathbb{R}^d$ such that $\|x_1\|_2 \leq R$. There exist constants K and N with polynomial dependence on M, Σ such that

$$\Pr\left(\left\|\mu_1-\mathbb{E}_{-1}[\mu_1]\right\|_2^2 \geq \frac{KR^4p}{n}\right) \leq \delta,$$

provided that $n \geq N$.

Proof Define

$$Z = \left\|\mu_1-\mathbb{E}_{-1}[\mu_1]\right\|_2.$$

Applying Markov's inequality, we observe that for any $t \geq 0$ and any $p \geq 1$,

$$\begin{aligned}\Pr\left(Z^2 \geq t\right) &= \Pr\left(Z \geq \sqrt{t}\right)\\ &\leq \left(\frac{\|Z\|_p}{\sqrt{t}}\right)^p,\end{aligned}$$

where we define

$$\|Z\|_p = \left(\mathbb{E}_{-1}[Z^p]\right)^{1/p}.$$

Set $p = \max(2, \log(1/\delta))$. Suppose that $\|Z\|_p \leq \frac{1}{3}\sqrt{K}R^2\sqrt{\frac{p}{n}}$ when n is sufficiently large, for some $K \geq 0$. Setting $t = KR^4\frac{p}{n}$, we see that

$$\begin{aligned}\Pr\left(Z \geq \sqrt{t}\right) &\leq 3^{-p} \\ &\leq 3^{\log(\delta)} \\ &= \delta^{\log 3} \\ &\leq \delta,\end{aligned}$$

where we used the fact that $\delta < 1$ and $\log 3 > 1$. We now show that when n is sufficiently large, $\|Z\|_p \leq \frac{1}{3}\sqrt{K}R^2\sqrt{\frac{p}{n}}$ for some $K \geq 0$. Applying Lemma 18, we see that

$$\|Z\|_p^p \leq C_p \tau_1^{2p^2} 2^{p-1}(n-1)^{-p/2}\|\Sigma\|_{\mathrm{op}}^{p/2}\left(2^p\|B\|_{\mathrm{op}}^p\left(\mathbb{E}_{-1}[M_n^{4p}]\right)^{1/2} + d^{p/2}\right),$$

where $\tau_1 = \exp\left(\frac{1}{2}x_1^\top B\Sigma B^\top x_1\right)$. It is clear that under the assumption that $\|x_1\|_2 \leq R$,

$$\begin{aligned}\tau_1^{2p} &\leq \exp\left(pR^2\|B\|_{\mathrm{op}}^2\|\Sigma\|_{\mathrm{op}}\right) \\ &\leq e,\end{aligned}$$

where we used the assumption that $\|B\|_{\mathrm{op}} \leq \gamma$.

Recall from Lemma 22 that

$$\mathbb{E}_{-1}[M_n^k] \leq \left(R + \sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{\|\Sigma\|_{\mathrm{op}}}\left(\sqrt{2\log n} + \sqrt{k+2}\right)\right)^k.$$

Assume that n is chosen such that $\sqrt{2\log n} \geq \sqrt{4p+2}$. We see that

$$\mathbb{E}_{-1}[M_n^{4p}] \leq (2R)^{4p}.$$

We also recall the elementary estimates $C_p^{1/p} \leq \frac{\pi}{2}\sqrt{p}$ and $(n-1)^{-1/2} \leq \sqrt{2}n^{-1/2}$ for $n \geq 2$. Putting the pieces together and using the fact that $R \geq 1$, we see that

$$\begin{aligned}\|Z\|_p &\leq 2\pi e\|\Sigma\|_{\mathrm{op}}^{1/2}\max\left(2\|B\|_{\mathrm{op}},\, \sqrt{d}\right)R^2\sqrt{\frac{p}{n}} \\ &= \frac{1}{3}\sqrt{K}R^2\sqrt{\frac{p}{n}},\end{aligned}$$

where we set

$$K = 9 \cdot 2^2\pi^2 e^2\|\Sigma\|_{\mathrm{op}}\max\left(4\|B\|_{\mathrm{op}}^2,\, d\right).$$

■

Theorem 21 *Set*

$$R = \max\left(1,\, \sqrt{\mathrm{Tr}(\Sigma)} + 2\sqrt{2\|\Sigma\|_{\mathrm{op}}\log n}\right), \qquad p = \max(2, \log(1/\delta)).$$

Let γ be defined as in Lemma 24 and assume that $\|B\|_F \leq \gamma$. Fix $\delta \in (0,1)$ and $x_1 \in \mathbb{R}^d$ such that $\|x_1\|_2 \leq R$. There exist constants K and N with polynomial dependence on M, Σ such that

$$\Pr\left(\left\|\Sigma_1 - \mathbb{E}_{-1}[\Sigma_1]\right\|_F^2 \geq \frac{KR^6 p}{n}\right) \leq \delta,$$

provided that $n \geq N$.

Proof Define

$$Z = \left\|\Sigma_1 - \mathbb{E}_{-1}[\Sigma_1]\right\|_F.$$

Applying Markov's inequality, we observe that for any $t \geq 0$ and any $p \geq 1$,

$$\begin{aligned} \Pr\left(Z^2 \geq t\right) &= \Pr\left(Z \geq \sqrt{t}\right) \\ &\leq \left(\frac{\|Z\|_p}{\sqrt{t}}\right)^p, \end{aligned}$$

where we define

$$\|Z\|_p = \left(\mathbb{E}_{-1}[Z^p]\right)^{1/p}.$$

Set $p = \max(2, \log(1/\delta))$. Suppose that $\|Z\|_p \leq \frac{1}{3}\sqrt{K}R^3\sqrt{\frac{p}{n}}$ when n is sufficiently large, for some $K \geq 0$. Setting $t = KR^6\frac{p}{n}$, we see that

$$\begin{aligned} \Pr\left(Z \geq \sqrt{t}\right) &\leq 3^{-p} \\ &\leq 3^{\log(\delta)} \\ &= \delta^{\log 3} \\ &\leq \delta, \end{aligned}$$

where we used the fact that $\delta < 1$ and $\log 3 > 1$. We now show that when n is sufficiently large, $\|Z\|_p \leq \frac{1}{3}\sqrt{K}R^3\sqrt{\frac{p}{n}}$ for some $K \geq 0$. Applying Lemma 19, we see that

$$\|Z\|_p^p \leq C_p \tau_1^{2p^2} 4^{p-1}(n-1)^{-p/2}\|\Sigma\|_{\mathrm{op}}^{p/2}\left(2^{2p+1}\|B\|_{\mathrm{op}}^p\left(\mathbb{E}_{-1}[M_n^{6p}]\right)^{1/2} + 2^{p+1}d^{p/2}\left(\mathbb{E}_{-1}[M_n^{2p}]\right)^{1/2}\right),$$

where $\tau_1 = \exp\left(\frac{1}{2}x_1^\top B\Sigma B^\top x_1\right)$. It is clear that under the assumption that $\|x_1\|_2 \leq R$,

$$\begin{aligned} \tau_1^{2p} &\leq \exp\left(pR^2\|B\|_{\mathrm{op}}^2\|\Sigma\|_{\mathrm{op}}\right) \\ &\leq e, \end{aligned}$$

where we used the assumption that $\|B\|_{\mathrm{op}} \leq \gamma$.

Recall from Lemma 22 that

$$\mathbb{E}_{-1}[M_n^k] \leq \left(R + \sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{\|\Sigma\|_{\mathrm{op}}}\left(\sqrt{2\log n} + \sqrt{k+2}\right)\right)^k.$$

Notice that this bound increases monotonically in k. Assume that n is chosen such that $\sqrt{2\log n} \geq \sqrt{6p+2}$. We see that

$$\max\left(\mathbb{E}_{-1}[M_n^{2p}], \mathbb{E}_{-1}[M_n^{6p}]\right) \leq (2R)^{6p}.$$

We also recall the elementary estimates $C_p^{1/p} \leq \frac{\pi}{2}\sqrt{p}$ and $(n-1)^{-1/2} \leq \sqrt{2}n^{-1/2}$ for $n \geq 2$. Putting the pieces together and using the fact that $R \geq 1$, we see that

$$\begin{aligned} \|Z\|_p &\leq 32\pi e \|\Sigma\|_{\mathrm{op}}^{1/2} \max(2\|B\|_{\mathrm{op}}, \sqrt{d}) R^3 \sqrt{\frac{p}{n}} \\ &= \frac{1}{3}\sqrt{K} R^3 \sqrt{\frac{p}{n}}, \end{aligned}$$

where we set

$$K = 9 \cdot 32^2 \pi^2 e^2 \|\Sigma\|_{\mathrm{op}} \max(4\|B\|_{\mathrm{op}}^2, d).$$

■

Lemma 22 *Let $R \geq 0$ an fix $x_1 \in \mathbb{R}^d$ such that $\|x_1\|_2 \leq R$. Let $M_n = \sup_{i\in[n]} \|x_i\|_2$. The following inequality holds:*

$$\mathbb{E}_{-1}[M_n^k] \leq \left(R + \sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{\|\Sigma\|_{\mathrm{op}}}\big(\sqrt{2\log n} + \sqrt{k+2}\big) \right)^k.$$

In particular, if n is chosen such that $2\log n \geq k+2$, and R is chosen such that

$$R \geq \sqrt{\mathrm{Tr}(\Sigma)} + 2\sqrt{2\|\Sigma\|_{\mathrm{op}} \log n},$$

then the following inequality holds:

$$\mathbb{E}_{-1}[M_n^k] \leq (2R)^k.$$

Proof It is clear that

$$M_n \leq R + \max_{2\leq i\leq n} \|x_i\|_2.$$

Applying Minkowski's inequality, we see that

$$\big(\mathbb{E}_{-1}[M_n^k]\big)^{1/k} \leq R + \left(\mathbb{E}\left[\max_{2\leq i\leq n} \|x_i\|_2^k \right] \right)^{1/k}.$$

Gaussian concentration of Lipschitz functions (Vershynin, 2018) lets us easily bound the second term:

$$\left(\mathbb{E}\left[\max_{2\leq i\leq n} \|x_i\|_2^k \right] \right)^{1/k} \leq \mathbb{E}\left[\max_{2\leq i\leq n} \|x_i\|_2 \right] + \sqrt{\|\Sigma\|_{\mathrm{op}}(k+2)}.$$

The Gaussian maximal inequality yields the bound

$$\mathbb{E}\left[\max_{2\leq i\leq n} \|x_i\|_2 \right] \leq \sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{2\|\Sigma\|_{\mathrm{op}} \log n}.$$

The claim follows immediately. ■

Lemma 23 *Let R be chosen such that*

$$R \geq \sqrt{\mathrm{Tr}(\Sigma)} + 2\sqrt{2\|\Sigma\|_{\mathrm{op}} \log n},$$

and let $x \sim \mathcal{N}(0, \Sigma)$. The following tail bound holds:

$$\Pr(\|x\|_2 \geq R) \leq n^{-4}.$$

Proof This claim follows immediately from the standard Gaussian tail bound

$$\Pr(\|x\|_2 \geq \sqrt{\mathrm{Tr}(\Sigma)} + t) \leq \exp\left(-\frac{t^2}{2\|\Sigma\|_{\mathrm{op}}}\right).$$

■

Appendix E. Initialization

Lemma 24 *Fix any $\delta \in (0, 1)$. Set*

$$\varepsilon_\Sigma = \min\left\{\|\Sigma\|_{\mathrm{op}}^{-1},\ \frac{\alpha}{8\sqrt{\beta}}\|\Sigma\|_{\mathrm{op}}^{-1/2}\right\}.$$

Let $\alpha, \beta, \varepsilon_0$, and $\mathcal{S}$ be defined as in Theorem 2, and let $\bar{\mathcal{S}}$ denote the set of points which are ε_0-close to $\mathcal{S}$ in P-norm. The probability that any of the following events fail to occur is at most δ, provided that n is sufficiently large:

(E_1) *The empirical covariance is close to the true covariance, and the inverse of the empirical covariance is close to the inverse of the true covariance:*

$$\max\left\{\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}},\ \|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_{\mathrm{op}}\right\} \leq \varepsilon_\Sigma.$$

(E_2) *The initial choice of parameters θ_0 lies in $\bar{\mathcal{S}}$.*

(E_3) *The renormalization factor $\hat{\gamma}$ is at most γ, where we define*

$$\begin{aligned}
\nu &= \max\left(1,\ \sqrt{\mathrm{Tr}(\Sigma)} + 2\sqrt{2\|\Sigma\|_{\mathrm{op}} \log n}\right), \\
\hat{\nu} &= \max\left(1,\ \sqrt{\mathrm{Tr}(\hat{\Sigma})} + 2\sqrt{2\|\hat{\Sigma}\|_{\mathrm{op}} \log n}\right), \\
p &= \max(2, \log(1/\delta))
\end{aligned}$$

and set

$$\gamma = \sqrt{\frac{1}{p\nu^2\|\Sigma\|_{\mathrm{op}}}}, \qquad \hat{\gamma} = \frac{1}{2}\sqrt{\frac{1}{p\hat{\nu}^2\|\Sigma\|_{\mathrm{op}}}}.$$

Proof We show that each event E_i occurs with probability tending to 1 as n increases; the union bound implies that the probability that any event fails to occur tends to zero, and in particular can be made at most δ by taking n sufficiently large.

(E_1) It is a standard result that

$$\max\left\{\|\hat{\Sigma}-\Sigma\|_F^2,\, \|\hat{\Sigma}^{-1}-\Sigma^{-1}\|_F^2\right\} \leq K \max\{\|\Sigma\|_{\mathrm{op}}^2,\, \|\Sigma^{-1}\|_{\mathrm{op}}^2\}\frac{d+\log(1/\delta)}{n}$$

with probability at least $1-\delta$, for some $K \geq 0$ and sufficiently large n. We refer to (Koltchinskii and Lounici, 2017; Vershynin, 2018) for details.

(E_2) Recall that the parameters are initialized at the point

$$\theta_0 = (A_0, B_0) = \begin{bmatrix} \hat{U}\hat{\Gamma}^{1/2}\hat{\Sigma}^{-1/2} \\ \hat{\Sigma}^{-1/2}\hat{V}\hat{\Gamma}^{1/2}\hat{\Sigma}^{-1/2} \end{bmatrix},$$

where $\hat{U}\hat{\Gamma}\hat{V}^\top$ is a singular value decomposition of $\hat{M}\hat{\Sigma}^{1/2}$ and

$$\hat{M} = \frac{1}{n}\sum_{i=1}^n y_i x_i^\top \hat{\Sigma}^{-1}.$$

Using the fact that $y_i = Mx_i + z_i$ for each $i \in [n]$, we see that

$$\begin{aligned} \hat{M} &= \frac{1}{n}\sum_{i=1}^n \left(Mx_ix_i^\top + z_ix_i^\top\right)\hat{\Sigma}^{-1} \\ &= \frac{1}{n}\sum_{i=1}^n \left(Mx_ix_i^\top + z_ix_i^\top\right)\hat{\Sigma}^{-1} \\ &= M + \frac{1}{n}\sum_{i=1}^n z_ix_i^\top\hat{\Sigma}^{-1}. \end{aligned}$$

It follows that

$$\|M-\hat{M}\|_{\mathrm{op}} \leq \left\|\frac{1}{n}\sum_{i=1}^n z_ix_i^\top\hat{\Sigma}^{-1}\right\|_{\mathrm{op}}.$$

It is clear that $\hat{M}$ converges to M in probability, because the random variables $\{z_i\}_{i=1}^n$ are i.i.d. zero mean subgaussian random variables which are independent of the covariates, so

$$\left\|\frac{1}{n}\sum_{i=1}^n z_ix_i^\top\hat{\Sigma}^{-1}\right\|_{\mathrm{op}}$$

converges to zero in probability. It follows that $\hat{M}\hat{\Sigma}^{1/2}$ converges to $M\Sigma^{1/2}$ in probability. Lemma 5.14 from Tu et al. (2016) implies that the factors $\hat{U}, \hat{\Gamma}, \hat{V}$ converge (up to orthogonal transformations) to U, Γ, V, where $U\Gamma V^\top$ is a singular value decomposition of $M\Sigma^{1/2}$. It follows that $\|\theta_0 - \theta_0^\star\|_P^2$ converges in probability to zero, where $\theta_0^\star$ is the projection of θ_0 onto $\mathcal{S}$.

(E_3) An identical argument shows that $\hat{\gamma}$ converges in probability to $\frac{1}{2}\gamma$ as n increases, and in particular satisfies $\hat{\gamma} \leq \gamma$ with probability $1 - \delta$ for sufficiently large n.

■

References

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). *arXiv preprint arXiv:2310.01082*, 2023.

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36, 2024.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pages 110–133, 2017.

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. In *International Conference on Machine Learning*, pages 19689–19729. PMLR, 2023.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 964–973, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/tu16.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.