Analysis of Catastrophic Forgetting for Random Orthogonal Transformation Tasks in the Overparameterized Regime

Daniel Goldfarb
Northeastern University

Paul Hand
Northeastern University

Abstract

Overparameterization is known to permit strong generalization performance in neural networks. In this work, we provide an initial theoretical analysis of its effect on catastrophic forgetting in a continual learning setup. We show experimentally that in Permuted MNIST image classification tasks, the generalization performance of multilayer perceptrons trained by vanilla stochastic gradient descent can be improved by overparameterization, and the extent of the performance increase achieved by overparameterization is comparable to that of state-of-the-art continual learning algorithms. We provide a theoretical explanation of this effect by studying a qualitatively similar two-task linear regression problem, where each task is related by a random orthogonal transformation. We show that when a model is trained on the two tasks in sequence without any additional regularization, the risk gain on the first task is small if the model is sufficiently overparameterized.

1 INTRODUCTION

Continual learning is the ability of a model to learn continuously from a stream of data, building on what was previously learned and retaining previously learned skills without the need for retraining. A major obstacle for neural networks to learn continually is the catastrophic forgetting problem: the abrupt drop in performance on previous tasks upon learning new ones. Modern neural networks are typically trained to greedily minimize a loss objective on a training set, and without any regularization, the model's performance on a previously trained task may degrade. Techniques for mitigating catastrophic forgetting fall under a wide family of groups (Delange et al., 2021). In this paper, we focus on comparison to regularization techniques. Generally, the goal of regularization methods is to determine important parameters from previous tasks and constrain them so that they do not get modified too much while training subsequent tasks. Two common state-of-the-art regularization methods are Synaptic Intelligence (SI) (Zenke et al., 2017) and Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017). See Appendix A for a detailed description of them.

It is well-known that strong generalization performance for neural networks is typically obtained in the overparameterized regime, where the number of learnable parameters is greater than the number of training examples. Work on overparameterized machine learning has led to research on the so-called double descent phenomenon, where test error improves as model complexity increases beyond the level needed to fit the training data, outperforming all underparameterized versions of the model (Belkin et al., 2019). One of the first observations of this behavior in modern neural networks was in extremely wide ResNet18 models that generalize better than their underparameterized counterparts on CIFAR-10 despite fitting to label noise (Nakkiran et al., 2021). This model-wise double descent phenomenon has been demonstrated analytically in a variety of machine learning models (Hastie et al., 2019, Belkin et al., 2020, Bartlett et al., 2020), including some as simple as linear regression. Such linear models will be the basis of theoretical analysis in the present paper.

There exists work that shows that the catastrophic forgetting of a multilayer perceptron (MLP) on the Rotated MNIST benchmark can be reduced simply by increasing the width of the architecture (Mirzadeh et al., 2022a). We replicate this work in the setting of 10 Permuted MNIST tasks, where each task has data given by random permutations of the original MNIST images (LeCun, 1998). We train 2-layer MLPs with a variety of layer widths given by $[400w, 400w]$, where $w = 1, 3, 5, 7, 9$, using vanilla stochastic gradient descent and the continual learning algorithms, SI and EWC. We compare the average test accuracy on all seen tasks. As expected, and as shown in Figure 1, average accuracy with SGD drops significantly after learn-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

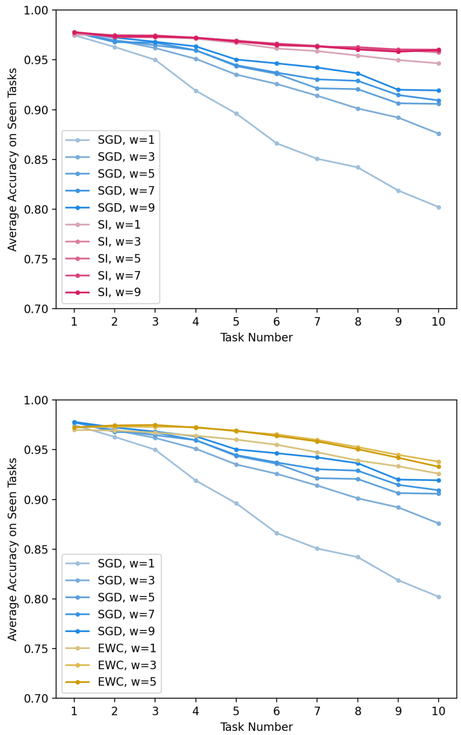

Figure 1: Results of Permuted MNIST experiment. Red curves denote performance of SI, yellow curves denote performance of EWC, blue curves denote performance of Vanilla SGD. Bolder saturation of lines corresponds to larger width parameters (more overparameterization). Specific hyperparameters are reported in Appendix B. Curves for $w = 7, 9$ for EWC are omitted due to computational constraints on Fisher matrix estimates.

ing multiple tasks, and that drop is mitigated by using SI or EWC. Interestingly, we observe that a significant fraction of the accuracy gain achieved by SI or EWC can be obtained with vanilla SGD by simply overparameterizing the model. This can be seen by comparing the $w = 1$ and $w = 9$ curves with SGD to the curves with SI and EWC. See Appendix B for more details on the experiments.

The goal of the present paper is to analytically illustrate the effect overparameterization can have on catastrophic forgetting. As with the latest illustrations of double descent (Hastie et al., 2019, Belkin et al., 2020, Bartlett et al., 2020), we choose to study the effect with a linear regression problem for simplicity and mathematical convenience. While the experiments above study the case of tasks related by random permutations, our analysis will instead study the qualitatively similar case where two tasks are related by a random orthogonal transformation, which results in simpler mathematical analysis, as we discuss in Section 3. The relation given by the random orthogonal transformation gives two data feature spaces corresponding to each task that are approximately but not exactly orthogonal. We construct two tasks, A and B. Let task A be defined by data matrix $X_A \in \mathbb{R}^{n \times p}$, with rows being p noisy random projections of some low-dimensional latent features, and responses $y \in \mathbb{R}^n$ that are noiseless and linear in the latent features. Let O be a random $p \times p$ orthogonal matrix. Then task B is defined by data $X_B = X_A O^\top$ and the same responses y. Learning tasks A and B involves estimating a $\beta \in \mathbb{R}^p$ for the predictor $f_\beta : x \mapsto x^\top \beta$ to fit the data (X_A, y) and (X_B, y), respectively.

We analyze the increase in statistical risk on task A between an estimator trained on task A by minimizing square loss, with initialization at zero, and one sequentially trained on task A and then task B with no explicit regularization. Let $R(f_\beta)$ be the risk on task A of an estimator f with parameters β. Let $\hat{\beta}_A$ be the parameters of the model that is trained on task A. Let $\hat{\beta}_{BA}$ be the parameters of a model that is initialized at $\hat{\beta}_A$ and then trained on task B. Our main result is that if there are more training examples than the intrinsic (latent) dimensionality of the data and if there is not too much noise in the observed features, then

$$R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) \lesssim \sqrt{\frac{n}{p}} \tag{1}$$

with high probability. The result asserts that under our linear model, the extent of catastrophic forgetting is arbitrarily small if the overparameterization ratio, p/n, is sufficiently large. We thus see an analytical illustration that catastrophic forgetting can be ameliorated by overparametization in the case of a suitable linear model. The full theorem is stated in Section 2.4 and its proof is provided in Appendix F.

The contributions of this paper are:

- We provide a linear regression problem that exhibits the effect that overparameterization can account for a majority of the performance drop due to catastrophic forgetting.
- We establish a non-asymptotic bound on the performance drop of this linear model in an orthogonal transformation task setting using results from random matrix theory. This result provides a formal illustration that continual learning can in some cases be ameliorated by overparameterization.

2 ANALYSIS OF CATASTROPHIC FORGETTING IN A LINEAR MODEL

In this section, we present a latent space model for linear regression that we will analyze in order to illustrate that overparameterization can ameliorate catastrophic forgetting. Our single task model is the latent space model of Hastie et al. (2019) without label noise. Then, we present the analogy between this linear model and neural networks. Next, we empirically demonstrate that under this model, overparameterization ameliorates catastrophic forgetting. Finally, we present a theorem that establishes that observation with high probability.

2.1 Latent Space Models for Two Linear Regression tasks

Let $\mathcal{Z} = \mathbb{R}^d$, which we call the latent feature space. Consider data for regression generated by a noiseless linear response to standard Gaussian latent features. That is, for some $\theta \in \mathbb{R}^d$, let an example be given by

$$z \sim \mathcal{N}(0, I_d), \tag{2}$$

$$y = z^\top \theta. \tag{3}$$

Let $\mathcal{X} = \mathbb{R}^p$, which we call the observed feature space. We consider the case where, for each example, we have access only to p observed features, given by noisy random projections of the latent features:

$$x = Wz + u \tag{4}$$

where $W \in \mathbb{R}^{p \times d}$ and $u \sim \mathcal{N}(0, I_p)$. We could take W to have i.i.d. $\mathcal{N}(0, \gamma)$ entries, but for mathematical convenience, we will instead study the idealization in which W has columns that form a scaled orthonormal basis of a random d-dimensional subspace of $\mathbb{R}^p$. Namely, $W^\top W = p\gamma I_d$. For large p, this idealization is approximately satisfied under the above Gaussian model for W.

We consider two tasks, A and B, both with n examples. Task A has data $(X_A, y) \in \mathbb{R}^{n \times p} \times \mathbb{R}^n$ where each of the n rows of X_A and entries of y are sampled independently by (2) - (4). Let O be sampled uniformly from the set of $p \times p$ orthogonal matrices. Task B has data (X_B, y) where $X_B = X_A O^\top$.

We study estimators that are linear in the observed features x:

$$f_{\hat{\beta}} : x \mapsto x^\top \hat{\beta}, \tag{5}$$

and we will sometimes refer to the parameters $\hat{\beta}$ as the estimator. We estimate the parameters of this model by gradient descent with a square loss. We are interested in the case of $d < n < p$. As $n < p$, the solution to this problem depends on initialization and solves the following optimization problem:

$$\arg\min_{\hat{\beta}} \frac{1}{2}\|\hat{\beta} - \beta_0\|^2 \text{ s.t. } y = X\hat{\beta}, \tag{6}$$

where β_0 is the initialization, and X is either X_A or X_B, depending on the task being solved. To study the sequential training of tasks A and B, we define the following estimators:

- $\hat{\beta}_A$ is the solution to task A when initialized at 0,
- $\hat{\beta}_B$ is the solution to task B when initialized at 0,
- $\hat{\beta}_{BA}$ is the solution to task B when initialized at $\hat{\beta}_A$.

These parameters are found by solving the following optimization problems:

$$\hat{\beta}_A = \arg\min_{\hat{\beta}} \frac{1}{2}\|\hat{\beta}\|^2 \text{ s.t. } y = X_A\hat{\beta}, \tag{7}$$

$$\hat{\beta}_B = \arg\min_{\hat{\beta}} \frac{1}{2}\|\hat{\beta}\|^2 \text{ s.t. } y = X_B\hat{\beta}, \tag{8}$$

$$\hat{\beta}_{BA} = \arg\min_{\hat{\beta}} \frac{1}{2}\|\hat{\beta} - \hat{\beta}_A\|^2 \text{ s.t. } y = X_B\hat{\beta}. \tag{9}$$

The optimization problem in (6) has the following closed form solution when X has rank n:

$$\hat{\beta} = \beta_0 + X^\top (XX^\top)^{-1} y - X^\top (XX^\top)^{-1} X \beta_0 \tag{10}$$

$$= \beta_0 + X^\top (XX^\top)^{-1} y - \mathcal{P}_{X^\top}\beta_0, \tag{11}$$

where $\mathcal{P}_{X^\top}$ is the orthogonal projector onto the range of $X^\top$. As $n < p$, X_A and X_B have rank n with probability 1, and this gives the following closed forms for $\hat{\beta}_A, \hat{\beta}_B, \hat{\beta}_{BA}$:

$$\hat{\beta}_A = X_A^\top (X_A X_A^\top)^{-1} y, \tag{12}$$

$$\hat{\beta}_B = X_B^\top (X_B X_B^\top)^{-1} y, \tag{13}$$

$$\hat{\beta}_{BA} = \hat{\beta}_A + \hat{\beta}_B - \mathcal{P}_{X_B^\top}\hat{\beta}_A. \tag{14}$$

We evaluate these estimators on task A. The risk on task A of an estimator f with parameters $\hat{\beta}$ is given by

$$R(f_{\hat{\beta}}) = \sigma^2 + (\hat{\beta} - \beta)^\top \Sigma (\hat{\beta} - \beta) \tag{15}$$

where

$$\Sigma = WW^\top + I_p \tag{16}$$

$$\beta = (I + WW^\top)^{-1} W\theta \tag{17}$$

$$\sigma^2 = \theta^\top (W^\top W + I_d)^{-1}\theta \tag{18}$$

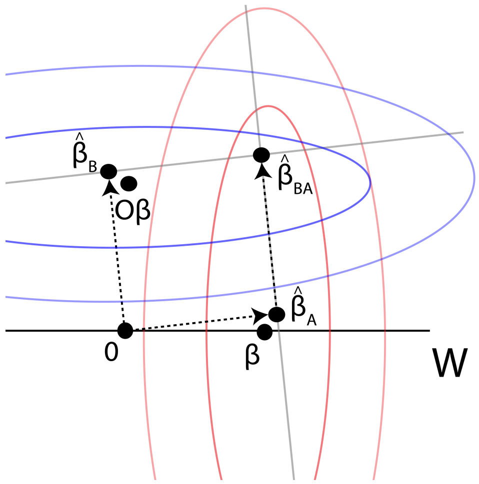

Figure 2: The solid black line depicts the span of W. The true parameters corresponding to tasks A and B are given by $\beta \in W$ and $O\beta$. The gray lines depict the set of solutions to $X_A\beta = y$ and $X_B\beta = y$. The estimators $\hat{\beta}_A, \hat{\beta}_B, \hat{\beta}_{BA}$ are given by orthogonal projections of an initialization on the respective consistent solutions. The red and blue ellipses depict lines of constant risk for tasks A and B, respectively. Note that the grey line through $\hat{\beta}_A$ is nearly (but not exactly) orthogonal to W due to noise in X_A. Consequently, β does not lie exactly in this set and the gray lines do not run precisely through the red and blue ellipses.

See Appendix E for the derivation of (15)–(18). It follows from showing that the latent space model described above is equivalent to an anisotropic regression model where X_A has i.i.d. rows $X_{A_i} \sim \mathcal{N}(0, \Sigma)$ and labels $y = X_A\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

We aim to bound $R(f_{\hat{\beta}_{BA}})$ relative to $R(f_{\hat{\beta}_A})$. Figure 2 illustrates the estimators $\hat{\beta}_A, \hat{\beta}_B, \hat{\beta}_{BA}$ and curves of constant risk. As depicted, if p is large enough, $\hat{\beta}_{BA}$ has low risk on Task A and Task B simultaneously.

2.2 Analogy of Linear Model to Neural Networks

The linear model we study is intended to be a mathematically tractable idealization of a neural network, and it is meant to analytically illustrate that overparameteriation can ameliorate catastrophic forgetting. The analogy of this linear model and neural network training on image data is as follows:

Natural images in a neural network's training distribution can be (approximately) modeled as being on a nonlinear manifold and having a low-dimensional latent representation. Instead of observing the latent representation of an image, the neural network only sees a high-dimensional representation either directly in pixel space or perhaps in a representation computed from pixel space. Either of these representations contain noise in the features used for prediction. Responses can be approximated by a neural network.

In our linear model, the low-dimensional representation of an input image is in a d-dimensional latent feature linear space. The responses are linear in the latent features. We assume the response is noiseless for the sake of simplicity, though our results could be extended to the noisy case. In our linear model, predictions are made off of a p-dimensional model given by noisy random projections of the latent features. We constrain W to have orthonormal columns which is a mathematical idealization of Gaussian measurements. We study two tasks with the same responses like in the permutation task setup, but for mathematical convenience we study tasks that are related by a random orthogonal transformation instead.

2.3 Numerical Experiment

Before we establish our theoretical result about the system described in Section 2.1, we provide empirical evidence that the latent space linear regression model above exhibits the phenomenon that overparameterization can ameliorate catastrophic forgetting. Specifically, we provide empirical evidence that $R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A})$ decreases with p.

Let $d = 20$, $n = 100$, $\gamma = 1$, $\beta_0 = \vec{0}$, and $\theta \sim \mathcal{N}(0, I_d)$. We plot $R(f_{\hat{\beta}_{BA}})$, $R(f_{\hat{\beta}_A})$, $R(f_{\beta_0})$ as a function of $p \in (n, 2000)$ averaged over 100 samplings of W, X_A, O, u.

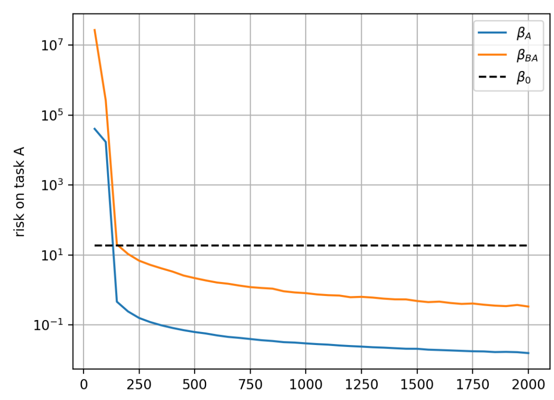

Figure 3: Result of simulated numerical experiment for random orthogonal transformation tasks. Risk is plotted as a function of model complexity p for model (2)–(14). Dashed black line denotes null risk (corresponding to the zero estimator), blue line denotes risk of estimator trained on task A, orange line denotes risk of estimator trained on task A then task B.

Figure 3 shows the results of the experiment. We first note that both $\hat{\beta}_A$ and $\hat{\beta}_{BA}$ outperform the null risk, given by $\beta = 0$. The null risk, $R(f_{\beta_0})$, defines a baseline that any reasonable model must beat. We also observe that $R(f_{\hat{\beta}_A})$ and $R(f_{\hat{\beta}_{BA}})$ are decreasing with p in the overparameterized regime, and that the difference between these risks is decreasing in p as well. Due to the log-scale of the vertical-axis, the latter observation is non-obvious. Appendix C includes a plot of the difference between $R(f_{\hat{\beta}_{BA}})$ and $R(f_{\hat{\beta}_A})$ to better illustrate the behavior. This provides evidence that catastrophic forgetting is alleviated in the overparameterized regime in our two-task learning setup.

2.4 Main Result

Our main result is an upper bound on the performance drop, defined as $R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A})$, for the two-task latent space linear regression model described above and inspired by the double descent literature. As described in Section 2.1, we consider (2)–(14), where W satisfies the following assumption.

Assumption 2.1. *All non-zero singular values of W are equal. Namely, $W^\top W = p\gamma I_d$.*

We begin with a proposition that defines the unlearned baseline for the problem.

Proposition 2.2. *Fix $\theta \in \mathbb{R}^d$. Let $W \in \mathbb{R}^{p \times d}$ satisfy Assumption 2.1. Then*

$$R(f_0) = \|\theta\|^2. \tag{19}$$

This risk calculation agrees with the numerical experiment in Section 2.3 where $\|\theta\|^2 \approx d = 20$. The result is formally stated and proven in Lemma F.3. For our main result, we prove that if the number of examples exceeds the problem's latent dimensionality, if the number of parameters is sufficiently large relative to the number of examples and relative to the noise level of the observable features, then with high probability, the performance drop is small.

Theorem 2.3. *Fix $\theta \in \mathbb{R}^d$. Let tasks A, B be given by (2)–(14). Let $W \in \mathbb{R}^{p \times d}$ satisfy Assumption 2.1 and $n \geq d, p \geq max(17n, 1/\gamma)$. Then there exists constant $c > 0$ such that with probability at least $1 - 10e^{-cd}$, the following holds:*

$$R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) \leq \left(66\sqrt{\frac{n}{p}} + \frac{12}{p\gamma}\right)\|\theta\|^2 \tag{20}$$

Theorem 2.3 provides an upper bound on the amount of risk gained on task A after subsequential training on task B given by two terms. The first term provides dependence on the overparameterization ratio p/n and decreases as overparameterization becomes more extreme. The second term is given by the signal to noise ratio of the noisy features. This term dominates only when $\gamma \ll 1/\sqrt{np}$. Based on the theorem, we observe that the overparameterization needs only to be linear in n to achieve a negligible performance drop in unregularized sequential task training compared to the baseline of $\|\theta\|^2$ in Proposition 2.2. This shows that catastrophic forgetting is ameliorated in the overparameterized regime. This result is formally stated in Lemma F.11. A formal proof and supporting lemmas are supplied in Appendix F. We provide a proof sketch here to outline the techniques used.

2.5 Proof Sketch of Theorem 2.3

For readability, we write X_A as A and X_B as B. As shown in Appendix E, the latent space model described above is equivalent to an anisotropic regression model where A has i.i.d. rows $A_i \sim \mathcal{N}(0, \Sigma)$ and labels $y = A\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

Performance drop is given by

$$\begin{aligned} R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) &= (\hat{\beta}_{BA} - \beta)^\top \Sigma (\hat{\beta}_{BA} - \beta) \\ &\quad - (\hat{\beta}_A - \beta)^\top \Sigma (\hat{\beta}_A - \beta). \end{aligned} \tag{21}$$

After substituting the closed form solutions for $\hat{\beta}_A, \hat{\beta}_{BA}$, distributing terms, and applying simple Cauchy-Schwarz and triangle inequalities, we get the following bound:

$$\begin{aligned} R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) &\leq 8\frac{p\gamma\sqrt{p\gamma}}{p\gamma+1}\|\theta\|\|\mathcal{P}_W\hat{\beta}_B\| \\ &\quad + 14\sqrt{p\gamma}\|\theta\|\|\mathcal{P}_{B^\top}\hat{\beta}_A\| \\ &\quad + 12\frac{p\gamma}{(p\gamma+1)^2}\|\theta\|^2 \qquad (22) \\ &= I + II + III \qquad (23) \end{aligned}$$

Lemmas F.5, F.6 establish results for orthogonal transformations to help bound $\|\mathcal{P}_W\hat{\beta}_B\|, \|\mathcal{P}_{B^\top}\hat{\beta}_A\|$ respectively. $\mathcal{P}_W$ is a projection onto a d-dimensional space which scales the norm in I by d/p. $\mathcal{P}_{B^\top}$ is a projection onto an n-dimensional space which scales the norm in II by n/p. As $d \leq n$ by assumption, II dominates I in the final bound. Using these results and simplifying gives the following bound with probability at least $1 - 10e^{-cd}$ for constant $c > 0$:

$$I + II \leq 66\sqrt{\frac{n}{p}}\|\theta\|^2 \tag{24}$$

We directly obtain

$$III \leq \frac{12}{p\gamma}\|\theta\|^2 \quad (25)$$

Combining these bounds completes the proof.

3 DISCUSSION

Overparameterization is a necessity for continual learning so that there can exist an infinity of potential optima for each task (Kirkpatrick et al., 2017). This makes it likely that there exists an optimum for some task B that is close to the solutions of some task A. We provide experimental evidence that overparameterization can provide additional benefits in combatting catastrophic forgetting for neural networks solving permutation tasks. However, we are not the first to observe an experimental relationship between overparameterization and forgetting. Prior work has shown that wide models can mitigate forgetting on a number of continual learning benchmarks (Mirzadeh et al., 2022a, Mirzadeh et al., 2022b, Ramasesh et al., 2021). We use a linear model with clear analogies to neural networks in order to study this behavior theoretically. In our analysis of the linear model in the overparameterized regime, non-asymptotic matrix estimates and results for orthogonal transformations provide bounds on the performance drop. Our main result shows that, under our model, catastrophic forgetting is ameliorated for sufficiently large overparameterization. For the linear setting we study, the behavior we observe can be explained geometrically: overparameterization causes the random orthogonal transformation tasks to live in approximately orthogonal subspaces, so training on subsequent tasks does not interrupt performance on learned tasks.

We view the present work as helping to establish initial results for continual learning theory. Before the field can rigorously understand machine learning algorithms in practice, the behavior of simple systems should be well understood. In particular, the behavior of linear systems with only gradient descent is the most natural initial result. Our work remarks that future theory should establish that continual learning algorithms beat not only a moderately parameterized baseline, but also the performance of extremely overparameterized models.

First we address the concern for using permutation tasks as realistic benchmarks for continual learning methods. Researchers believe that permutation tasks only provide a best-case for real world scenarios (Farquhar and Gal, 2018). Also, on a number of image classification datasets, MLPs do not experience forgetting when only two permutation tasks are being learned (Pfülb and Gepperth, 2019). Our experiments confirm this effect while also showing that overparameterization mitigates the observable forgetting on 10 task Permuted MNIST. Despite these critiques, we use permutation tasks as a launching point for theory because each task is of the same ‘difficulty’ and is amenable to mathematical analysis.

Next we discuss our choice to study the problem with a linear model. Linear regression is the simplest setting, for which we know, that exhibits double descent. The consensus of several works that study double descent in linear models is that the risk of a model is monotonically decreasing in the overparameterized regime with respect to number of parameters only if the data has low effective dimension and high ambient dimension compared to the number of training samples (Dar et al., 2021, Bartlett et al., 2020, Hastie et al., 2019). In order to have a model that has monotonically decreasing performance drop for a particular continual learning problem, it is a necessity that it exhibits monotonically decreasing risk on a single task. Additionally in recent work, connections have been made between neural networks and linear models using the so-called neural tangent kernel (NTK) phenomenon (Jacot et al., 2018). The parameterization of a neural network can be so large that training only changes its parameters slightly from its initialization, resulting in functions that can be accurately approximated linearly. Hence it is reasonable that the analysis of linear models can explain the behavior of neural networks.

We now remark at a technical level two choices in our analysis. The first is why we studied the case of random orthogonal transformation tasks instead of permutation tasks. The empirical performance between orthogonal and permutation tasks is similar; they both create tasks that are equally ’difficult’ for an MLP to learn, which spares us from needing to quantify problem difficulty. Appendix D provides evidence that permutation and orthogonal transformation tasks have the same difficulty in the linear setting. Also, the mathematical analysis is easier when studying orthogonal transformation tasks. With random orthogonal transformations, any subspace gets mapped to a random subspace, for which the values of coefficients are typically well spread out. With random permutations, some subspaces (e.g. those aligned with the standard basis elements) do not exhibit the same spreading effect, making the technical analysis more involved. Secondly, we do not present a bound on $R(f_{\hat{\beta}_A})$, though it is expected to approach zero for large p, as suggested by Figure 3. Whether or not this risk goes to 0 in p, the performance drop goes to 0 in p while the null estimator remains with constant risk. So the regression problem is being solved arbitrarily well for sufficiently large p.

With the growing popularity of continual learning, much of recent work is focused on developing new algorithms to mitigate catastrophic forgetting (Kirkpatrick et al., 2017, Zenke et al., 2017, Shin et al., 2017, Li and Hoiem, 2017). Only a few papers study the problem theoretically (Knoblauch et al., 2020, Bennani et al., 2020, Doan et al., 2021, Heckel, 2022, Benzing, 2022, Lee et al., 2021, Evron et al., 2022). Knoblauch et al. (2020) uses set theory to prove that, in general, continual learning problems are NP-hard, explaining why generative replay methods perform so well. Bennani et al. (2020) uses the NTK regime to prove generalisation guarantees for an existing continual learning method. Doan et al. (2021) uses an NTK overlap matrix to define a notion of task similarity and show that catastrophic forgetting is more severe when tasks have high similarity. Heckel (2022) studies a family of continual learning methods that uses approximations of the Hessian to determine parameter importance, presenting scenarios where continual learning provably fails and succeeds. Benzing (2022) shows that a number of regularization techniques that seem to be derived from differing philosophies actually all study a variation of the Fisher information matrix. Lee et al. (2021) studies a teacher-student setup where they derive the dynamics of test error to show that continual learning is most difficult when tasks have intermediate similarity. Evron et al. (2022) uses angle as a proxy for task similarity to study the best and worst cases of forgetting on task sequences. While this work is the most similar to our analysis, Evron et al. evaluate forgetting using training error. We believe that it is important to study statistical risk (expected test error) since the going goal for analysis of supervised models is to bound the performance on unseen data.

A natural next step is to study the regimes in which catastrophic forgetting is most problematic. This includes the setting where tasks do not have a nearly orthogonal relationship but also when data does not necessarily live on a low-dimensional manifold. We are also interested in understanding how the ideas of this paper generalize to other continual learning benchmarks and for more general neural network architectures. While our experiments use networks that have been trained using SGD, our analysis studies minimum-norm interpolators. Even when these two methods are solving for the same objective, there may be a gap in the resulting estimators (Zou et al., 2021). It would be interesting to learn what implications this has for our theory but we leave this for future work. Prior work found experimental evidence that catastrophic forgetting is most severe not when tasks are very dissimilar but when they only have an intermediate level of similarity (Ramasesh et al., 2020). Using orthogonality as a proxy for task similarity, this agrees with our work that shows that nearly orthogonal tasks are less prone to catastrophic forgetting. An interesting future work would be to formalize this notion of task similarity for our model. Moving forward, one goal of theory in continual learning is to be able to analytically compare algorithms. Our work provides a foundation of understanding this behavior in a simple linear regression setting. In order to push this work forward, either non-linear models need to be studied or tasks that are related by something more complex than permutations.

Acknowledgements

PH acknowledges support from NSF Awards DMS-2053448, DMS-1848087, and DMS-2022205.

References

[Bartlett et al., 2020] Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

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

[Belkin et al., 2020] Belkin, M., Hsu, D., and Xu, J. (2020). Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180.

[Bennani et al., 2020] Bennani, M. A., Doan, T., and Sugiyama, M. (2020). Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*.

[Benzing, 2022] Benzing, F. (2022). Unifying importance based regularisation methods for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2372–2396. PMLR.

[Dar et al., 2021] Dar, Y., Muthukumar, V., and Baraniuk, R. G. (2021). A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning. *arXiv preprint arXiv:2109.02355*.

[Delange et al., 2021] Delange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., and Tuytelaars, T. (2021). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Doan et al., 2021] Doan, T., Bennani, M. A., Mazoure, B., Rabusseau, G., and Alquier, P. (2021). A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *International Conference on Artificial Intelligence and Statistics*, pages 1072–1080. PMLR.

[Evron et al., 2022] Evron, I., Moroshko, E., Ward, R., Srebro, N., and Soudry, D. (2022). How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*, pages 4028–4079. PMLR.

[Farquhar and Gal, 2018] Farquhar, S. and Gal, Y. (2018). Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*.

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

[Heckel, 2022] Heckel, R. (2022). Provable continual learning via sketched jacobian approximations. In *International Conference on Artificial Intelligence and Statistics*, pages 10448–10470. PMLR.

[Jacot et al., 2018] Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*.

[Kirkpatrick et al., 2017] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

[Knoblauch et al., 2020] Knoblauch, J., Husain, H., and Diethe, T. (2020). Optimal continual learning has perfect memory and is np-hard. In *International Conference on Machine Learning*, pages 5327–5337. PMLR.

[LeCun, 1998] LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

[Lee et al., 2021] Lee, S., Goldt, S., and Saxe, A. (2021). Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pages 6109–6119. PMLR.

[Li and Hoiem, 2017] Li, Z. and Hoiem, D. (2017). Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

[Mirzadeh et al., 2022a] Mirzadeh, S. I., Chaudhry, A., Yin, D., Hu, H., Pascanu, R., Gorur, D., and Farajtabar, M. (2022a). Wide neural networks forget less catastrophically. In *International Conference on Machine Learning*, pages 15699–15717. PMLR.

[Mirzadeh et al., 2022b] Mirzadeh, S. I., Chaudhry, A., Yin, D., Nguyen, T., Pascanu, R., Gorur, D., and Farajtabar, M. (2022b). Architecture matters in continual learning. *arXiv preprint arXiv:2202.00275*.

[Nakkiran et al., 2021] Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003.

[Pfülb and Gepperth, 2019] Pfülb, B. and Gepperth, A. (2019). A comprehensive, application-oriented study of catastrophic forgetting in dnns. *arXiv preprint arXiv:1905.08101*.

[Ramasesh et al., 2020] Ramasesh, V. V., Dyer, E., and Raghu, M. (2020). Anatomy of catastrophic forgetting: Hidden representations and task semantics. *arXiv preprint arXiv:2007.07400*.

[Ramasesh et al., 2021] Ramasesh, V. V., Lewkowycz, A., and Dyer, E. (2021). Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*.

[Shin et al., 2017] Shin, H., Lee, J. K., Kim, J., and Kim, J. (2017). Continual learning with deep generative replay. *Advances in neural information processing systems*, 30.

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

[Zenke et al., 2017] Zenke, F., Poole, B., and Ganguli, S. (2017). Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR.

[Zou et al., 2021] Zou, D., Wu, J., Braverman, V., Gu, Q., and Kakade, S. (2021). Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on Learning Theory*, pages 4633–4635. PMLR.

Table 1: Hyperparameters for the MNIST experiments

Hyperparameter	SI	EWC	SGD
learning rate	0.1	0.01	0.1
batch size	64	128	64
epochs / dataset	5	20	5
c	0.1		
ξ	0.1		
λ		150	
fisher sample size		1,000	

A DESCRIPTION OF CONTINUAL LEARNING TECHNIQUES

Synaptic Intelligence (SI) is a regularization technique that assigns to each parameter of the network an estimate of importance for learned tasks (Zenke et al., 2017). This weight is determined in an online manner by tracking the amount that each parameter contributed to the decrease in loss during training. The weight is then used to penalize changes to the network parameters during subsequent training in the form of a regularization term added to the loss function.

Elastic Weight Consolidation (EWC) is a regularization technique that determines the importance of network weights using an estimation of the Fisher Information Matrix (Kirkpatrick et al., 2017). Near a minimum of the loss function, the diagonals of the Fisher matrix act as an estimate of the second order derivative of the loss with respect to each parameter. The magnitude of this derivative is used as a proxy for how sensitive the loss function is to fluctuation of the parameter. Constraining parameters according to their corresponding Fisher diagonal entries shows as an effective way of retaining the values of important weights from previous tasks while training on new ones.

B MNIST EXPERIMENTS

Table 1 reports the hyperparameters used in the MNIST experiments. All architectures used ReLU activation functions for the hidden layers and softmax for the output layers. Weights were initialized as $Unif(\frac{-1}{\sqrt{i}}, \frac{1}{\sqrt{i}})$ where i is the input dimension of the given layer. We adopted the same hyperparameters for SI as in the original paper (Zenke et al., 2017). To our surprise, EWC with default hyperparameters (Kirkpatrick et al., 2017) did not compete with SI. A basic grid search gave us a model that was more competitive. Blank entries mean that the hyperparameter is not relevant for the particular method. Curves for $w = 7, 9$ are omitted due to computational constraints in computing Fisher matrix estimates.

C PLOT OF DIFFERENCE IN RISK FOR LINEAR MODEL

In our paper, we define performance drop to be $R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A})$. This is the quantity that we analyze in our main theoretical result. However, it is not obvious from Figure 3 that this difference is decreasing in p due to the log-scale of the vertical-axis. Figure 4 plots this quantity directly (run independently from the experiment in Figure 3). We can observe the monotonically decreasing behavior in the purple curve, which provides evidence that catastrophic forgetting is alleviated in the overparameterized regime in our two-task learning setup.

D PERMUTED NUMERICAL EXPERIMENT

Figure 5 shows the result of the numerical experiment in Figure 3 but run independently with a random permutation matrix instead of a random orthogonal matrix. We observe that the same behavior holds in this scenario.

E EQUIVALENCE OF MODELS AND DERIVATION OF RISK

Recall in Section 2.1 where we defined the LSM model for linear regression. In this section we show that LSM is equivalent to an anisotropic regression model (ARM). We then use ARM to define the risk expression that we analyze theoretically.

We begin by defining ARM. Define data matrix $X \in \mathbb{R}^{n\times p}$ and responses $y = X\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, $\sigma^2 = \theta^\top(W^\top W + I_d)^{-1}\theta$, and $\beta = (I + WW^\top)^{-1}W\theta$ for some $\theta \in \mathbb{R}^d, W \in \mathbb{R}^{p\times d}$. Let rows X_i be independent random

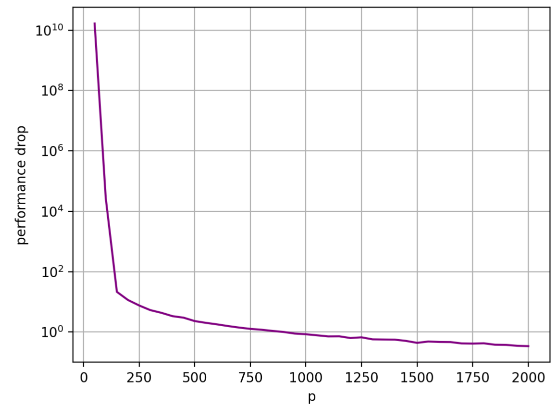

Figure 4: Result of simulated numerical experiment for orthogonal transformation tasks. Purple curve denotes the value of performance drop as a function of model complexity p for model (2)–(14).

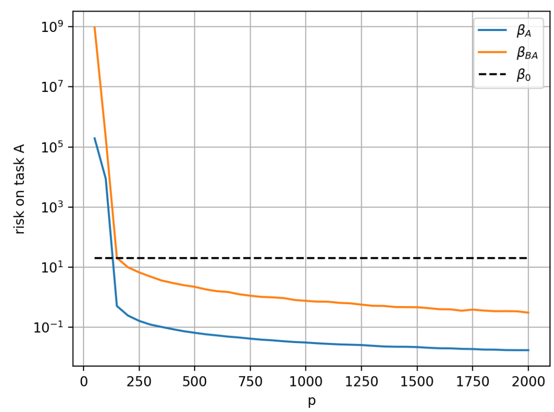

Figure 5: Result of simulated numerical experiment for random permutation tasks. Dotted black line denotes risk of null estimator, blue line denotes risk of estimator trained on task A, orange line denotes risk of estimator trained on task A then task B.

vectors in $\mathbb{R}^p$ with covariance $\Sigma = WW^\top + I_p$. Then the model is defined by the distribution over (X, y).

Next we show that ARM is equivalent to LSM. First observe that for both models, $(y_i, x_i^\top) \in \mathbb{R}^{p+1}$ are centered Gaussian vectors. Thus to show that they induce the same distribution, it suffices to show that they have the same covariance.

$$Cov((y_i, x_i^\top)^\top) = \mathbb{E}(y_i, x_i^\top)^\top (y_i, x_i^\top) = \mathbb{E}\begin{bmatrix} y_i^2 & y_i x_i^\top \\ y_i x_i & x_i x_i^\top \end{bmatrix} \tag{26}$$

We then compute the covariance matrices for each model.

Under LSM, we have:

$$\mathbb{E}[y_i^2] = \mathbb{E}(\theta^\top z_i)(z_i^\top \theta) = \theta^\top I \theta \tag{27}$$

$$\mathbb{E}[y_i x_i] = \mathbb{E}(W z_i + u_i)(z_i^\top \theta) = W\theta \tag{28}$$

$$\mathbb{E}[x_i x_i^\top] = \mathbb{E}(W x_i + u_i)(z_i^\top W^\top + u_i) = WW^\top + I \tag{29}$$

Plugging these quantities into (26) gives:

$$Cov((y_i, x_i^\top)^\top) = \begin{bmatrix} \|\theta\|^2 & (W\theta)^\top \\ W\theta & I + WW^\top \end{bmatrix} \tag{30}$$

Under ARM, we have:

$$\mathbb{E}[y_i^2] = \mathbb{E}(\beta^\top x_i + \epsilon_i)(x_i^\top \beta + \epsilon_i) = \beta^\top (I + WW^\top)\beta + \sigma^2 \tag{31}$$

$$\mathbb{E}[y_i x_i] = \mathbb{E}(x_i(x_i^\top \beta + \epsilon_i)) = \mathbb{E}(x_i x_i^\top \beta + x_i \epsilon_i) = (I + WW^\top)\beta \tag{32}$$

$$\mathbb{E}[x_i x_i^\top] = I + WW^\top \tag{33}$$

Plugging these quantities into (26) gives:

$$Cov((y_i, x_i^\top)^\top) = \begin{bmatrix} \beta^\top (I + WW^\top)\beta + \sigma^2 & ((I + WW^\top)\beta)^\top \\ (I + WW^\top)\beta & I + WW^\top \end{bmatrix} \tag{34}$$

We now show equivalence of the covariance matrices. Recall that for ARM, $\beta = (I + WW^\top)^{-1}W\theta$ and $\sigma^2 = \theta^\top (I + W^\top W)^{-1}\theta$. We first show equivalence of the first row, first column entries of the covariance matrices:

$$\begin{aligned} \beta^\top (I + WW^\top)\beta + \sigma^2 &= \theta^\top W^\top (I + WW^\top)^{-1} W\theta + \theta^\top (I + W^\top W)^{-1}\theta & (35) \\ &= \theta^\top (W^\top (I + WW^\top)^{-1} W + (I + W^\top W)^{-1})\theta & (36) \end{aligned}$$

By Lemma F.12, $(I + WW^\top)^{-1}W = W(I + W^\top W)^{-1}$. This gives

$$\begin{aligned} \beta^\top (I + WW^\top)\beta + \sigma^2 &= \theta^\top (W^\top W(I + W^\top W)^{-1} + (I + W^\top W)^{-1})\theta & (37) \\ &= \theta^\top (W^\top W + I)(I + W^\top W)^{-1}\theta & (38) \\ &= \theta^\top \theta = \|\theta\|^2 & (39) \end{aligned}$$

Next we show equivalence of the second row, first column entries of the covariance matrices:

$$(I + WW^\top)\beta = (I + WW^\top)(I + WW^\top)^{-1}W\theta = W\theta \tag{40}$$

The equivalence of the first row, second column entries also follows from this equality. The equivalence of the second row, second column entries is trivial.

Finally we derive the expression for the risk of ARM. By definition, the risk of an estimator f with parameters $\hat{\beta}$ has the following form:

$$\begin{aligned} R(f_{\hat{\beta}}) &= \mathbb{E}_{x,y}\|\hat{\beta}^\top x - y\|^2 & (41) \\ &= \mathbb{E}_{x,\epsilon}\|\hat{\beta}^\top x - \beta^\top x - \epsilon\|^2 & (42) \\ &= \mathbb{E}_x\|\hat{\beta}^\top x - \beta^\top x\|^2 + \mathbb{E}_\epsilon\|\epsilon\|^2 & (43) \\ &= \mathbb{E}_x\|(\hat{\beta} - \beta)^\top x\|^2 + \mathbb{E}_\epsilon\|\epsilon\|^2 & (44) \\ &= \mathbb{E}_x(\hat{\beta} - \beta)^\top x x^\top (\hat{\beta} - \beta) + \mathbb{E}_\epsilon\|\epsilon\|^2 & (45) \\ &= (\hat{\beta} - \beta)^\top \Sigma (\hat{\beta} - \beta) + \sigma^2 & (46) \end{aligned}$$

where the third equality holds from independence of ϵ and the sixth equality holds by definition of covariance.

We choose to study ARM with a slightly different but equivalent expression for β. Using Lemma F.12, $\beta = (I + WW^\top)^{-1}W\theta = W(W^\top W + I)^{-1}\theta$.

F SUPPORTING LEMMAS

We begin with an assumption, inspired by Hastie et al. (2019), that all non-zero singular values of W are equal.

Assumption F.1. *All non-zero singular values of W are equal. Namely, $W^\top W = p\gamma I_d$.*

Lemma F.2. *Assume $W \in \mathbb{R}^{p\times d}$ satisfies Assumption F.1. Then*

$$WW^\top = p\gamma \mathcal{P}_W \tag{47}$$

where $\mathcal{P}_W$ is the orthogonal projection onto the range of W.

Proof. We have that

$$WW^\top = WW^\top \frac{p\gamma}{p\gamma} = p\gamma W\left(\frac{1}{p\gamma}I_p\right)W^\top \tag{48}$$

By Assumption F.1, we have

$$WW^\top = p\gamma W(W^\top W)^{-1}W^\top \tag{49}$$

$W^\top W$ has full rank with probability 1, so $W(W^\top W)^{-1}W^\top$ is given explicitly by $\mathcal{P}_W$, which completes the proof. □

Lemma F.3. *Let $\Sigma = WW^\top + I_p$ where $W \in \mathbb{R}^{p\times d}$ satisfies Assumption F.1. For some $\theta \in \mathbb{R}^d$, let $\beta = W(W^\top W + I_d)^{-1}\theta$. Then*

$$R(f_{\vec{0}}) = \|\theta\|^2 \tag{50}$$

Proof. We have that

$$R(f_{\vec{0}}) = (\vec{0} - \beta)^\top \Sigma(\vec{0} - \beta) + \sigma^2 \tag{51}$$
$$= \beta^\top \Sigma\beta + \sigma^2 \tag{52}$$

By Lemma F.2, $\Sigma = p\gamma\mathcal{P}_W + I_p$ where $\mathcal{P}_W$ is the orthogonal projection onto the range of W, which gives

$$R(f_{\vec{0}}) = \beta^\top(p\gamma\mathcal{P}_W + I_p)\beta + \sigma^2 \tag{53}$$

Since $\beta \in range(W)$,

$$R(f_{\vec{0}}) = (p\gamma + 1)\|\beta\|^2 + \sigma^2 \tag{54}$$
$$= (p\gamma + 1)\theta^\top(W^\top W + I_d)^{-1}W^\top W(W^\top W + I_d)^{-1}\theta + \theta^\top(W^\top W + I_d)^{-1}\theta \tag{55}$$
$$= (p\gamma + 1)\theta^\top((p\gamma + 1)I_d)^{-1}p\gamma I_d((p\gamma + 1)I_d)^{-1}\theta + \theta^\top((p\gamma + 1)I_d)^{-1}\theta \tag{56}$$
$$= \left(\frac{p\gamma}{p\gamma + 1} + \frac{1}{p\gamma + 1}\right)\|\theta\|^2 = \|\theta\|^2 \tag{57}$$

□

Lemma F.4. *Let $x \sim \mathcal{N}(0, I_d)$, and $\epsilon \leq 1$, then*

$$\mathbb{P}\left(d(1-\epsilon) \leq \|x\|_2^2 \leq d(1+\epsilon)\right) \geq 1 - e^{-c\epsilon^2 d}$$

where $c > 0$ is an absolute constant.

Proof. This statement follows from Corollary 5.17 in Vershynin (2010), concerning concentration of sub-exponential random variables. □

Lemma F.5. *Assume $W \in \mathbb{R}^{p\times d}$ satisfies Assumption F.1. Let O be a random $p \times p$ orthogonal matrix. Fix $v \in \mathbb{R}^p$. Then, with probability at least $1 - 2e^{-c_1 d}$,*

$$\|\mathcal{P}_W O v\|^2 \leq \frac{2d}{p}\|v\|^2, \tag{58}$$

for some universal constant $c_1 > 0$.

Proof. Let $x = Ov$, and note that $\|x\| = \|v\|$, $\|x\| > 0$ with probability 1 and $\frac{x}{\|x\|} \sim \text{Uniform}(\mathbb{S}^{p-1})$. Letting $z \sim \mathcal{N}(0, I_p)$, we have that

$$\|\mathcal{P}_W O v\| \leq \left\|\mathcal{P}_W \frac{x}{\|x\|}\right\|\|v\| \overset{d}{=} \left\|\mathcal{P}_W \frac{z}{\|z\|}\right\|\|v\| \tag{59}$$

where the symbol $\overset{d}{=}$ means equality in distribution. Applying Lemma F.4 twice, we get that for any $\epsilon < 1$, with probability at least $1 - e^{-c\epsilon^2 p} - e^{-c\epsilon^2 d}$,

$$\|\mathcal{P}_W \frac{z}{\|z\|}\|\|v\| \leq \frac{\|\mathcal{P}_W z\|}{\sqrt{p}\sqrt{1-\epsilon}}\|v\| \leq \frac{\sqrt{d}\sqrt{1+\epsilon}}{\sqrt{p}\sqrt{1-\epsilon}}\|v\| \tag{60}$$

for some universal constant $c > 0$. By choosing suitable ϵ, we obtain that for $c_1 = c\epsilon^2$, with probability at least $1-2e^{-c_1 d}$, $\|\mathcal{P}_W O v\|^2 \leq \frac{2d}{p}\|v\|^2$. □

Lemma F.6. *Define $A \in \mathbb{R}^{n\times p}$ with rows A_i as independent random vectors in $\mathbb{R}^p$ with covariance $\Sigma = WW^\top + I_p$ where $W \in \mathbb{R}^{p\times d}$ satisfies Assumption F.1. Let O be a random $p\times p$ orthogonal matrix. Fix $v \in range(A^\top)$. Then with probability at least $1 - 2e^{-c_1 n}$*

$$\|\mathcal{P}_{OA^\top} v\|^2 \leq \frac{2n}{p}\|v\|^2 \tag{61}$$

for some universal constant $c_1 > 0$.

Proof. We have that

$$\begin{aligned}
\|\mathcal{P}_{OA^\top}\mathcal{P}_{A^\top} v\| &= \|O\mathcal{P}_{A^\top}O^\top \mathcal{P}_{A^\top} v\| = \|\mathcal{P}_{A^\top}O^\top\mathcal{P}_{A^\top} v\| && (62)\\
&= \|\mathcal{P}_{A^\top}\frac{O^\top\mathcal{P}_{A^\top}v}{\|O^\top\mathcal{P}_{A^\top}v\|}\|\|O^\top\mathcal{P}_{A^\top}v\| && (63)\\
&\leq \|\mathcal{P}_{A^\top}\frac{O^\top\mathcal{P}_{A^\top}v}{\|O^\top\mathcal{P}_{A^\top}v\|}\|\|v\| && (64)\\
&\overset{d}{=} \|\mathcal{P}_{A^\top}\frac{z}{\|z\|}\|\|v\| && (65)
\end{aligned}$$

where $z \sim \mathcal{N}(0, I_p)$ and the last equality follows from the rotational invariance of O. Applying Lemma F.4 twice, we get that for any $\epsilon < 1$, with probability at least $1 - e^{-c\epsilon^2 p} - e^{-c\epsilon^2 n}$,

$$\|\mathcal{P}_{A^\top}\frac{z}{\|z\|}\|\|v\| \leq \frac{\|\mathcal{P}_{A^\top} z\|}{\sqrt{p}\sqrt{1-\epsilon}}\|v\| \leq \frac{\sqrt{n}\sqrt{1+\epsilon}}{\sqrt{p}\sqrt{1-\epsilon}}\|v\| \tag{66}$$

for some universal constant $c > 0$. By choosing suitable ϵ, we obtain that for $c_1 = c\epsilon^2$, with probability at least $1-2e^{-c_1 n}$, $\|\mathcal{P}_{OA^\top} v\|^2 \leq \frac{2n}{p}\|v\|^2$. □

Lemma F.7. *Let $a \in \mathbb{R}^p$ be generated by $\mathcal{N}(0, \Sigma)$, $\Sigma = WW^\top + I_p$ where $W \in \mathbb{R}^{p \times d}$ satisfies Assumption F.1. Then*

$$\mathbb{E}\|a\|_2^2 = dp\gamma + p \tag{67}$$

Proof. It holds that $\mathbb{E}\|a\|_2^2 = \|\Sigma^{1/2}\|_F^2$ (Vershynin, 2010).

$$\|\Sigma^{1/2}\|_F^2 = \|\Sigma\|_* \tag{68}$$

where $\|\|_*$ denotes the nuclear norm. By Lemma F.2, $\Sigma = p\gamma\mathcal{P}_W + I_p$, which has d singular values of $p\gamma + 1$ and $p - d$ singular values of 1. So

$$\|\Sigma\|_* = d(p\gamma + 1) + p - d = dp\gamma + p \tag{69}$$

□

Lemma F.8. *Define $A \in \mathbb{R}^{n \times p}$ with rows A_i independent random vectors in $\mathbb{R}^p$ with covariance $\Sigma = p\gamma\mathcal{P}_W + I_p$ where $W \in \mathbb{R}^{p \times d}$ satisfies Assumption F.1. Then with probability at least $1 - 2e^{-n}$,*

$$\sigma_{min}(A)^2 \geq (\sqrt{p-d} - 2\sqrt{n})^2 \tag{70}$$

Proof. WLOG let $range(W) = span(e_1, ..., e_d)$. Then we can decompose A into two pieces: $A_{(1)} \in \mathbb{R}^{n \times d}$ with i.i.d. $\mathcal{N}(0, p\gamma + 1)$ entries and $A_{(2)} \in \mathbb{R}^{n \times p-d}$ with i.i.d. $\mathcal{N}(0, 1)$ entries. This gives

$$\sigma_{min}(A)^2 = \sigma_{min}(AA^\top) = \sigma_{min}(A_{(1)}A_{(1)}^\top + A_{(2)}A_{(2)}^\top) \tag{71}$$
$$\geq \sigma_{min}(A_{(2)}A_{(2)}^\top) = \sigma_{min}(A_{(2)})^2 = \sigma_{min}(A_{(2)}^\top)^2 \tag{72}$$

By Theorem 5.39 in Vershynin (2010), $\sigma_{min}(A_{(2)}^\top)^2 \geq (\sqrt{p-d} - 2\sqrt{n})^2$ with probability at least $1 - 2e^{-n}$. □

Lemma F.9. *Define $A \in \mathbb{R}^{n \times p}$ with rows A_i independent random vectors in $\mathbb{R}^p$ with covariance $\Sigma = WW^\top + I_p$ where $W \in \mathbb{R}^{p \times d}$ satisfies Assumption F.1. Let $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\sigma^2 = \theta^\top(W^\top W + I_d)^{-1}\theta$ for some $\theta \in \mathbb{R}^d$. Then with probability at least $1 - 2e^{-n}$,*

$$\|A^\top(AA^\top)^{-1}\epsilon\|^2 \leq \frac{n\|\theta\|^2}{p\gamma(\sqrt{p-d} - 2\sqrt{n})^2}, \tag{73}$$

Proof. We have that

$$\|A^\top(AA^\top)^{-1}\epsilon\|^2 \leq \|A^\top(AA^\top)^{-1}\|^2\|\epsilon\|^2 \tag{74}$$

We have that $\|\epsilon\|^2 = n\sigma^2 = n\theta^\top(W^\top W + I_d)^{-1}\theta$. Under Assumption F.1, this gives $\|\epsilon\|^2 = \frac{n\|\theta\|^2}{p\gamma+1}$,

$$\|A^\top(AA^\top)^{-1}\epsilon\|^2 \leq \|A^\top(AA^\top)^{-1}\|^2 \frac{n\|\theta\|^2}{p\gamma + 1} \tag{75}$$

It holds that $\|A^\top(AA^\top)^{-1}\| = 1/\sigma_{min}(A^\top)$ where $\sigma_{min}(A^\top)$ is the smallest singular value of $A^\top$. By Lemma F.8, $\sigma_{min}(A) \geq \sqrt{p-d} - 2\sqrt{n}$ with probability at least $1 - 2e^{-n}$. This gives

$$\|A^\top(AA^\top)^{-1}\epsilon\|^2 \leq \frac{1}{(\sqrt{p-d} - 2\sqrt{n})^2} \cdot \frac{n\|\theta\|^2}{p\gamma + 1} \leq \frac{n\|\theta\|^2}{p\gamma(\sqrt{p-d} - 2\sqrt{n})^2} \tag{76}$$

□

Lemma F.10. *Suppose $W \in \mathbb{R}^{p\times d}$ satisfies Assumption F.1. Let $\beta = W(W^\top W + I_d)^{-1}\theta$ for some $\theta \in \mathbb{R}^d$. If $p \geq 1/\gamma$ and $p \geq 16n + d$, then*

$$\frac{\sqrt{n}\|\theta\|}{\sqrt{p\gamma}(\sqrt{p-d} - 2\sqrt{n})} \leq \|\beta\| \tag{77}$$

Proof. We have that $\|\beta\|^2 = \theta^\top(W^\top W + I_d)^{-1}W^\top W(W^\top W + I_d)\theta$. Using Assumption F.1, this gives $\|\beta\| = \frac{\sqrt{p\gamma}}{p\gamma+1}\|\theta\|$. Suppose $p \geq 1/\gamma$, then we have that $\|\beta\| \geq \frac{1}{2\sqrt{p\gamma}}\|\theta\|$. When $p \geq 16n + d$, $\frac{\sqrt{n}}{\sqrt{p-d}-2\sqrt{n}} \leq \frac{1}{2}$, which gives

$$\frac{\sqrt{n}\|\theta\|}{\sqrt{p\gamma}(\sqrt{p-d} - 2\sqrt{n})} \leq \frac{1}{2\sqrt{p\gamma}}\|\theta\| \leq \|\beta\| \tag{78}$$

□

Theorem F.11. *Suppose $W \in \mathbb{R}^{p\times d}$ follows Assumption F.1. Define data matrix $A \in \mathbb{R}^{n\times p}$ and responses $y = A\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, $\sigma^2 = \theta^\top(W^\top W + I_d)^{-1}\theta$, and $\beta = W(W^\top W + I_d)^{-1}\theta$ for some $\theta \in \mathbb{R}^d$. Let rows A_i be independent random vectors in $\mathbb{R}^p$ with covariance $\Sigma = WW^\top + I_p$ and $n \geq d, p \geq max(17n, 1/\gamma)$. Let O be a random $p \times p$ orthogonal matrix and $B = AO^\top$. Let $\hat{\beta}_A$ be the parameters of the minimum norm estimator on A, and $\hat{\beta}_{BA}$ be the parameters of the estimator on B using $\hat{\beta}_A$ as initialization as defined in Section 2.1. Let $R(f_{\hat{\beta}})$ be the risk on task A of an estimator with parameters $\hat{\beta}$. Then there exists constant $c > 0$ such that with probability at least $1 - 10e^{-cd}$, the following holds:*

$$R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) \leq \left(\frac{66\sqrt{n}}{\sqrt{p}} + \frac{12}{p\gamma}\right)\|\theta\|^2 \tag{79}$$

Proof. We have that

$$\begin{aligned} R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) &= (\hat{\beta}_{BA} - \beta)^\top\Sigma(\hat{\beta}_{BA} - \beta) - (\hat{\beta}_A - \beta)^\top\Sigma(\hat{\beta}_A - \beta) & (80)\\ &= (\hat{\beta}_A + \hat{\beta}_B - \mathcal{P}_{B^\top}\hat{\beta}_A - \beta)^\top\Sigma(\hat{\beta}_A + \hat{\beta}_B - \mathcal{P}_{B^\top}\hat{\beta}_A - \beta)\\ &\quad - (\hat{\beta}_A - \beta)^\top\Sigma(\hat{\beta}_A - \beta) & (81) \end{aligned}$$

Distributing terms with $\hat{\beta}_B$ and $\mathcal{P}_{B^\top}\hat{\beta}_A$ gives

$$\begin{aligned} R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) &= (\hat{\beta}_A - \beta)^\top\Sigma(\hat{\beta}_A - \beta) - (\hat{\beta}_A - \beta)^\top\Sigma(\hat{\beta}_A - \beta) + 2\hat{\beta}_A^\top\Sigma\hat{\beta}_B - 2\hat{\beta}_B^\top\Sigma\beta\\ &\quad + \hat{\beta}_B^\top\Sigma\hat{\beta}_B - 2\hat{\beta}_A^\top\Sigma\mathcal{P}_{B^\top}\hat{\beta}_A - 2\hat{\beta}_B^\top\Sigma\mathcal{P}_{B^\top}\hat{\beta}_A + 2\beta^\top\Sigma\mathcal{P}_{B^\top}\hat{\beta}_A\\ &\quad + (\mathcal{P}_{B^\top}\hat{\beta}_A)^\top\Sigma\mathcal{P}_{B^\top}\hat{\beta}_A & (82)\\ &= 2\hat{\beta}_A^\top\Sigma\hat{\beta}_B - 2\hat{\beta}_B^\top\Sigma\beta + \hat{\beta}_B^\top\Sigma\hat{\beta}_B - 2\hat{\beta}_A^\top\Sigma\mathcal{P}_{B^\top}\hat{\beta}_A - 2\hat{\beta}_B^\top\Sigma\mathcal{P}_{B^\top}\hat{\beta}_A\\ &\quad + 2\beta^\top\Sigma\mathcal{P}_{B^\top}\hat{\beta}_A + (\mathcal{P}_{B^\top}\hat{\beta}_A)^\top\Sigma\mathcal{P}_{B^\top}\hat{\beta}_A & (83) \end{aligned}$$

By Lemma F.2, $\Sigma = p\gamma\mathcal{P}_W + I_p$ where $\mathcal{P}_W$ is the orthogonal projection onto the range of W. This implies that $\|\Sigma\| = p\gamma + 1$, giving

$$\begin{aligned} R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) &= 2\hat{\beta}_A^\top(p\gamma\mathcal{P}_W + I_p)\hat{\beta}_B - 2\hat{\beta}_B^\top(p\gamma\mathcal{P}_W + I_p)\beta + \hat{\beta}_B^\top(p\gamma\mathcal{P}_W + I_p)\hat{\beta}_B \\ &\quad + (\mathcal{P}_{B^\top}\hat{\beta}_A)^\top(p\gamma\mathcal{P}_W + I_p)(\mathcal{P}_{B^\top}\hat{\beta}_A - 2\hat{\beta}_A - 2\hat{\beta}_B + 2\beta) \quad (84) \\ &\leq 2p\gamma\hat{\beta}_A^\top\mathcal{P}_W\hat{\beta}_B + 2\hat{\beta}_A^\top\hat{\beta}_B - 2p\gamma\hat{\beta}_B^\top\mathcal{P}_W\beta - 2\hat{\beta}_B^\top\beta + p\gamma\hat{\beta}_B^\top\mathcal{P}_W\hat{\beta}_B + \hat{\beta}_B^\top\hat{\beta}_B \\ &\quad + (p\gamma + 1)\|\mathcal{P}_{B^\top}\hat{\beta}_A\|\|\mathcal{P}_{B^\top}\hat{\beta}_A - 2\hat{\beta}_A - 2\hat{\beta}_B + 2\beta\| \quad (85) \end{aligned}$$

Applying Cauchy-Schwarz and triangle inequality gives the following bound:

$$\begin{aligned} R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) &\leq 2p\gamma\|\hat{\beta}_A\|\|\mathcal{P}_W\hat{\beta}_B\| + 2\|\hat{\beta}_A\|\|\hat{\beta}_B\| + 2p\gamma\|\beta\|\|\mathcal{P}_W\hat{\beta}_B\| + 2\|\hat{\beta}_B\|\|\beta\| \\ &\quad + p\gamma\|\hat{\beta}_B\|\|\mathcal{P}_W\hat{\beta}_B\| + \|\hat{\beta}_B\|^2 \\ &\quad + (p\gamma + 1)\|\mathcal{P}_{B^\top}\hat{\beta}_A\|(\|\mathcal{P}_{B^\top}\hat{\beta}_A\| + 2\|\hat{\beta}_A\| + 2\|\hat{\beta}_B\| + 2\|\beta\|) \quad (86) \end{aligned}$$

By definition in Section 2, $\hat{\beta}_A = \mathcal{P}_{A^\top}\beta + A^\top(AA^\top)^{-1}\epsilon$ and $\hat{\beta}_B = O\mathcal{P}_{A^\top}\beta + OA^\top(AA^\top)^{-1}\epsilon$ and it holds that $\|\mathcal{P}_{A^\top}\beta\| \leq \|\beta\|$. By Lemma F.9, $\|A^\top(AA^\top)^{-1}\epsilon\| \leq \frac{\sqrt{n}\|\theta\|}{\sqrt{p\gamma}(\sqrt{p-d} - 2\sqrt{n})}$ with probability at least $1 - 2e^{-n}$ (call this Event E). So by Lemma F.10 if $p \geq max(17n, 1/\gamma)$, then $\|\hat{\beta}_A\| \leq 2\|\beta\|$ and $\|\hat{\beta}_B\| \leq 2\|\beta\|$, which gives

$$R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) \leq 8p\gamma\|\beta\|\|\mathcal{P}_W\hat{\beta}_B\| + 14(p\gamma + 1)\|\beta\|\|\mathcal{P}_{B^\top}\hat{\beta}_A\| + 12\|\beta\|^2 \quad (87)$$

We have that $\|\beta\|^2 = \theta^\top(W^\top W + I_d)^{-1}W^\top W(W^\top W + I_d)\theta$. Using Assumption F.1, this gives $\|\beta\| = \frac{\sqrt{p\gamma}}{p\gamma+1}\|\theta\|$,

$$\begin{aligned} R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) &\leq 8\frac{p\gamma\sqrt{p\gamma}}{p\gamma + 1}\|\theta\|\|\mathcal{P}_W\hat{\beta}_B\| + 14\sqrt{p\gamma}\|\theta\|\|\mathcal{P}_{B^\top}\hat{\beta}_A\| + 12\frac{p\gamma}{(p\gamma + 1)^2}\|\theta\|^2 \quad (88) \\ &= I + II + III \quad (89) \end{aligned}$$

We will bound each of these terms separately, starting with I:

Substituting $\hat{\beta}_B = B^\top(BB^\top)^{-1}y$ into this expression and distributing accordingly, we get that $\mathcal{P}_W\hat{\beta}_B = \mathcal{P}_W O\mathcal{P}_{A^\top}\beta + \mathcal{P}_W OA^\top(AA^\top)^{-1}\epsilon$,

$$\begin{aligned} I &= 8\frac{p\gamma\sqrt{p\gamma}}{p\gamma + 1}\|\theta\|\|\mathcal{P}_W O\mathcal{P}_{A^\top}\beta + \mathcal{P}_W OA^\top(AA^\top)^{-1}\epsilon\| \quad (90) \\ &\leq 8\frac{p\gamma\sqrt{p\gamma}}{p\gamma + 1}\|\theta\|\|\mathcal{P}_W O\mathcal{P}_{A^\top}\beta\| + 8\frac{p\gamma\sqrt{p\gamma}}{p\gamma + 1}\|\theta\|\|\mathcal{P}_W OA^\top(AA^\top)^{-1}\epsilon\| \quad (91) \end{aligned}$$

By Lemma F.5, there exists constant $c_1 > 0$ such that $\|\mathcal{P}_W O\mathcal{P}_{A^\top}\beta\| \leq 1.5\sqrt{\frac{d}{p}}\|\beta\|$ and $\|\mathcal{P}_W OA^\top(AA^\top)^{-1}\epsilon\| \leq 1.5\sqrt{\frac{d}{p}}\|A^\top(AA^\top)^{-1}\epsilon\|$ with probability at least $1 - 2e^{-c_1 d}$ each. By Lemma F.9, $\|A^\top(AA^\top)^{-1}\epsilon\| \leq \frac{\sqrt{n}\|\theta\|}{\sqrt{p\gamma}(\sqrt{p-d} - 2\sqrt{n})}$ (failure probability already accounted for on Event E). This gives the following bound with probability at least $1 - 4e^{-c_1 d}$:

$$I \leq 12\frac{p\gamma\sqrt{d\gamma}}{p\gamma + 1}\|\theta\|\|\beta\| + 12\frac{\gamma\sqrt{ndp}}{(p\gamma + 1)(\sqrt{p-d} - 2\sqrt{n})}\|\theta\|^2 \quad (92)$$

Substituting $\|\beta\| = \frac{\sqrt{p\gamma}}{p\gamma+1}\|\theta\|$ gives

$$I \leq 12\frac{p\gamma^2\sqrt{dp}}{(p\gamma+1)^2}\|\theta\|^2 + 12\frac{\gamma\sqrt{ndp}}{(p\gamma+1)(\sqrt{p-d}-2\sqrt{n})}\|\theta\|^2 \tag{93}$$

Using $p\gamma + 1 > p\gamma$ gives

$$I \leq 12\frac{\sqrt{d}}{\sqrt{p}}\|\theta\|^2 + 12\frac{\sqrt{nd}}{\sqrt{p}(\sqrt{p-d}-2\sqrt{n})}\|\theta\|^2 \tag{94}$$

Now we bound term II:

Substituting $\hat{\beta}_A = A^\top(AA^\top)^{-1}y$ into this expression and distributing accordingly, we get that $\mathcal{P}_{B^\top}\hat{\beta}_A = \mathcal{P}_{B^\top}\mathcal{P}_{A^\top}\beta + \mathcal{P}_{B^\top}A^\top(AA^\top)^{-1}\epsilon$. This gives the following bound:

$$II = 14\sqrt{p\gamma}\|\theta\|\|\mathcal{P}_{B^\top}\mathcal{P}_{A^\top}\beta + \mathcal{P}_{B^\top}A^\top(AA^\top)^{-1}\epsilon\| \tag{95}$$

$$\leq 14\sqrt{p\gamma}\|\theta\|\|\mathcal{P}_{B^\top}\mathcal{P}_{A^\top}\beta\| + 14\sqrt{p\gamma}\|\theta\|\|\mathcal{P}_{B^\top}A^\top(AA^\top)^{-1}\epsilon\| \tag{96}$$

By Lemma F.6, there exists constant $c_2 > 0$ such that $\|\mathcal{P}_{B^\top}\mathcal{P}_{A^\top}\beta\| \leq 1.5\sqrt{\frac{n}{p}}\|\beta\|$ and $\|\mathcal{P}_{B^\top}A^\top(AA^\top)^{-1}\epsilon\| \leq 1.5\sqrt{\frac{n}{p}}\|A^\top(AA^\top)^{-1}\epsilon\|$ with probability at least $1-2e^{-c_2 n}$ each. By Lemma F.9, $\|A^\top(AA^\top)^{-1}\epsilon\| \leq \frac{\sqrt{n}\|\theta\|}{\sqrt{p\gamma}(\sqrt{p-d}-2\sqrt{n})}$ (failure probability already accounted for on Event E). This gives the following bound with probability $1-4e^{-c_2 n}$:

$$II \leq 21\sqrt{n\gamma}\|\theta\|\|\beta\| + 21\frac{n}{\sqrt{p}(\sqrt{p-d}-2\sqrt{n})}\|\theta\|^2 \tag{97}$$

Substituting $\|\beta\| = \frac{\sqrt{p\gamma}}{p\gamma+1}\|\theta\|$ gives

$$II \leq 21\frac{\gamma\sqrt{np}}{p\gamma+1}\|\theta\|^2 + 21\frac{n}{\sqrt{p}(\sqrt{p-d}-2\sqrt{n})}\|\theta\|^2 \tag{98}$$

Using $p\gamma + 1 > p\gamma$ gives

$$II \leq 21\frac{\sqrt{n}}{\sqrt{p}}\|\theta\|^2 + 21\frac{n}{\sqrt{p}(\sqrt{p-d}-2\sqrt{n})}\|\theta\|^2 \tag{99}$$

Lastly we bound term III. Using $p\gamma + 1 > p\gamma$ gives the following bound:

$$III \leq \frac{12}{p\gamma}\|\theta\|^2 \tag{100}$$

Putting all three terms together gives the following bound with probability $1-10e^{-cd}$ where $c = min(c_1, c_2)$:

$$R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) \leq 12\frac{\sqrt{d}}{\sqrt{p}}\|\theta\|^2 + 12\frac{\sqrt{nd}}{\sqrt{p}(\sqrt{p-d}-2\sqrt{n})}\|\theta\|^2 + 21\frac{\sqrt{n}}{\sqrt{p}}\|\theta\|^2 \tag{101}$$

$$+ 21\frac{n}{\sqrt{p}(\sqrt{p-d}-2\sqrt{n})}\|\theta\|^2 + \frac{12}{p\gamma}\|\theta\|^2 \tag{102}$$

$$= \frac{12\sqrt{d}+21\sqrt{n}}{\sqrt{p}}\|\theta\|^2 + \frac{12\sqrt{nd}+21n}{\sqrt{p}(\sqrt{p-d}-2\sqrt{n})}\|\theta\|^2 + \frac{12}{p\gamma}\|\theta\|^2 \tag{103}$$

Using $d \leq n$ gives the following bound:

$$R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) \leq \left(\frac{33\sqrt{n}}{\sqrt{p}} + \frac{33n}{\sqrt{p}(\sqrt{p-n}-2\sqrt{n})} + \frac{12}{p\gamma}\right)\|\theta\|^2 \tag{104}$$

By assumption, $p \geq 17n$, so $\frac{n}{\sqrt{p}(\sqrt{p-n}-2\sqrt{n})} \leq \frac{\sqrt{n}}{\sqrt{p}}$, which gives the following bound:

$$R(f_{\hat{\beta}_{BA}}) - R(f_{\hat{\beta}_A}) \leq \left(\frac{66\sqrt{n}}{\sqrt{p}} + \frac{12}{p\gamma}\right)\|\theta\|^2 \tag{105}$$

□

Lemma F.12. *Let* $W \in \mathbb{R}^{p\times d}$. *Then*

$$(I + WW^\top)^{-1}W = W(W^\top W + I)^{-1} \tag{106}$$

Proof. Let $W = USV$ be the SVD of W where $U \in \mathbb{R}^{p\times d}, S \in \mathbb{R}^{d\times d}, V \in \mathbb{R}^{d\times d}$. Then we have

$$(I + WW^\top)^{-1}W = (I + USVV^\top SU^\top)^{-1}USV \tag{107}$$

$$= (I + US^2U^\top)^{-1}USV \tag{108}$$

Let $\tilde{U} \in \mathbb{R}^{p\times p}$ have the first d columns be U and the last $p-d$ columns be the rest of the orthonormal basis. Then we have

$$(I + WW^\top)^{-1}W = (\tilde{U}\tilde{U}^\top + US^2U^\top)^{-1}USV \tag{109}$$

$$= (\tilde{U}\tilde{U}^\top + \tilde{U}\begin{bmatrix} S^2 & 0 \\ 0 & 0 \end{bmatrix}\tilde{U}^\top)^{-1}USV \tag{110}$$

$$= \left(\tilde{U}\begin{bmatrix} I+S^2 & 0 \\ 0 & I \end{bmatrix}\tilde{U}^\top\right)^{-1}USV \tag{111}$$

$$= \tilde{U}\left(\begin{bmatrix} I+S^2 & 0 \\ 0 & I \end{bmatrix}\right)^{-1}\tilde{U}^\top USV \tag{112}$$

$$= \tilde{U}\left(\begin{bmatrix} I+S^2 & 0 \\ 0 & I \end{bmatrix}\right)^{-1}(I_d, 0)^\top SV \tag{113}$$

$$= \tilde{U}((I+S^2)^{-1}, 0)^\top SV \tag{114}$$

$$= U(I+S^2)^{-1}SV \tag{115}$$

$$= US(I+S^2)^{-1}V \tag{116}$$

$$= USVV^\top(I+S^2)^{-1}V \tag{117}$$

$$= W(I+W^\top W)^{-1} \tag{118}$$

□

G SOCIETAL IMPACT

We are interested in studying theory so that we can better understand existing methods in practice. This allows for a more thorough understanding of the limitations and failure modes for computational systems that learn throughout their lifetimes. Thus we believe that theoretical works of artificial intelligence are positive for society and help in the goal of developing safer and more ethical technology.