How Data Augmentation affects Optimization for Linear Regression

Boris Hanin[*]
Department of Operations Research
and Financial Engineering
Princeton University
bhanin@princeton.edu

Yi Sun[*]
Department of Statistics
University of Chicago
yisun@statistics.uchicago.edu

Abstract

Though data augmentation has rapidly emerged as a key tool for optimization in modern machine learning, a clear picture of how augmentation schedules affect optimization and interact with optimization hyperparameters such as learning rate is nascent. In the spirit of classical convex optimization and recent work on implicit bias, the present work analyzes the effect of augmentation on optimization in the simple convex setting of linear regression with MSE loss.

We find joint schedules for learning rate and data augmentation scheme under which augmented gradient descent provably converges and characterize the resulting minimum. Our results apply to arbitrary augmentation schemes, revealing complex interactions between learning rates and augmentations even in the convex setting. Our approach interprets augmented (S)GD as a stochastic optimization method for a time-varying sequence of proxy losses. This gives a unified way to analyze learning rate, batch size, and augmentations ranging from additive noise to random projections. From this perspective, our results, which also give rates of convergence, can be viewed as Monro-Robbins type conditions for augmented (S)GD.

1 Introduction

Data augmentation, a popular set of techniques in which data is augmented (i.e. modified) at every optimization step, has become increasingly crucial to training models using gradient-based optimization. However, in modern overparametrized settings where there are many different minimizers of the training loss, the specific minimizer selected by training and the quality of the resulting model can be highly sensitive to choices of augmentation hyperparameters. As a result, practitioners use methods ranging from simple grid search to Bayesian optimization and reinforcement learning (8; 9; 17) to select and schedule augmentations by changing hyperparameters over the course of optimization. Such approaches, while effective, often require extensive compute and lack theoretical grounding.

These empirical practices stand in contrast to theoretical results from the implicit bias and stochastic optimization literature. The extensive recent literature on implicit bias (15; 29; 32) gives provable guarantees on which minimizer of the training loss is selected by GD and SGD in simple settings, but considers cases without complex scheduling. On the other hand, classical theorems in stochastic optimization, building on the Monro-Robbins theorem in (25), give provably optimal learning rate schedules for strongly convex objectives. However, neither line of work addresses the myriad augmentation and hyperparameter choices crucial to gradient-based training effective in practice.

The present work takes a step towards bridging this gap. We consider two main questions for a learning rate schedule and data augmentation policy:

[*] Equal contribution

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

1. When and at what rate will optimization converge?
2. Assuming optimization converges, what point does it converge to?

To isolate the effect *on optimization* of jointly scheduling learning rate and data augmentation schemes, we consider these questions in the simple convex model of linear regression with MSE loss:

$$\mathcal{L}(W;\mathcal{D}) = \frac{1}{N} \left\| Y - WX \right\|_F^2 . \tag{1.1}$$

In this setting, we analyze the effect of the data augmentation policy on the optimization trajectory W_t of augmented (stochastic) gradient descent[2], which follows the update equation

$$W_{t+1} = W_t - \eta_t \nabla_W \mathcal{L}(W;\mathcal{D}_t)\big|_{W=W_t}. \tag{1.2}$$

Here, the dataset $\mathcal{D} = (X, Y)$ contains N datapoints arranged into data matrices $X \in \mathbb{R}^{n\times N}$ and $Y \in \mathbb{R}^{p\times N}$ whose columns consist of inputs $x_i \in \mathbb{R}^n$ and outputs $y_i \in \mathbb{R}^p$. In this context, we take a flexible definition *data augmentation scheme* as any procedure that consists, at every step of optimization, of replacing the dataset $\mathcal{D}$ by a randomly augmented variant which we denote by $\mathcal{D}_t = (X_t, Y_t)$. This framework is flexible enough to handle SGD and commonly used augmentations such as additive noise (14), CutOut (12), SpecAugment (23), Mixup (35), and label-preserving transformations (e.g. color jitter, geometric transformations (26))).

We give a general answer to Questions 1 and 2 for arbitrary data augmentation schemes. Our main result (Theorem 3.1) gives sufficient conditions for optimization to converge in terms of the learning rate schedule and simple 2nd and 4th order moment statistics of augmented data matrices. When convergence occurs, we explicitly characterize the resulting optimum in terms of these statistics. We then specialize our results to (S)GD with modern augmentations such as additive noise (14) and random projections (e.g. CutOut (12) and SpecAugment (23)). In these cases, we find learning rate and augmentation parameters which ensure convergence with rates to the minimum norm optimum for overparametrized linear regression. To sum up, our main contributions are:

1. We analyze *arbitrary* data augmentation schemes for linear regression with MSE loss, obtaining explicit sufficient conditions on the joint schedule of the data augmentation policy and the learning rate for (stochastic) gradient descent that guarantee convergence with rates in Theorems 3.1 and 3.2. The resulting augmentation-dependent optimum encodes the ultimate effect of augmentation on optimization, and we characterize it in Theorem 3.1. Our results generalize Monro-Robbins theorems (25) to situations where the stochastic optimization objective may change at each step.
2. We specialize our results to (stochastic) gradient descent with additive input noise (§4) or random projections of the input (§5), a proxy for the popular CutOut and SpecAugment augmentations (12; 23). In each case, we find that jointly scheduling learning rate and augmentation strength is critical for allowing convergence with rates to the minimum norm optimizer. We find specific schedule choices which guarantee this convergence with rates (Theorems 4.1, 4.2, and 5.1) and validate our results empirically (Figure 4.1). This suggests explicitly adding learning rate schedules to the search space for learned augmentations as in (8; 9), which we leave to future work.

2 Related Work

In addition to the extensive empirical work on data augmentation cited elsewhere in this article, we briefly catalog other theoretical work on data augmentation and learning rate schedules. The latter were first considered in the seminal work (25). This spawned a vast literature on *rates* of convergence for GD, SGD, and their variants. We mention only the relatively recent articles (1; 11; 4; 27; 22) and the references therein. The last of these, namely (22), finds optimal choices of learning rate and batch size for SGD in the overparametrized linear setting.

A number of articles have also pointed out in various regimes that data augmentation and more general transformations such as feature dropout correspond in part to ℓ_2-type regularization on model parameters, features, gradients, and Hessians. The first article of this kind of which we are aware is (3),

[2] Both GD and SGD fall into our framework. To implement SGD, we take $\mathcal{D}_t$ to be a subset of $\mathcal{D}$.

which treats the case of additive Gaussian noise (see §4). More recent work in this direction includes (5; 30; 19; 21). There are also several articles investigating *optimal* choices of ℓ_2-regularization for linear models (cf e.g. (33; 31; 2)). These articles focus directly on the generalization effects of ridge-regularized minima but not on the dynamics of optimization. We also point the reader to (20), which considers optimal choices for the weight decay coefficient empirically in neural networks and analytically in simple models.

We also refer the reader to a number of recent attempts to characterize the benefits of data augmentation. In (24), for example, the authors quantify how much augmented data, produced via additive noise, is needed to learn positive margin classifiers. (6), in contrast, focuses on the case of data invariant under the action of a group. Using the group action to generate label-preserving augmentations, the authors prove that the variance of any function depending only on the trained model will decrease. This applies in particular to estimators for the trainable parameters themselves. (10) shows augmented k-NN classification reduces to a kernel method for augmentations transforming each datapoint to a finite orbit of possibilities. It also gives a second order expansion for the proxy loss of a kernel method under such augmentations and interprets how each term affects generalization. Finally, the article (34) considers both label preserving and noising augmentations, pointing out the conceptually distinct roles such augmentations play.

3 Time-varying Monro-Robbins for linear models under augmentation

We seek to isolate the impact of data augmentation on optimization in the simple setting of augmented (stochastic) gradient descent for linear regression with the MSE loss (1.1). Since the augmented dataset $\mathcal{D}_t$ at time t is a stochastic function of $\mathcal{D}$, we may view the update rule (1.2) as a form of stochastic optimization for the *proxy loss at time* t

$$\overline{\mathcal{L}}_t(W) := \mathbb{E}_{\mathcal{D}_t}\left[\mathcal{L}(W; \mathcal{D}_t)\right] \tag{3.1}$$

which uses an unbiased estimate of the gradient of $\mathcal{L}(W; \mathcal{D}_t)$ from a single draw of $\mathcal{D}_t$. The connection between data augmentation and this proxy loss was introduced in (3; 5), but we now consider it in the context of stochastic optimization. In particular, we consider scheduling the augmentation, which allows the distribution of $\mathcal{D}_t$ to change with t and thus enables optimization to converge to points which are not minimizers of the proxy loss $\overline{\mathcal{L}}_t(W)$ at any fixed time.

Our main results, Theorems 3.1 and 3.2, provide sufficient conditions for jointly scheduling learning rates and general augmentation schemes to guarantee convergence of augmented gradient descent in the linear regression model (1.1). Before stating them, we first give examples of augmentations falling into our framework, which we analyze using our general results in §4 and §5.

- **Additive Gaussian noise:** For SGD with batch size B_t and noise level $\sigma_t > 0$, this corresponds to $X_t = c_t(XA_t + \sigma_t \cdot G_t)$ and $Y_t = c_t Y A_t$, where G_t is a matrix of i.i.d. standard Gaussians, $A_t \in \mathbb{R}^{N \times B_t}$ has i.i.d. columns with a single non-zero entry equal to 1 chosen uniformly at random and $c_t = \sqrt{N/B_t}$ is a normalizing factor. The proxy loss is

$$\overline{\mathcal{L}}_t(W) = \mathcal{L}(W; \mathcal{D}) + \sigma_t^2 \left\| W \right\|_F^2, \tag{3.2}$$

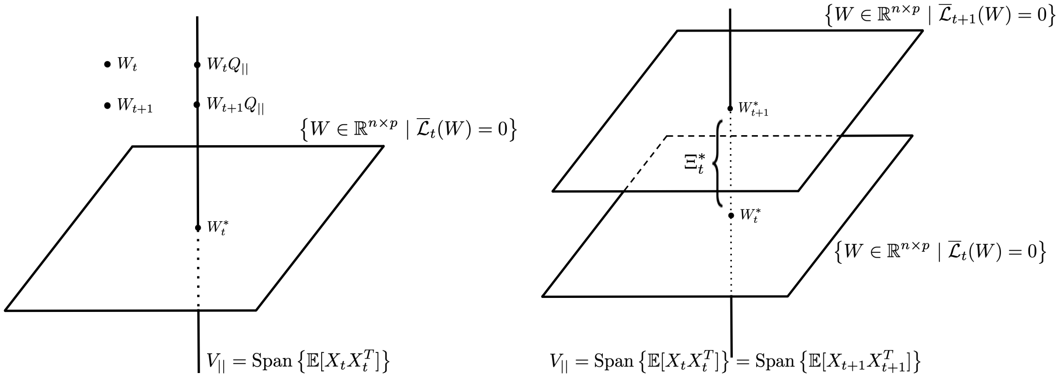
  which adds an ℓ_2 penalty. We analyze this case in §4.

- **Random projection:** This corresponds to $X_t = \Pi_t X$ and $Y_t = Y$, where Π_t is an orthogonal projection onto a random subspace. For $\gamma_t = \mathrm{tr}(\Pi_t)/n$, the proxy loss is

$$\overline{\mathcal{L}}_t(W) = \frac{1}{N}\|Y - \gamma_t W X\|_F^2 + \frac{1}{N}\gamma_t(1-\gamma_t)\frac{1}{n}\|X\|_F^2 \cdot \|W\|_F^2 + O(n^{-1}),$$

  adding a data-dependent ℓ_2 penalty and applying Stein-type shrinkage on input data. We analyze this in §5.

In addition to these augmentations, the augmentations below also fit into our framework, and Theorems 3.1 and 3.2 apply. However, in these cases, explicitly characterizing the learned minimum beyond the general description given in Theorems 3.1 and 3.2 is more difficult, and we thus leave interpretion of these specializations to future work.


$\bullet\, W_t$
$W_t Q_{||}$
$\bullet\, W_{t+1}$
$W_{t+1} Q_{||}$
$\{W \in \mathbb{R}^{n\times p} \mid \overline{\mathcal{L}}_t(W) = 0\}$
W_t^*
$V_{||} = \mathrm{Span}\,\{\mathbb{E}[X_t X_t^T]\}$
$\{W \in \mathbb{R}^{n\times p} \mid \overline{\mathcal{L}}_{t+1}(W) = 0\}$
W_{t+1}^*
Ξ_t^*
W_t^*
$\{W \in \mathbb{R}^{n\times p} \mid \overline{\mathcal{L}}_t(W) = 0\}$
$V_{||} = \mathrm{Span}\,\{\mathbb{E}[X_t X_t^T]\} = \mathrm{Span}\,\{\mathbb{E}[X_{t+1} X_{t+1}^T]\}$

(a) The time t proxy loss $\overline{\mathcal{L}}_t$ is non-degenerate on $V_{||}$ with minimal norm minimizer W_t^*. The increment $W_{t+1} - W_t$ is in $V_{||}$, so only the projection $W_t Q_{||}$ of W_t onto $V_{||}$ changes.

(b) The proxy losses $\overline{\mathcal{L}}_t$ and $\overline{\mathcal{L}}_{t+1}$ at consecutive times will generally have different minimal norm minimizers W_t^*, W_{t+1}^*. The increment $\Xi_t^* = W_{t+1}^* - W_t^*$ measures this change.

Figure 3.1: Schematic diagrams of augmented optimization in the parameter space $\mathbb{R}^{n\times p}$.

- **Label-preserving transformations:** For a 2-D image viewed as a vector $x \in \mathbb{R}^n$, geometric transforms (with pixel interpolation) or other label-preserving transforms such as color jitter take the form of linear transforms $\mathbb{R}^n \to \mathbb{R}^n$. We may implement such augmentations in our framework by $X_t = A_t X$ and $Y_t = Y$ for some random transform matrix A_t.
- **Mixup:** To implement Mixup, we can take $X_t = X A_t$ and $Y_t = Y A_t$, where $A_t \in \mathbb{R}^{N\times B_t}$ has i.i.d. columns containing two random non-zero entries equal to $1 - c_t$ and c_t with mixing coefficient c_t drawn independently from a Beta(α_t, α_t) distribution for a parameter α_t.

3.1 A general time-varying Monro-Robbins theorem

Given an augmentation scheme for the overparameterized linear model (1.1), the time t gradient update at learning rate η_t is

$$W_{t+1} := W_t + \frac{2\eta_t}{N} \cdot (Y_t - W_t X_t) X_t^{\mathsf{T}}, \tag{3.3}$$

where $\mathcal{D}_t = (X_t, Y_t)$ is the augmented dataset at time t. The minimum norm minimizer of the corresponding proxy loss $\overline{\mathcal{L}}_t$ (see (3.1)) is

$$W_t^* := \mathbb{E}[Y_t X_t^{\mathsf{T}}]\mathbb{E}[X_t X_t^{\mathsf{T}}]^+, \tag{3.4}$$

where $\mathbb{E}[X_t X_t^{\mathsf{T}}]^+$ denotes the Moore-Penrose pseudo-inverse (see Figure 3.1a). In this section we state a rigorous result, Theorem 3.1, giving sufficient conditions on the learning rate η_t and distributions of the augmented matrices X_t, Y_t under which augmented gradient descent converges. To state it, note that (3.3) implies that each row of $W_{t+1} - W_t$ is contained in the column span of the Hessian $X_t X_t^{\mathsf{T}}$ of the augmented loss and therefore almost surely belongs to the subspace

$$V_{||} := \text{column span of } \mathbb{E}[X_t X_t^{\mathsf{T}}] \subseteq \mathbb{R}^n, \tag{3.5}$$

as illustrated in Figure 3.1a. The reason is that, in the orthogonal complement to $V_{||}$, the augmented loss $\mathcal{L}(W; \mathcal{D}_t)$ has zero gradient with probability 1. To ease notation, we assume that $V_{||}$ is independent of t. This assumption holds for additive Gaussian noise, random projection, MixUp, SGD, and their combinations. It is not necessary in general, however, and we refer the interested reader to Remark B.2 in the Appendix for how to treat the general case.

Let us denote by $Q_{||} : \mathbb{R}^n \to \mathbb{R}^n$ the orthogonal projection onto $V_{||}$ (see Figure 3.1a). As we already pointed out, at time t, gradient descent leaves the matrix of projections $W_t(\mathrm{Id} - Q_{||})$ of each row of W_t onto the orthogonal complement of $V_{||}$ unchanged. In contrast, $\|W_t Q_{||} - W_t^*\|_F$ decreases at a rate governed by the smallest positive eigenvalue $\lambda_{\min, V_{||}}\left(\mathbb{E}\left[X_t X_t^{\mathsf{T}}\right]\right)$ of the Hessian for the proxy loss $\overline{\mathcal{L}}_t$, which is obtained by restricting its full Hessian $\mathbb{E}\left[X_t X_t^\mathsf{T}\right]$ to $V_\parallel$. Moreover, whether and at what rate $W_t Q_\parallel - W_t^*$ converges to 0 depends on how quickly the distance

$$\Xi_t^* := W_{t+1}^* - W_t^* \tag{3.6}$$

between proxy loss optima at successive steps tends to zero (see Figure 3.1b).

Theorem 3.1 (Special case of Theorem B.1)**.** *Suppose that $V_\parallel$ is independent of t, that the learning rate satisfies $\eta_t \to 0$, that the proxy optima satisfy*

$$\sum_{t=0}^{\infty} \|\Xi_t^*\|_F < \infty, \tag{3.7}$$

ensuring the existence of a limit $W_\infty^* := \lim_{t\to\infty} W_t^*$, that

$$\sum_{t=0}^{\infty} \eta_t \lambda_{min, V_\parallel}(\mathbb{E}[X_t X_t^\mathsf{T}]) = \infty \tag{3.8}$$

and finally that

$$\sum_{t=0}^{\infty} \eta_t^2 \mathbb{E}\Big[\|X_t X_t^\mathsf{T} - \mathbb{E}[X_t X_t^\mathsf{T}]\|_F^2 + \|Y_t X_t^\mathsf{T} - \mathbb{E}[Y_t X_t^\mathsf{T}]\|_F^2\Big] < \infty. \tag{3.9}$$

Then, for any initialization W_0, we have that $W_t Q_\parallel$ converges in probability to W_∞^*.

If the same augmentation is applied with different strength parameters at each step t (e.g. the noise level σ_t^2 for additive Gaussian noise), we may specialize Theorem 3.1 to this augmentation scheme. More precisely, translating conditions (3.7), (3.8), (3.9) into conditions on the learning rate and augmentation strength gives conditions on the schedule for η_t and these strength parameters to ensure convergence. In §4 and §5, we do this for additive Gaussian noise and random projections.

When the augmentation scheme is static in t, Theorem 3.1 reduces to a standard Monro-Robbins theorem (25) for the (static) proxy loss $\overline{\mathcal{L}}_t(W)$. As in that setting, condition (3.8) enforces that the learning trajectory travels far enough to reach an optimum, and the summand in Condition (3.9) bounds the variance of the gradient of the augmented loss $\mathcal{L}(W; \mathcal{D}_t)$ to ensure the total variance of the stochastic gradients is summable. Condition (3.7) is new and enforces that the minimizers W_t^* of the proxy losses $\overline{\mathcal{L}}_t(W)$ change slowly enough for augmented optimization procedure to keep pace.

3.2 Convergence rates and scheduling for data augmentation

Refining the proof of Theorem 3.1 gives rates of convergence for the projections $W_t Q_\parallel$ of the weights onto $V_\parallel$ to the limiting optimum W_∞^*. When the quantities in Theorem 3.1 have power law decay, we obtain the following result.

Theorem 3.2 (Special case of Theorem B.4)**.** *Suppose $V_\parallel$ is independent of t and the learning rate satisfies $\eta_t \to 0$. Moreover assume that for some $0 < \alpha < 1 < \beta_1, \beta_2$ and $\gamma > \alpha$ we have*

$$\eta_t \lambda_{min, V_\parallel}(\mathbb{E}[X_t X_t^\mathsf{T}]) = \Omega(t^{-\alpha}), \qquad \|\Xi_t^*\|_F = O(t^{-\beta_1}) \tag{3.10}$$

as well as

$$\eta_t^2 \mathbb{E}[\|X_t X_t^\mathsf{T} - \mathbb{E}[X_t X_t^\mathsf{T}]\|_2^2] = O(t^{-\gamma}) \tag{3.11}$$

and

$$\eta_t^2 \mathbb{E}\Big[\|\mathbb{E}[W_t](X_t X_t^\mathsf{T} - \mathbb{E}[X_t X_t^\mathsf{T}]) - (Y_t X_t^\mathsf{T} - \mathbb{E}[Y_t X_t^\mathsf{T}])\|_F^2\Big] = O(t^{-\beta_2}). \tag{3.12}$$

Then, for any initialization W_0, we have for any $\varepsilon > 0$ the following convergence in probability:

$$t^{\min\{\beta_1 - 1, \frac{\beta_2 - \alpha}{2}\} - \varepsilon} \|W_t Q_\parallel - W_\infty^*\|_F \xrightarrow{p} 0.$$

It may be surprising that $\mathbb{E}[W_t]$ appears in condition (3.12). Note that $\mathbb{E}[W_t]$ is the gradient descent trajectory for the time-varying sequence of deterministic proxy losses $\overline{\mathcal{L}}_t(W)$. To apply Theorem 3.2, one may first study this deterministic problem to show that $\mathbb{E}[W_t]$ converges to W_∞^* at some rate and then use (3.12) to obtain rates of convergence of the true stochastic trajectory W_t to W_∞^*.

In §4 and §5 below, we specialize Theorems 3.1 and 3.2 to obtain rates of convergence for specific augmentations. Optimizing the learning rate and augmentation parameter schedules in Theorem 3.2 yields power law schedules with convergence rate guarantees in these settings.

4 Special Case: Additive Gaussian Noise

We now specialize our main results Theorem 3.1 and 3.2 to the commonly used additive noise augmentation (14). Under gradient descent, this corresponds to taking augmented data matrices

$$X_t = X + \sigma_t G_t \qquad \text{and} \qquad Y_t = Y,$$

where $G_t \in \mathbb{R}^{n\times N}$ are independent matrices with i.i.d. standard Gaussian entries, and $\sigma_t > 0$ is a strength parameter. Under SGD (with replacement), this corresponds to augmented data matrices

$$X_t = c_t(XA_t + \sigma_t \cdot G_t) \qquad \text{and} \qquad Y_t = c_t Y A_t,$$

where $A_t \in \mathbb{R}^{N\times B_t}$ has i.i.d. columns with a single non-zero entry equal to 1 chosen uniformly at random and $c_t = \sqrt{N/B_t}$ is a normalizing factor. In both cases, the proxy loss is

$$\mathcal{L}_{\sigma_t}(W) := \frac{1}{N}\|Y - WX\|_F^2 + \sigma_t^2\|W\|_F^2, \tag{4.1}$$

which to our knowledge was first discovered in (3).

Before stating our precise results, we first illustrate how jointly scheduling learning rate and augmentation strength impacts GD for overparameterized linear regression, where

$$N \;=\; \#\text{data points} \;<\; \text{input dimension} \;=\; n. \tag{4.2}$$

The inequality (4.2) ensures $\mathcal{L}(W;\mathcal{D})$ has infinitely many minima, of which we consider the *minimum norm* minimizer

$$W_{\min} := YX^{\mathsf{T}}(XX^{\mathsf{T}})^{+}$$

most desirable. Notice that steps of vanilla gradient descent

$$W_{t+1} = W_t - \frac{2\eta_t}{N}\cdot(Y - W_tX)X^{\mathsf{T}} \tag{4.3}$$

change the rows of the weight matrix W_t only in the column space $V_{\|} = \operatorname{colspan}(XX^{\mathsf{T}}) \subseteq \mathbb{R}^n$. Because $V_{\|} \neq \mathbb{R}^n$ by the overparameterization assumption (4.2), minimizing $\mathcal{L}(W;\mathcal{D})$ *without augmentation* cannot change $W_{t,\perp}$, the matrix whose rows are the components of the rows of W_t orthogonal to $V_{\|}$. This means that GD converges to the minimal norm optimizer $W_{\min}$ only when each row of W_0 belongs to $V_{\|}$. As this explicit initialization may not be available for more general models, we seek to find augmentation schedules which allow GD or SGD to converge to $W_{\min}$ without it, in the spirit of recent studies on implicit bias of GD.

4.1 Joint schedules for augmented GD with additive noise to converge to $W_{\min}$

We specialize Theorems 3.1 and 3.2 to additive Gaussian noise to show that when the learning rate η_t and noise strength σ_t are jointly scheduled to converge to 0 at appropriate rates, augmented gradient descent can find the minimum norm optimizer $W_{\min}$.

Theorem 4.1. *Consider any joint schedule of the learning rate η_t and noise variance σ_t^2 in which both η_t and σ_t^2 tend to 0 and σ_t is non-decreasing. If the joint schedule satisfies*

$$\sum_{t=0}^{\infty}\eta_t\sigma_t^2 = \infty \qquad \textit{and} \qquad \sum_{t=0}^{\infty}\eta_t^2\sigma_t^2 < \infty, \tag{4.4}$$

then the weights W_t converge in probability to the minimal norm optimum W_{min} regardless of the initialization. Moreover, the first condition in (4.4) *is necessary for* $\mathbb{E}[W_t]$ *to converge to* $W_{\min}$.

If we further have $\eta_t = \Theta(t^{-x})$ and $\sigma_t^2 = \Theta(t^{-y})$ with $x, y > 0$, $x + y < 1$, and $2x + y > 1$ so that η_t and σ_t^2 satisfy (4.4), then for small $\varepsilon > 0$, we have $t^{\min\{y,\frac{1}{2}x\}-\varepsilon}\|W_t - W_{min}\|_F \xrightarrow{p} 0$.

The conditions of (4.4) require that η_t and σ_t be jointly scheduled correctly to ensure convergence to $W_{\min}$ and are akin to the Monro-Robbins conditions (25) for convergence of stochastic gradient methods. We now give an heuristic explanation for why the first condition from (4.4) is necessary. In this setting, the *average* trajectory of augmented gradient descent

$$\mathbb{E}[W_{t+1}] = \mathbb{E}[W_t] - \eta\nabla_W\mathcal{L}_{\sigma_t}(W)\big|_{W=\mathbb{E}[W_t]} \tag{4.5}$$

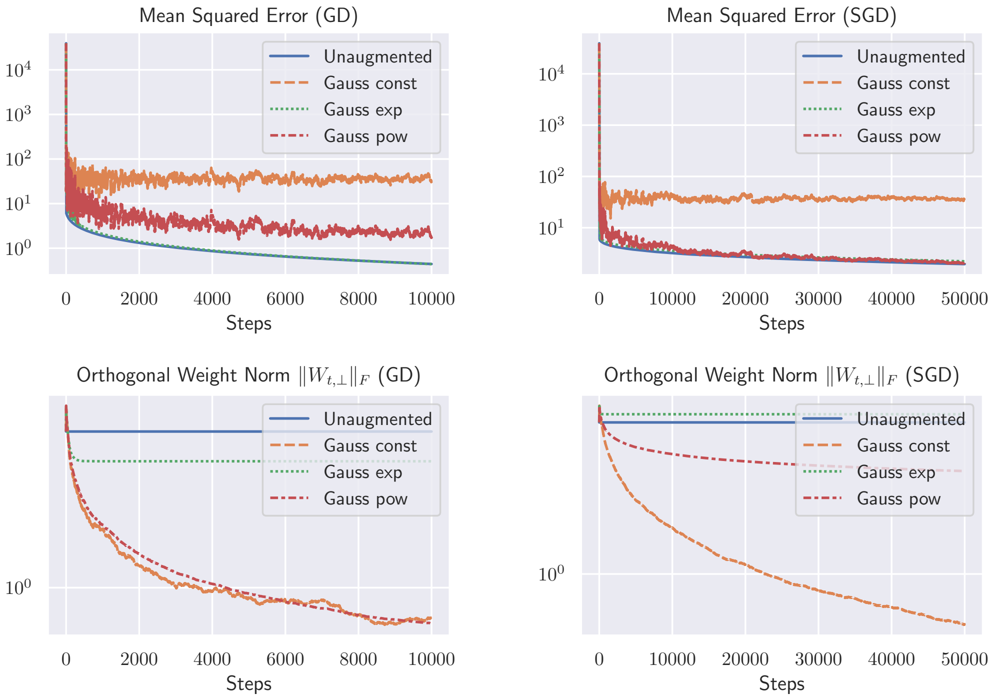


Figure 4.1: MSE and $\|W_{t,\perp}\|_F$ for optimization trajectories of GD with additive Gaussian noise augmentation and SGD with additive Gaussian noise augmentation under different augmentation schedules. For both GD and SGD, jointly scheduling learning rate and noise variance to have polynomial decay is necessary for optimization to converge to the minimal norm solution $W_{\min}$. Gauss const, Gauss exp, and Gauss pow have Gaussian noise augmentation with $\sigma_t^2 = 2, 2e^{-0.02t}, 2(1+\frac{t}{50})^{-0.33}$, respectively. All other details are given in §4.3.

is given by gradient descent on the ridge regularized losses $\mathcal{L}_{\sigma_t}(W)$. If $\sigma_t \equiv \sigma > 0$ is constant, then $\mathbb{E}[W_t]$ will converge to the unique minimizer W_σ^* of the ridge regularized loss $\mathcal{L}_\sigma$. This point W_σ^* has zero orthogonal component, but *does not* minimize the original loss $\mathcal{L}$.

To instead minimize $\mathcal{L}$, we must choose a schedule satisfying $\sigma_t \to 0$. For the expected optimization trajectory to converge to $W_{\min}$ for such a schedule, the matrix $\mathbb{E}[W_{t,\perp}]$ of components of rows of $\mathbb{E}\left[W_t\right]$ orthogonal to $V_{\|}$ must converge to 0. The GD steps for this matrix yield

$$\mathbb{E}[W_{t+1,\perp}] = (1 - \eta_t \sigma_t^2)\mathbb{E}[W_{t,\perp}] = \prod_{s=0}^{t}(1 - \eta_s \sigma_s^2)\mathbb{E}[W_{0,\perp}]. \tag{4.6}$$

Because $\eta_t \sigma_t^2$ approaches 0, this implies the necessary condition $\sum_{t=0}^{\infty} \eta_t \sigma_t^2 = \infty$ for $\mathbb{E}[W_{t,\perp}] \to 0$.

This argument illustrates a key intuition behind the conditions (4.4). The augmentation strength σ_t must decay to 0 to allow convergence to a true minimizer of the training loss, but this convergence must be carefully tuned to allow the implicit regularization of the noising augmentation to kill the orthogonal component of W_t in expectation. In a similar manner, the second expression in (4.4) measures the total variance of the gradients and ensures that only a finite amount of noise is injected into the optimization.

Although Theorem 4.1 is stated for additive Gaussian noise, an analogous version holds for arbitrary additive noise with bounded moments. Moreover, optimizing over x, y, the fastest rate of convergence guaranteed by Theorem 4.1 is obtained by setting $\eta_t = t^{-2/3+\varepsilon}$, $\sigma_t^2 = t^{-1/3}$ and results in a $O(t^{-1/3+\varepsilon})$ rate of convergence. It is not evident that this is best possible, however.

4.2 Joint schedules for augmented SGD with additive noise to converge to $W_{\min}$

To conclude our discussion of additive noise augmentations, we present the following result on convergence to $W_{\min}$ in the presence of both Gaussian noise and SGD (where datapoints in each batch are selected with replacement).

Theorem 4.2. *Suppose $\sigma_t^2 \to 0$ is decreasing, $\eta_t \to 0$, and we have*

$$\sum_{t=0}^{\infty} \eta_t \sigma_t^2 = \infty \qquad \textit{and} \qquad \sum_{t=0}^{\infty} \eta_t^2 < \infty. \tag{4.7}$$

Then the trajectory W_t of SGD with additive noise converges in probability to W_{min} for any initialization. If we further have $\eta_t = \Theta(t^{-x})$ and $\sigma_t^2 = \Theta(t^{-y})$ with $x > \frac{1}{2}$, $y > 0$ and $x + y < 1$, then for any $\varepsilon > 0$ we have that $t^{\min\{y, \frac{1}{2}x\} - \varepsilon} \|W_t - W_{min}\|_F \xrightarrow{p} 0$.

Theorem 4.2 is the analog of Theorem 4.1 for mini-batch SGD and provides an example where our framework can handle the *composition* of two augmentations, namely additive noise and mini-batching. The difference between conditions (4.7) for SGD and (4.4) for GD accounts for the fact that the batch selection changes the scale of the gradient variance at each step. Finally, Theorem 4.2 reveals a qualitative difference between SGD with and without additive noise. If η_t has power law decay, the convergence of noiseless SGD (Theorem F.1) is exponential in t, while Theorem 4.2 gives power law rates.

4.3 Experimental validation

To validate Theorems 4.1 and 4.2, we ran augmented GD and SGD with additive Gaussian noise on $N = 100$ simulated datapoints. Inputs were i.i.d. Gaussian vectors in dimension $n = 400$, and outputs in dim $p = 1$ were generated by a random linear map with i.i.d Gaussian coefficients drawn from $\mathcal{N}(1, 1)$. The learning rate followed a fixed polynomially decaying schedule $\eta_t = \frac{0.005}{100} \cdot (\text{batch size}) \cdot (1 + \frac{t}{20})^{-0.66}$, and the batch size used for SGD was 20. Figure 4.1 shows MSE and $\|W_{t,\perp}\|_F$ along a single optimization trajectory with different schedules for the variance σ_t^2 used in Gaussian noise augmentation. Complete code to generate this figure is provided in `supplement.zip` in the supplement. It ran in 30 minutes on a standard laptop CPU.

For both GD and SGD, Figure 4.1 shows that the optimization trajectory reaches $W_{\min}$ only when both learning rate and noise variance decay polynomially to zero. Indeed, Figure 4.1 shows that if σ_t^2 is zero (blue) or exponentially decaying (green), then while the MSE tends to zero, the orthogonal component $W_{t,\perp}$ does not tend to zero. Thus these choices of augmentation schedule cause W_t to converge to an optimum which does not have minimal norm.

On the other hand, if σ_t^2 remains constant (orange), then while $W_{t,\perp}$ tends to zero, the MSE is not minimized. Only by decaying both noise strength and learning rate to 0 at sufficiently slow polynomial rates (red) prescribed by Theorem 4.1 do we find both MSE and $W_{t,\perp}$ tending to 0, meaning that augmented (S)GD finds the minimum norm optimum $W_{\min}$ under this choice of parameter scheduling.

5 Special Case: Augmentation with Random Projections

We further illustrate our results by specializing them to a class of augmentations which replace each input x in a batch by its orthogonal projection $\Pi_t X$ onto a random subspace. In practice (e.g. when using CutOut (12) or SpecAugment (23)), the subspace is chosen based on a prior about correlations between components of X, but we consider the simplified case of a uniformly random subspace of $\mathbb{R}^n$ of given dimension.

At each time step t we fix a dimension k_t and a fixed k_t-dimensional subspace $\widetilde{S}_t$ of $\mathbb{R}^n$. Define the random subspace S_t by

$$S_t := Q_t(\widetilde{S}_t) = \{Q_t x \mid x \in \widetilde{S}_t\},$$

where $Q_t \in O(n)$ is a Haar random orthogonal matrix. Thus, S_t is uniformly distributed among all k_t-dimensional subspaces in $\mathbb{R}^n$. At step t, we take the augmentation given by

$$X_t = \Pi_t X \qquad Y_t = Y, \qquad \Pi_t := Q_t \widetilde{\Pi}_t Q_t^{\mathsf{T}},$$

where $\widetilde{\Pi}_t$ is the orthogonal projection onto $\widetilde{S}_t$ and hence Π_t is the orthogonal projection onto S_t.

Denoting by $\gamma_t = k_t/n$ the relative dimension of S_t, a direct computation (see Lemma E.1) reveals that the proxy loss $\overline{\mathcal{L}}_t(W)$ equals $\mathcal{L}(\gamma_t W; \mathcal{D})$ plus

$$\frac{1}{N}\frac{\gamma_t(1-\gamma_t)}{n}\|X\|_F^2 \cdot \|W\|_F^2 + \frac{\gamma_t(1-\gamma_t)(1/n - 2/n^2)}{N(1+1/n-2/n^2)}(\|WX\|_F^2 + \frac{1}{n}\|X\|_F^2 \cdot \|W\|_F^2). \quad (5.1)$$

Neglecting terms of order $O(n^{-1})$, this proxy loss applies a Stein-type shrinkage on input data by γ_t and adds a data-dependent ℓ_2 penalty. For $\gamma_t < 1$, the minimizer of the proxy loss (5.1) is

$$W^*_{\gamma_t} = YX^\mathsf{T}\Big(\frac{\gamma_t + 1/n - 2/n^2}{1+1/n-2/n^2}XX^\mathsf{T} + \frac{1-\gamma_t}{1+1/n-2/n^2}\frac{\|X\|_F^2}{n}\,\mathrm{Id}\Big)^{-1}.$$

Again, although $W^*_{\gamma_t}$ does not minimize the original objective for any $\gamma_t < 1$, the sequence of these proxy optima converges to the minimal norm optimum in the weak regularization limit. Namely, we have $\lim_{\gamma_t \to 1^-} W^*_{\gamma_t} = W_{\min}$. Specializing our general result Theorem 3.1 to this setting, we obtain explicit conditions under which joint schedules of the normalized rank of the projection and the learning rate guarantee convergence to the minimum norm optimizer $W_{\min}$.

Theorem 5.1. *Suppose that $\eta_t \to 0, \gamma_t \to 1$ with γ_t non-decreasing and*

$$\sum_{t=0}^{\infty}\eta_t(1-\gamma_t) = \infty \text{ and } \sum_{t=0}^{\infty}\eta_t^2(1-\gamma_t) < \infty. \quad (5.2)$$

Then, $W_t \xrightarrow{p} W_{min}$. Further, if $\eta_t = \Theta(t^{-x})$ and $\gamma_t = 1 - \Theta(t^{-y})$ with $x, y > 0$, $x + y < 1$, and $2x + y > 1$, then for small $\varepsilon > 0$, we have that $t^{\min\{y, \frac{1}{2}x\}-\varepsilon}\|W_t - W_{\min}\|_F \xrightarrow{p} 0$.

Comparing the conditions (5.2) of Theorem 5.1 to the conditions (4.4) of Theorem 4.1, we see that $1 - \gamma_t$ is a measure of the strength of the random projection preconditioning. As in that setting, the fastest rates of convergence guaranteed by Theorem 5.1 are obtained by setting $\eta_t = t^{-2/3+\varepsilon}$ and $\gamma_t = 1 - t^{-1/3}$, yielding a $O(t^{-1/3+\varepsilon})$ rate of convergence.

6 Discussion and Limitations

We have presented a theoretical framework to rigorously analyze the effect of data augmentation. As can be seen in our main results, our framework applies to completely general augmentations and relies only on analyzing the first few moments of the augmented dataset. This allows us to handle augmentations as diverse as additive noise and random projections as well as their composition in a uniform manner. We have analyzed some representative examples in detail in this work, but many other commonly used augmentations may be handled similarly: label-preserving transformations (e.g. color jitter, geometric transformations) and Mixup (35), among many others. Another line of investigation left to future work is to compare different methods of combining augmentations such as mixing, alternating, or composing, which often improve performance in the empirical literature (16).

Though our results provide a rigorous baseline to compare to more complex settings, the restriction of the present work to linear models is a significant constraint. In future work, we hope to extend our general analysis to models closer to those used in practice. Most importantly, we intend to consider more complex models such as kernels (including the neural tangent kernel) and neural networks by making similar connections to stochastic optimization. In an orthogonal direction, our analysis currently focuses on the mean square loss for regression, and we aim to extend it to other losses such as cross-entropy. Finally, our study has thus far been restricted to the effect of data augmentation on optimization, and it would be of interest to derive consequences for generalization with more complex models. We hope our framework can provide the theoretical underpinnings for a more principled understanding of the effect and practice of data augmentation.

Broader Impact

Our work provides a new theoretical approach to data augmentation for neural networks. By giving a better understanding of how this common practice affects optimization, we hope that it can lead to more robust and interpretable uses of data augmentation in practice. Because our work is theoretical and generic, we do not envision negative impacts aside from those arising from improving learning algorithms in general.

Acknowledgments and Disclosure of Funding

It is a pleasure to thank Daniel Park, Ethan Dyer, Edgar Dobriban, and Pokey Rule for a number of insightful conversations about data augmentation. B.H. was partially supported by NSF grants DMS-1855684 and DMS-2133806 and ONR MURI "Theoretical Foundations of Deep Learning." Y. S. was partially supported by NSF grants DMS-1701654/2039183 and DMS-2054838.

References

[1] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In *Advances in neural information processing systems*, pages 773–781, 2013.

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

[3] C. M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

[4] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[5] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik. Vicinal risk minimization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 416–422. MIT Press, 2001.

[6] S. Chen, E. Dobriban, and J. H. Lee. Invariance reduces variance: Understanding data augmentation in deep learning and beyond. *stat*, 1050:25, 2019.

[7] B. Collins and P. Śniady. Integration with respect to the Haar measure on unitary, orthogonal and symplectic group. *Communications in Mathematical Physics*, 264(3):773–795, Mar 2006.

[8] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 113–123, 2019.

[9] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[10] T. Dao, A. Gu, A. Ratner, V. Smith, C. De Sa, and C. Re. A kernel theory of modern data augmentation. volume 97 of *Proceedings of Machine Learning Research*, pages 1528–1537, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[11] A. Défossez and F. Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Artificial Intelligence and Statistics*, pages 205–213, 2015.

[12] T. DeVries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[13] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch SGD: Training Imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[14] Y. Grandvalet and S. Canu. Noise injection for inputs relevance determination. In *Advances in intelligent systems*, pages 378–382. IOS Press, 1997.

[15] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry, 2020.

[16] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[17] D. Ho, E. Liang, X. Chen, I. Stoica, and P. Abbeel. Population based augmentation: Efficient learning of augmentation policy schedules. In *International Conference on Machine Learning*, pages 2731–2741, 2019.

[18] D. Huang, J. Niles-Weed, J. A. Tropp, and R. Ward. Matrix concentration for products, 2020.

[19] D. LeJeune, R. Balestriero, H. Javadi, and R. G. Baraniuk. Implicit rugosity regularization via data augmentation. *arXiv preprint arXiv:1905.11639*, 2019.

[20] A. Lewkowycz and G. Gur-Ari. On the training dynamics of deep networks with L_2 regularization. *arXiv preprint arXiv:2006.08643*, 2020.

[21] F. Liu, A. Najmi, and M. Sundararajan. The penalty imposed by ablated data augmentation. *arXiv preprint arXiv:2006.04769*, 2020.

[22] S. Ma, R. Bassily, and M. Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

[23] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. *Proc. Interspeech 2019*, pages 2613–2617, 2019.

[24] S. Rajput, Z. Feng, Z. Charles, P.-L. Loh, and D. Papailiopoulos. Does data augmentation lead to positive margin? *arXiv preprint arXiv:1905.03177*, 2019.

[25] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951.

[26] P. Simard, D. Steinkraus, and J. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, volume 2, pages 958–958, 2003.

[27] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le. Don't decay the learning rate, increase the batch size. In *International Conference on Learning Representations*, 2018.

[28] S. L. Smith and Q. V. Le. A Bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.

[29] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data, 2018.

[30] S. Wager, S. Wang, and P. S. Liang. Dropout training as adaptive regularization. In *Advances in neural information processing systems*, pages 351–359, 2013.

[31] D. Wu and J. Xu. On the optimal weighted L_2 regularization in overparameterized linear regression. *arXiv preprint arXiv:2006.05800*, 2020.

[32] J. Wu, D. Zou, V. Braverman, and Q. Gu. Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate, 2021.

[33] L. Wu, C. Ma, and E. Weinan. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, pages 8279–8288, 2018.

[34] S. Wu, H. R. Zhang, G. Valiant, and C. Ré. On the generalization effects of linear transformations in data augmentation, 2020.

[35] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

A Analytic lemmas

In this section, we present several basic lemmas concerning convergence for certain matrix-valued recursions that will be needed to establish our main results. For clarity, we first collect some matrix notations used in this section and throughout the paper.

A.1 Matrix notations

Let $M \in \mathbb{R}^{m\times n}$ be a matrix. We denote its Frobenius norm by $\|M\|_F$ and its spectral norm by $\|M\|_2$. If $m = n$ so that M is square, we denote by $\mathrm{diag}(M)$ the diagonal matrix with $\mathrm{diag}(M)_{ii} = M_{ii}$. For matrices A, B, C of the appropriate shapes, define

$$A \circ (B \otimes C) := BAC \tag{A.1}$$

and

$$\mathrm{Var}(A) := \mathbb{E}[A^{\mathsf{T}} \otimes A] - \mathbb{E}[A^{\mathsf{T}}] \otimes \mathbb{E}[A]. \tag{A.2}$$

Notice in particular that

$$\mathrm{tr}[\mathrm{Id} \circ \mathrm{Var}(A)] = \mathbb{E}[\|A - \mathbb{E}[A]\|_F^2].$$

A.2 One- and two-sided decay

Definition A.1. *Let $A_t \in \mathbb{R}^{n\times n}$ be a sequence of independent random non-negative definite matrices with*

$$\sup_t \|A_t\| \leq 2 \quad \textit{almost surely},$$

let $B_t \in \mathbb{R}^{p\times n}$ be a sequence of arbitrary matrices, and let $C_t \in \mathbb{R}^{n\times n}$ be a sequence of non-negative definite matrices. We say that the sequence of matrices $X_t \in \mathbb{R}^{p\times n}$ has one-sided decay of type $(\{A_t\}, \{B_t\})$ if it satisfies

$$X_{t+1} = X_t(\mathrm{Id} - \mathbb{E}[A_t]) + B_t. \tag{A.3}$$

We say that a sequence of non-negative definite matrices $Z_t \in \mathbb{R}^{n\times n}$ has two-sided decay of type $(\{A_t\}, \{C_t\})$ if it satisfies

$$Z_{t+1} = \mathbb{E}[(\mathrm{Id} - A_t)Z_t(\mathrm{Id} - A_t)] + C_t. \tag{A.4}$$

Intuitively, if a sequence of matrices X_t (resp. Z_t) satisfies one decay of type $(\{A_t\}, \{B_t\})$ (resp. two-sided decay of type $(\{A_t\}, \{C_t\})$), then in those directions $u \in \mathbb{R}^n$ for which $\|A_t u\|$ does not decay too quickly in t we expect that X_t (resp. Z_t) will converge to 0 provided B_t (resp. C_t) are not too large. More formally, let us define

$$V_{\|} := \bigcap_{t=0}^{\infty} \ker\left[\prod_{s=t}^{\infty}(\mathrm{Id} - \mathbb{E}[A_s])\right] = \left\{u \in \mathbb{R}^n \;\middle|\; \lim_{T\to\infty}\prod_{s=t}^{T}(\mathrm{Id} - \mathbb{E}[A_s])u = 0, \quad \forall t \geq 1\right\},$$

and let $Q_{\|}$ be the orthogonal projection onto $V_{\|}$. It is on the space $V_{\|}$ that that we expect X_t, Z_t to tend to zero if they satisfy one or two-side decay, and the precise results follows.

A.3 Lemmas on Convergence for Matrices with One and Two-Sided Decay

We state here several results that underpin the proofs of our main results. We begin by giving in Lemmas A.2 and A.3 two slight variations of the same simple argument that matrices with one or two-sided decay converge to zero.

Lemma A.2. *If a sequence $\{X_t\}$ has one-sided decay of type $(\{A_t\}, \{B_t\})$ with*

$$\sum_{t=0}^{\infty} \|B_t\|_F < \infty, \tag{A.5}$$

then $\lim_{t\to\infty} X_t Q_{\|} = 0$.

Proof. For any $\varepsilon > 0$, choose T_1 so that $\sum_{t=T_1}^{\infty} \|B_t\|_F < \frac{\varepsilon}{2}$ and T_2 so that for $t > T_2$ we have

$$\left\| \Big(\prod_{s=T_1}^{t} (\mathrm{Id} - \mathbb{E}[A_s]) \Big) Q_{\|} \right\|_2 < \frac{\varepsilon}{2} \frac{1}{\|X_0\|_F + \sum_{s=0}^{T_1-1} \|B_s\|_F}.$$

By (A.3), we find that

$$X_{t+1} = X_0 \prod_{s=0}^{t} (\mathrm{Id} - \mathbb{E}[A_s]) + \sum_{s=0}^{t} B_s \prod_{r=s+1}^{t} (\mathrm{Id} - \mathbb{E}[A_r]),$$

which implies for $t > T_2$ that

$$\|X_{t+1} Q_{\|}\|_F \leq \|X_0\|_F \left\| \Big(\prod_{s=0}^{t} (\mathrm{Id} - \mathbb{E}[A_s]) \Big) Q_{\|} \right\|_2 + \sum_{s=0}^{t} \|B_s\|_F \left\| \Big(\prod_{r=s+1}^{t} (\mathrm{Id} - \mathbb{E}[A_r]) \Big) Q_{\|} \right\|_2. \tag{A.6}$$

Our assumption that $\|A_t\| \leq 2$ almost surely implies that for any $T \leq t$

$$\left\| \Big(\prod_{s=0}^{t} (\mathrm{Id} - \mathbb{E}[A_s]) \Big) Q_{\|} \right\|_2 \leq \left\| \Big(\prod_{s=0}^{T} (\mathrm{Id} - \mathbb{E}[A_s]) \Big) Q_{\|} \right\|_2$$

since each term in the product is non-negative-definite. Thus, we find

$$\|X_{t+1} Q_{\|}\|_F \leq \left[\|X_0\|_F + \sum_{s=0}^{T_1-1} \|B_s\|_F \right] \left\| \Big(\prod_{s=T_1}^{t} (\mathrm{Id} - \mathbb{E}[A_s]) \Big) Q_{\|} \right\|_2 + \sum_{s=T_1}^{t} \|B_s\|_F < \varepsilon.$$

Taking $t \to \infty$ and then $\varepsilon \to 0$ implies that $\lim_{t \to \infty} X_t Q_{\|} = 0$, as desired. □

Lemma A.3. *If a sequence $\{Z_t\}$ has two-sided decay of type $(\{A_t\}, \{C_t\})$ with*

$$\lim_{T \to \infty} \mathbb{E} \left[\left\| \Big(\prod_{s=t}^{T} (\mathrm{Id} - A_s) \Big) Q_{\|} \right\|_2^2 \right] = 0 \quad \textit{for all } t \geq 0 \tag{A.7}$$

and

$$\sum_{t=0}^{\infty} tr(C_t) < \infty, \tag{A.8}$$

then $\lim_{t \to \infty} Q_{\|}^{\mathsf{T}} Z_t Q_{\|} = 0$.

Proof. The proof is essentially identical to that of Lemma A.2. That is, for $\varepsilon > 0$, choose T_1 so that $\sum_{t=T_1}^{\infty} \mathrm{tr}(C_t) < \frac{\varepsilon}{2}$ and choose T_2 by (A.7) so that for $t > T_2$ we have

$$\mathbb{E} \left[\left\| \Big(\prod_{s=T_1}^{t} (\mathrm{Id} - A_s) \Big) Q_{\|} \right\|_2^2 \right] < \frac{\varepsilon}{2} \frac{1}{\mathrm{tr}(Z_0) + \sum_{s=0}^{T_1-1} \mathrm{tr}(C_s)}.$$

Conjugating (A.4) by $Q_{\|}$, we have that

$$Q_{\|}^{\mathsf{T}} Z_{t+1} Q_{\|} = \mathbb{E} \left[Q_{\|}^{\mathsf{T}} \Big(\prod_{s=0}^{t} (\mathrm{Id} - A_s) \Big)^{\mathsf{T}} Z_0 \Big(\prod_{s=0}^{t} (\mathrm{Id} - A_s) \Big) Q_{\|} \right] + \sum_{s=0}^{t} \mathbb{E} \left[Q_{\|}^{\mathsf{T}} \Big(\prod_{r=s+1}^{t} (\mathrm{Id} - A_r) \Big)^{\mathsf{T}} C_s \Big(\prod_{r=s+1}^{t} (\mathrm{Id} - A_r) \Big) Q_{\|} \right].$$

Our assumption that $\|A_t\| \leq 2$ almost surely implies that for any $T \leq t$

$$\left\| \Big(\prod_{s=0}^{t} (\mathrm{Id} - A_s) \Big) Q \right\|_2 \leq \left\| \Big(\prod_{s=0}^{T} (\mathrm{Id} - A_s) \Big) Q \right\|_2.$$

For $t > T_2$, this implies by taking trace of both sides that

$$\operatorname{tr}(Q_{\|}^{\mathsf{T}} Z_{t+1} Q_{\|}) \leq \operatorname{tr}(Z_0)\mathbb{E}\left[\left\|\Big(\prod_{s=0}^{t}(\mathrm{Id}-A_s)\Big)Q_{\|}\right\|_2^2\right] + \sum_{s=0}^{t}\operatorname{tr}(C_s)\mathbb{E}\left[\left\|\Big(\prod_{r=s+1}^{t}(\mathrm{Id}-A_r)\Big)Q_{\|}\right\|_2^2\right] \tag{A.9}$$

$$\leq \left[\operatorname{tr}(Z_0) + \sum_{s=0}^{T_1-1}\operatorname{tr}(C_s)\right]\mathbb{E}\left[\left\|\Big(\prod_{s=T_1}^{t}(\mathrm{Id}-A_s)\Big)Q_{\|}\right\|_2^2\right] + \sum_{s=T_1}^{t}\operatorname{tr}(C_s)$$

$$< \varepsilon,$$

which implies that $\lim_{t\to\infty} Q_{\|}^{\mathsf{T}} Z_t Q_{\|} = 0$. □

The preceding Lemmas will be used to provide sufficient conditions for augmented gradient descent to converge as in Theorem B.1 below. Since we are also interested in obtaining rates of convergence, we record here two quantitative refinements of the Lemmas above that will be used in the proof of Theorem B.4.

Lemma A.4. *Suppose $\{X_t\}$ has one-sided decay of type $(\{A_t\}, \{B_t\})$. Assume also that for some $X \geq 0$ and $C > 0$, we have*

$$\log\left\|\Big(\prod_{r=s}^{t}(\mathrm{Id}-\mathbb{E}[A_r])\Big)Q_{\|}\right\|_2 < X - C\int_s^{t+1} r^{-\alpha}dr$$

and $\|B_t\|_F = O(t^{-\beta})$ for some $0 < \alpha < 1 < \beta$. Then, $\|X_t Q_{\|}\|_F = O(t^{\alpha-\beta})$.

Proof. Denote $\gamma_{s,t} := \int_s^t r^{-\alpha}dr$. By (A.6), we have for some constants $C_1, C_2 > 0$ that

$$\|X_{t+1}Q_{\|}\|_F < C_1 e^{-C\gamma_{1,t+1}} + C_2 e^X \sum_{s=1}^{t}(1+s)^{-\beta}e^{-C\gamma_{s+1,t+1}}. \tag{A.10}$$

The first term on the right hand side is exponentially decaying in t since $\gamma_{1,t+1}$ grows polynomially in t. To bound the second term, observe that the function

$$f(s) := C\gamma_{s+1,t+1} - \beta\log(s+1)$$

satisfies

$$f'(s) \geq 0 \quad\Leftrightarrow\quad C(s+1)^{-\alpha} - \frac{\beta}{1+s} \geq 0 \quad\Leftrightarrow\quad s \geq \left(\frac{\beta}{C}\right)^{1/(1-\alpha)} =: K.$$

Hence, the summands are monotonically increasing for s greater than a fixed constant K depending only on α, β, C. Note that

$$\sum_{s=1}^{K}(1+s)^{-\beta}e^{-C\gamma_{s+1,t+1}} \leq Ke^{-C\gamma_{K+1,t+1}} \leq Ke^{-C't^{1-\alpha}}$$

for some C' depending only on α and K, and hence sum is exponentially decaying in t. Further, using an integral comparison, we find

$$\sum_{s=K+1}^{t}(1+s)^{-\beta}e^{-C\gamma_{s+1,t+1}} \leq \int_K^t (1+s)^{-\beta}e^{-\frac{C}{1-\alpha}\left((t+1)^{1-\alpha}-(s+1)^{1-\alpha}\right)}ds. \tag{A.11}$$

Changing variables using $u = (1+s)^{1-\alpha}/(1-\alpha)$, the last integral has the form

$$e^{-Cg_t}(1-\alpha)^{-\xi}\int_{g_K}^{g_t} u^{-\xi}e^{Cu}du, \qquad g_x := \frac{(1+x)^{1-\alpha}}{1-\alpha},\ \xi := \frac{\beta-\alpha}{1-\alpha}. \tag{A.12}$$

Integrating by parts, we have

$$\int_{g_K}^{g_t} u^{-\xi}e^{u}du = C^{-1}\xi\int_{g_K}^{g_t} u^{-\xi-1}e^{Cu}du + (u^{-\xi}e^{Cu})|_{g_K}^{g_t}$$

Further, since on the range $g_K \leq u \leq g_t$ the integrand is increasing, we have

$$e^{-Cg_t}\xi \int_{g_K}^{g_t} u^{-\xi-1}e^{Cu}du \leq \xi g_t^{-\xi}.$$

Hence, e^{-Cg_t} times the integral in (A.12) is bounded above by

$$O(g_t^{-\xi}) + e^{-Cg_t}(u^{-\xi}e^{Cu})|_{g_K}^{g_t} = O(g_t^{-\xi}).$$

Using (A.11) and substituting the previous line into (A.12) yields the estimate

$$\sum_{s=K+1}^{t}(1+s)^{-\beta}e^{-C\gamma_{s+1,t+1}} \leq (1+t)^{-\beta+\alpha},$$

which completes the proof. □

Lemma A.5. *Suppose $\{Z_t\}$ has two-sided decay of type $(\{A_t\},\{C_t\})$. Assume also that for some $X \geq 0$ and $C > 0$, we have*

$$\log \mathbb{E}\left[\left\|\Big(\prod_{r=s}^{t}(\mathrm{Id}-A_r)\Big)Q_{\|}\right\|_2^2\right] < X - C\int_s^{t+1} r^{-\alpha}dr$$

as well as tr$(C_t) = O(t^{-\beta})$ for some $0 < \alpha < 1 < \beta$. Then tr$(Q_{\|}^T Z_t Q_{\|}) = O(t^{\alpha-\beta})$.

Proof. This argument is identical to the proof of Lemma A.4. Indeed, using (A.9) we have that

$$\mathrm{tr}\left(Q_{\|}^T Z_t Q_{\|}\right) \leq C_1 e^{-C\gamma_{1,t+1}} + C_2 e^X \sum_{s=1}^{t}(1+s)^{-\beta}e^{-C\gamma_{s+1,t+1}}.$$

The right hand side of this inequality coincides with the expression on the right hand side of (A.10), which we already bounded by $O(t^{\beta-\alpha})$ in the proof of Lemma A.4. □

In what follows, we will use a concentration result for products of matrices from (18). Let $Y_1, \ldots, Y_n \in \mathbb{R}^{N\times N}$ be independent random matrices. Suppose that

$$\|\mathbb{E}[Y_i]\|_2 \leq a_i \qquad \text{and} \qquad \mathbb{E}\left[\|Y_i - \mathbb{E}[Y_i]\|_2^2\right] \leq b_i^2 a_i^2$$

for some $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$. We will use the following result, which is a specialization of Theorem 5.1 in (18) for $p = q = 2$.

Theorem A.6 (Theorem 5.1 in (18))**.** *For $Z_0 \in \mathbb{R}^{N\times n}$, the product $Z_n = Y_n Y_{n-1}\cdots Y_1 Z_0$ satisfies*

$$\mathbb{E}\left[\|Z_n\|_2^2\right] \leq e^{\sum_{i=1}^n b_i^2}\prod_{i=1}^{n} a_i^2 \cdot \|Z_0\|_2^2$$

$$\mathbb{E}\left[\|Z_n - \mathbb{E}[Z_n]\|_2^2\right] \leq \Big(e^{\sum_{i=1}^n b_i^2} - 1\Big)a_i^2 \cdot \|Z_0\|_2^2.$$

Finally, we collect two simple analytic lemmas for later use.

Lemma A.7. *For any matrix $M \in \mathbb{R}^{m\times n}$, we have that*

$$\mathbb{E}[\|M\|_2^2] \geq \|\mathbb{E}[M]\|_2^2.$$

Proof. We find by Cauchy-Schwartz and the convexity of the spectral norm that

$$\mathbb{E}[\|M\|_2^2] \geq \mathbb{E}[\|M\|_2]^2 \geq \|\mathbb{E}[M]\|_2^2.$$ □

Lemma A.8. *For bounded $a_t \geq 0$, if we have $\sum_{t=0}^{\infty} a_t = \infty$, then for any $C > 0$ we have*

$$\sum_{t=0}^{\infty} a_t e^{-C\sum_{s=0}^{t} a_s} < \infty.$$

Proof. Define $b_t := \sum_{s=0}^{t} a_s$ so that

$$S := \sum_{t=0}^{\infty} a_t e^{-C\sum_{s=0}^{t} a_s} = \sum_{t=0}^{\infty}(b_t - b_{t-1})e^{-Cb_t} \leq \int_0^{\infty} e^{-Cx}dx < \infty,$$

where we use $\int_0^\infty e^{-Cx}dx$ to upper bound its right Riemann sum. □

B Analysis of data augmentation as stochastic optimization

In this section, we prove generalizations of our main theoretical results Theorems 3.1 and 3.2 giving Monro-Robbins type conditions for convergence and rates for augmented gradient descent in the linear setting.

B.1 Monro-Robbins type results

To state our general Monro-Robbins type convergence results, let us briefly recall the notation. We consider overparameterized linear regression with loss

$$\mathcal{L}(W;\mathcal{D}) = \frac{1}{N}\left\|WX - Y\right\|_F^2,$$

where the dataset $\mathcal{D}$ of size N consists of data matrices X, Y that each have N columns $x_i \in \mathbb{R}^n, y_i \in \mathbb{R}^p$ with $n > N$. We optimize $\mathcal{L}(W;\mathcal{D})$ by augmented gradient descent, which means that at each time t we replace $\mathcal{D} = (X, Y)$ by a random dataset $\mathcal{D}_t = (X_t, Y_t)$. We then take a step

$$W_{t+1} = W_t - \eta_t \nabla_W \mathcal{L}(W_t;\mathcal{D}_t)$$

of gradient descent on the resulting randomly augmented loss $\mathcal{L}(W;\mathcal{D}_t)$ with learning rate η_t. Recall that we set

$$V_{||} := \text{ column span of } \mathbb{E}[X_t X_t^{\mathsf{T}}]$$

and denoted by $Q_{||}$ the orthogonal projection onto $V_{||}$. As noted in §3, on $V_{||}$ the proxy loss

$$\overline{\mathcal{L}}_t = \mathbb{E}\left[\mathcal{L}(W;\mathcal{D}_t)\right]$$

is strictly convex and has a unique minimum, which is

$$W_t^* = \mathbb{E}\left[Y_t X_t^T\right](Q_{||}\mathbb{E}\left[X_t X_t^{\mathsf{T}}\right]Q_{||})^{-1}.$$

The change from one step of augmented GD to the next in these proxy optima is captured by

$$\Xi_t^* := W_{t+1}^* - W_t^*.$$

With this notation, we are ready to state Theorems B.1, which gives two different sets of time-varying Monro-Robbins type conditions under which the optimization trajectory W_t converges for large t. In Theorem B.4, we refine the analysis to additionally give rates of convergence. Note that Theorem B.1 is a generalization of Theorem 3.1 and that Theorem B.4 is a generalization of Theorem 3.2.

Theorem B.1. *Suppose that $V_{||}$ is independent of t, that the learning rate satisfies $\eta_t \to 0$, that the proxy optima satisfy*

$$\sum_{t=0}^{\infty}\|\Xi_t^*\|_F < \infty, \tag{B.1}$$

ensuring the existence of a limit $W_\infty^* := \lim_{t\to\infty} W_t^*$ and that

$$\sum_{t=0}^{\infty}\eta_t\lambda_{min,V_{||}}(\mathbb{E}[X_t X_t^{\mathsf{T}}]) = \infty. \tag{B.2}$$

Then if either

$$\sum_{t=0}^{\infty}\eta_t^2\mathbb{E}\left[\|X_t X_t^{\mathsf{T}} - \mathbb{E}[X_t X_t^{\mathsf{T}}]\|_F^2 + \|Y_t X_t^{\mathsf{T}} - \mathbb{E}[Y_t X_t^{\mathsf{T}}]\|_F^2\right] < \infty \tag{B.3}$$

or

$$\sum_{t=0}^{\infty}\eta_t^2\mathbb{E}\Big[\|X_t X_t^{\mathsf{T}} - \mathbb{E}[X_t X_t^{\mathsf{T}}]\|_F^2 + \Big\|\mathbb{E}[W_t](X_t X_t^{\mathsf{T}} - \mathbb{E}[X_t X_t^{\mathsf{T}}]) - (Y_t X_t^{\mathsf{T}} - \mathbb{E}[Y_t X_t^{\mathsf{T}}])\Big\|_F^2\Big] < \infty \tag{B.4}$$

hold, then for any initialization W_0, we have $W_t Q_{||} \xrightarrow{p} W_\infty^*$.

Remark B.2. *In the general case, the column span $V_{||}$ of $\mathbb{E}[X_tX_t^{\mathsf{T}}]$ may vary with t. This means that some directions in $\mathbb{R}^n$ may only have non-zero overlap with colspan$(\mathbb{E}[X_tX_t^{\mathsf{T}}])$ for some positive but finite collection of values of t. In this case, only finitely many steps of the optimization would move W_t in this direction, meaning that we must define a smaller space for convergence. The correct definition of this subspace turns out to be the following*

$$\begin{aligned} V_{||} &:= \bigcap_{t=0}^{\infty} \ker\left[\prod_{s=t}^{\infty}\Big(\mathrm{Id} - \frac{2\eta_s}{N}\mathbb{E}[X_sX_s^{\mathsf{T}}]\Big)\right] \qquad \text{(B.5)} \\ &= \bigcap_{t=0}^{\infty}\left\{u \in \mathbb{R}^n \;\middle|\; \lim_{T\to\infty}\prod_{s=t}^{T}\Big(\mathrm{Id} - \frac{2\eta_s}{N}\mathbb{E}[X_sX_s^{\mathsf{T}}]\Big)u = 0\right\}. \end{aligned}$$

With this re-definition of $V_{||}$ and with $Q_{||}$ still denoting the orthogonal projection to $V_{||}$, Theorem B.1 holds verbatim and with the same proof. Note that if $\eta_t \to 0$, $V_{||} =$ colspan$(\mathbb{E}[X_tX_t^{\mathsf{T}}])$ is fixed in t, and (B.2) holds, this definition of $V_{||}$ reduces to that defined in (3.5).

Remark B.3. *The condition* (B.4) *can be written in a more conceptual way as*

$$\sum_{t=0}^{\infty}\left[\|X_tX_t^{\mathsf{T}} - \mathbb{E}[X_tX_t^{\mathsf{T}}]\|_F^2 + \eta_t^2 tr\left[\mathrm{Id}\circ Var\Big((\mathbb{E}[W_t]X_t - Y_t)X_t^{\mathsf{T}}\Big)\right]\right] < \infty,$$

where we recognize that $(\mathbb{E}[W_t]X_t - Y_t)X_t^{\mathsf{T}}$ is precisely the stochastic gradient estimate at time t for the proxy loss $\overline{\mathcal{L}}_t$, evaluated at $\mathbb{E}[W_t]$, which is the location at time t for vanilla GD on $\overline{\mathcal{L}}_t$ since taking expectations in the GD update equation (3.3) coincides with GD for $\overline{\mathcal{L}}_t$. Moreover, condition (B.4) *actually implies condition* (B.3) *(see* (B.12) *below). The reason we state Theorem B.1 with both conditions, however, is that* (B.4) *makes explicit reference to the average $\mathbb{E}[W_t]$ of the augmented trajectory. Thus, when applying Theorem B.1 with this weaker condition, one must separately estimate the behavior of this quantity.*

Theorem B.1 gave conditions on joint learning rate and data augmentation schedules under which augmented optimization is guaranteed to converge. Our next result proves rates for this convergence.

Theorem B.4. *Suppose that $\eta_t \to 0$ and that for some $0 < \alpha < 1 < \beta_1, \beta_2$ and $C_1, C_2 > 0$, we have*

$$\log \mathbb{E}\left[\left\|\Big(\prod_{r=s}^{t}\Big(\mathrm{Id} - \frac{2\eta_r}{N}X_rX_r^{\mathsf{T}}\Big)\Big)Q_{||}\right\|_2^2\right] < C_1 - C_2\int_s^{t+1} r^{-\alpha}dr \qquad \text{(B.6)}$$

as well as

$$\|\Xi_t^*\|_F = O(t^{-\beta_1}) \qquad \text{(B.7)}$$

and

$$\eta_t^2 tr\left[\mathrm{Id}\circ Var(\mathbb{E}[W_t]X_tX_t^{\mathsf{T}} - Y_tX_t^{\mathsf{T}}\Big)\right] = O(t^{-\beta_2}). \qquad \text{(B.8)}$$

Then, for any initialization W_0, we have for any $\varepsilon > 0$ that

$$t^{\min\{\beta_1 - 1, \frac{\beta_2-\alpha}{2}\}-\varepsilon}\|W_tQ_{||} - W_\infty^*\|_F \xrightarrow{p} 0.$$

Remark B.5. *To reduce Theorem 3.2 to Theorem B.4, we notice that (3.10) and (3.11) mean that Theorem A.6 applies to $Y_t = \mathrm{Id} - 2\eta_t\frac{X_tX_t^{\mathsf{T}}}{N}$ with $a_t = 1 - \Omega(t^{-\alpha})$ and and $b_t^2 = O(t^{-\gamma})$, thus implying (B.6).*

The first step in proving both Theorem B.1 and Theorem B.4 is to obtain recursions for the mean and variance of the difference $W_t - W_t^*$ between the time t proxy optimum and the augmented optimization trajectory at time t. We will then complete the proof of Theorem B.1 in §B.3 and the proof of Theorem B.4 in §B.4.

B.2 Recursion relations for parameter moments

The following proposition shows that difference between the mean augmented dynamics $\mathbb{E}[W_t]$ and the time$-t$ optimum W_t^* satisfies, in the sense of Definition A.1, one-sided decay of type $(\{A_t\}, \{B_t\})$ with

$$A_t = \frac{2\eta_t}{N}X_tX_t^{\mathsf{T}}, \qquad B_t = -\Xi_t^*.$$

It also shows that the variance of this difference, which is non-negative definite, satisfies two-sided decay of type $(\{A_t\}, \{C_t\})$ with A_t as before and

$$C_t = \frac{4\eta_t^2}{N^2}\left[\mathrm{Id} \circ \mathrm{Var}\Big(\mathbb{E}[W_t]X_tX_t^{\mathsf{T}} - Y_tX_t^{\mathsf{T}}\Big)\right].$$

In terms of the notations of Appendix A.1, we have the following recursions.

Proposition B.6. *The quantity* $\mathbb{E}[W_t] - W_t^*$ *satisfies*

$$\mathbb{E}[W_{t+1}] - W_{t+1}^* = (\mathbb{E}[W_t] - W_t^*)\Big(\mathrm{Id} - \frac{2\eta_t}{N}\mathbb{E}[X_tX_t^{\mathsf{T}}]\Big) - \Xi_t^* \tag{B.9}$$

and $Z_t := \mathbb{E}[(W_t - \mathbb{E}[W_t])^{\mathsf{T}}(W_t - \mathbb{E}[W_t])]$ *satisfies*

$$Z_{t+1} = \mathbb{E}\left[(\mathrm{Id} - \frac{2\eta_t}{N}X_tX_t^{\mathsf{T}})Z_t(\mathrm{Id} - \frac{2\eta_t}{N}X_tX_t^{\mathsf{T}})\right] + \frac{4\eta_t^2}{N^2}\left[\mathrm{Id} \circ \mathit{Var}\Big(\mathbb{E}[W_t]X_tX_t^{\mathsf{T}} - Y_tX_t^{\mathsf{T}}\Big)\right]. \tag{B.10}$$

Proof. Notice that $\mathbb{E}[X_tX_t^{\mathsf{T}}]u = 0$ if and only if $X_t^{\mathsf{T}}u = 0$ almost surely, which implies that

$$W_t^*\mathbb{E}[X_tX_t^{\mathsf{T}}] = \mathbb{E}[Y_tX_t^{\mathsf{T}}]\mathbb{E}[X_tX_t^{\mathsf{T}}]^+\mathbb{E}[X_tX_t^{\mathsf{T}}] = \mathbb{E}[Y_tX_t^{\mathsf{T}}].$$

Thus, the learning dynamics (3.3) yield

$$\begin{aligned}\mathbb{E}[W_{t+1}] &= \mathbb{E}[W_t] - \frac{2\eta_t}{N}\Big(\mathbb{E}[W_t]\mathbb{E}[X_tX_t^{\mathsf{T}}] - \mathbb{E}[Y_tX_t^{\mathsf{T}}]\Big)\\ &= \mathbb{E}[W_t] - \frac{2\eta_t}{N}(\mathbb{E}[W_t] - W_t^*)\mathbb{E}[X_tX_t^{\mathsf{T}}].\end{aligned}$$

Subtracting W_{t+1}^* from both sides yields (B.9). We now analyze the fluctuations. Writing $\mathrm{Sym}(A) := A + A^{\mathsf{T}}$, we have

$$\begin{aligned}\mathbb{E}[W_{t+1}]^{\mathsf{T}}\mathbb{E}[W_{t+1}] &= \mathbb{E}[W_t]^{\mathsf{T}}\mathbb{E}[W_t] + \frac{2\eta_t}{N}\mathrm{Sym}\Big(\mathbb{E}[W_t]^{\mathsf{T}}\mathbb{E}[Y_tX_t^{\mathsf{T}}] - \mathbb{E}[W_t]^{\mathsf{T}}\mathbb{E}[W_t]\mathbb{E}[X_tX_t^{\mathsf{T}}]\Big)\\ &+\frac{4\eta_t^2}{N^2}\Big(\mathbb{E}[X_tX_t^{\mathsf{T}}]\mathbb{E}[W_t]^{\mathsf{T}}\mathbb{E}[W_t]\mathbb{E}[X_tX_t^{\mathsf{T}}] + \mathbb{E}[X_tY_t^{\mathsf{T}}]\mathbb{E}[Y_tX_t^{\mathsf{T}}] - \mathrm{Sym}(\mathbb{E}[X_tX_t^{\mathsf{T}}]\mathbb{E}[W_t]^{\mathsf{T}}\mathbb{E}[Y_tX_t^{\mathsf{T}}])\Big).\end{aligned}$$

Similarly, we have that

$$\begin{aligned}\mathbb{E}[W_{t+1}^{\mathsf{T}}W_{t+1}] &= \mathbb{E}[W_t^{\mathsf{T}}W_t] + \frac{2\eta_t}{N}\mathrm{Sym}(\mathbb{E}[W_t^{\mathsf{T}}Y_tX_t^{\mathsf{T}} - W_t^{\mathsf{T}}W_tX_tX_t^{\mathsf{T}}])\\ &+ \frac{4\eta_t^2}{N^2}\mathbb{E}[X_tX_t^{\mathsf{T}}W_t^{\mathsf{T}}W_tX_tX_t^{\mathsf{T}} - \mathrm{Sym}(X_tX_t^{\mathsf{T}}W_t^{\mathsf{T}}Y_tX_t^{\mathsf{T}}) + X_tY_t^{\mathsf{T}}Y_tX_t^{\mathsf{T}}].\end{aligned}$$

Noting that X_t and Y_t are independent of W_t and subtracting yields the desired. □

B.3 Proof of Theorem B.1

First, by Proposition B.6, we see that $\mathbb{E}[W_t] - W_t^*$ has one-sided decay with

$$A_t = 2\eta_t\frac{X_tX_t^{\mathsf{T}}}{N} \qquad \text{and} \qquad B_t = -\Xi_t^*.$$

Thus, by Lemma A.2 and (B.1), we find that

$$\lim_{t\to\infty}(\mathbb{E}[W_t]Q_\| - W_t^*) = 0, \tag{B.11}$$

which gives convergence in expectation.

For the second moment, by Proposition B.6, we see that Z_t has two-sided decay with

$$A_t = 2\eta_t\frac{X_tX_t^{\mathsf{T}}}{N} \qquad \text{and} \qquad C_t = \frac{4\eta_t^2}{N^2}\left[\mathrm{Id} \circ \mathrm{Var}\Big(\mathbb{E}[W_t]X_tX_t^{\mathsf{T}} - Y_tX_t^{\mathsf{T}}\Big)\right].$$

We now verify (A.7) and (A.8) in order to apply Lemma A.3.

For (A.7), for any $\varepsilon > 0$, notice that

$$\mathbb{E}[\|A_s - \mathbb{E}[A_s]\|_F^2] = \eta_s^2 \mathbb{E}[\|X_s X_s^\intercal - \mathbb{E}[X_s X_s^\intercal]\|_F^2]$$

so by either (B.3) or (B.4) we may choose $T_1 > t$ so that $\sum_{s=T_1}^{\infty} \mathbb{E}[\|A_s - \mathbb{E}[A_s]\|_F^2] < \frac{\varepsilon}{2}$. Now choose $T_2 > T_1$ so that for $T > T_2$, we have

$$\left\|\Big(\prod_{r=T_1}^{T} \mathbb{E}[\mathrm{Id} - A_r]\Big) Q_\|\right\|_2^2 < \frac{\varepsilon}{2} \frac{1}{\|\prod_{s=t}^{T_1-1} \mathbb{E}[\mathrm{Id} - A_s]\|_F^2 + \sum_{s=t}^{T_1-1} \mathbb{E}[\|A_s - \mathbb{E}[A_s]\|_F^2]}.$$

For $T > T_2$, we then have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\Big(\prod_{s=t}^{T} (\mathrm{Id} - A_s)\Big) Q_\|\right\|_2^2\right] \\
&\leq \left\|\Big(\prod_{s=t}^{T} \mathbb{E}[\mathrm{Id} - A_s]\Big) Q_\|\right\|^2 + \sum_{s=t}^{T} \mathbb{E}\left[\left\|\prod_{r=t}^{s} (\mathrm{Id} - A_r) \prod_{r=s+1}^{T} (\mathrm{Id} - \mathbb{E}[A_r]) Q_\|\right\|_F^2 - \left\|\prod_{r=t}^{s-1} (\mathrm{Id} - A_r) \prod_{r=s}^{T} (\mathrm{Id} - \mathbb{E}[A_r]) Q_\|\right\|_F^2\right] \\
&= \left\|\Big(\prod_{s=t}^{T} \mathbb{E}[\mathrm{Id} - A_s]\Big) Q_\|\right\|_F^2 + \sum_{s=t}^{T} \mathbb{E}\left[\left\|\prod_{r=t}^{s-1} (\mathrm{Id} - A_r)(A_s - \mathbb{E}[A_s]) \prod_{r=s+1}^{T} (\mathrm{Id} - \mathbb{E}[A_r]) Q_\|\right\|_F^2\right] \\
&\leq \left\|\prod_{s=t}^{T_1-1} \mathbb{E}[\mathrm{Id} - A_s]\right\|^2 \left\|_F \Big(\prod_{r=T_1}^{T} \mathbb{E}[\mathrm{Id} - A_r]\Big) Q_\|\right\|_2^2 + \sum_{s=t}^{T} \mathbb{E}[\|A_s - \mathbb{E}[A_s]\|_F^2] \left\|\Big(\prod_{r=s+1}^{T} \mathbb{E}[\mathrm{Id} - A_r]\Big) Q_\|\right\|_2^2 \\
&\leq \left(\left\|\prod_{s=t}^{T_1-1} \mathbb{E}[\mathrm{Id} - A_s]\right\|_F^2 + \sum_{s=t}^{T_1-1} \mathbb{E}[\|A_s - \mathbb{E}[A_s]\|_F^2]\right) \left\|\Big(\prod_{r=T_1}^{T} \mathbb{E}[\mathrm{Id} - A_r]\Big) Q_\|\right\|_2^2 + \sum_{s=T_1}^{T} \mathbb{E}[\|A_s - \mathbb{E}[A_s]\|_F^2] \\
&< \varepsilon,
\end{aligned}$$

which implies (A.7). Condition (A.8) follows from either (B.4) or (B.3) and the bounds

$$\begin{aligned}
\mathrm{tr}(C_t) &\leq \frac{8\eta_t^2}{N^2} \left(\|\mathbb{E}[W_t](X_t X_t^\intercal - \mathbb{E}[X_t X_t^\intercal])\|_F^2 + \|Y_t X_t^\intercal - \mathbb{E}[Y_t X_t^\intercal]\|_F^2\right) &\text{(B.12)} \\
&\leq \frac{8\eta_t^2}{N^2} \left(\|\mathbb{E}[W_t]\|^2 \|X_t X_t^\intercal - \mathbb{E}[X_t X_t^\intercal]\|_F^2 + \|Y_t X_t^\intercal - \mathbb{E}[Y_t X_t^\intercal]\|_F^2\right),
\end{aligned}$$

where in the first inequality we use the fact that $\|M_1 - M_2\|_F^2 \leq 2(\|M_1\|_F^2 + \|M_2\|_F^2)$. Furthermore, iterating (B.9) yields $\|\mathbb{E}[W_t] - W_t^*\|_F \leq \|W_0 - W_0^*\|_F + \sum_{t=0}^{\infty} \|\Xi_t^*\|_F$, which combined with (B.12) and either (B.3) or (B.4) therefore implies (A.8). We conclude by Lemma A.3 that

$$\lim_{t\to\infty} Q_\|^\intercal Z_t Q_\| = \lim_{t\to\infty} \mathbb{E}[Q_\|^\intercal (W_t - \mathbb{E}[W_t])^\intercal (W_t - \mathbb{E}[W_t]) Q_\|] = 0. \qquad \text{(B.13)}$$

Together, (B.11) and (B.13) imply that $W_t Q_\| - W_t^* \xrightarrow{p} 0$. The conclusion then follows from the fact that $\lim_{t\to 0} W_t^* = W_\infty^*$. This complete the proof of Theorem B.1. □

B.4 Proof of Theorem B.4

By Proposition B.6, $\mathbb{E}[W_t] - W_t^*$ has one-sided decay with

$$A_t = \frac{2\eta_t}{N} X_t X_t^\intercal, \qquad B_t = -\Xi_t^*.$$

By Lemma A.7 and (B.6), $\mathbb{E}[A_t]$ satisfies

$$\begin{aligned}
\log\left\|\prod_{r=s}^{t} \Big(\mathrm{Id} - 2\eta_r \frac{1}{N} \mathbb{E}[X_r X_r^\intercal]\Big) Q_\|\right\|_2 &\leq \frac{1}{2} \log \mathbb{E}\left[\left\|\Big(\prod_{r=s}^{t} \Big(\mathrm{Id} - 2\eta_r \frac{X_r X_r^\intercal}{N}\Big)\Big) Q_\|\right\|_2^2\right] \\
&< \frac{C_1}{2} - \frac{C_2}{2} \int_s^{t+1} r^{-\alpha} dr.
\end{aligned}$$

Applying Lemma A.4 using this bound and (B.7), we find that

$$\|\mathbb{E}[W_t]Q_\| - W_t^*\|_F = O(t^{\alpha-\beta_1}).$$

Moreover, because $\|\Xi_t^*\|_F = O(t^{-\beta_1})$, we also find that $\|W_t^* - W_\infty^*\|_F = O(t^{-\beta_1+1})$, and hence

$$\|\mathbb{E}[W_t]Q_\| - W_\infty^*\|_F = O(t^{-\beta_1+1}).$$

Further, by Proposition B.6, $\mathbb{E}[(W_t - \mathbb{E}[W_t])^\mathsf{T}(W_t - \mathbb{E}[W_t])]$ has two-sided decay with

$$A_t = \frac{2\eta_t}{N} X_t X_t^\mathsf{T}, \qquad C_t = \frac{4\eta_t^2}{N^2}\left[\mathrm{Id} \circ \mathrm{Var}\Big(\mathbb{E}[W_t]X_tX_t^\mathsf{T} - Y_tX_t^\mathsf{T}\Big)\right].$$

Applying Lemma A.5 with (B.6) and (B.8), we find that

$$\mathbb{E}\left[\|(W_t - \mathbb{E}[W_t])Q_\|\|_F^2\right] = O(t^{\alpha-\beta_2}).$$

By Chebyshev's inequality, for any $x > 0$ we have

$$\mathbb{P}\Big(\|W_tQ_\| - W_\infty^*\|_F \geq O(t^{-\beta_1+1}) + x \cdot O(t^{\frac{\alpha-\beta_2}{2}})\Big) \leq x^{-2}.$$

For any $\varepsilon > 0$, choosing $x = t^\delta$ for small $0 < \delta < \varepsilon$ we find as desired that

$$t^{\min\{\beta_1-1, \frac{\beta_2-\alpha}{2}\}-\varepsilon}\|W_tQ_\| - W_\infty^*\|_F \xrightarrow{p} 0,$$

thus completing the proof of Theorem B.4. □

C Intrinsic time

Theorem 3.2 measures rates in terms of optimization steps t, but a different measurement of time called the *intrinsic time* of the optimization will be more suitable for measuring the behavior of optimization quantities. This was introduced for SGD in (28; 27), and we now generalize it to our broader setting. For gradient descent on a loss $\mathcal{L}$, the intrinsic time is a quantity which increments by $\eta\lambda_{\min}(H)$ for a optimization step with learning rate η at a point where $\mathcal{L}$ has Hessian H. When specialized to our setting, it is given by

$$\tau(t) := \sum_{s=0}^{t-1} \frac{2\eta_s}{N}\lambda_{\min,V_\|}(\mathbb{E}[X_sX_s^\mathsf{T}]). \tag{C.1}$$

Notice that intrinsic time of augmented optimization for the sequence of proxy losses $\overline{\mathcal{L}}_s$ appears in Theorems 3.1 and 3.2, which require via conditions (3.8) and (3.11) that the intrinsic time tends to infinity as the number of optimization steps grows.

Intrinsic time will be a sensible variable in which to measure the behavior of quantities such as the fluctuations of the optimization path $f(t) := \mathbb{E}[\|(W_t - \mathbb{E}[W_t])Q_\|\|_F^2]$. In the proofs of Theorems 3.1 and 3.2, we show that the fluctuations satisfy an inequality of the form

$$f(t+1) \leq f(t)(1 - a(t))^2 + b(t) \tag{C.2}$$

for $a(t) := 2\eta_t \frac{1}{N}\lambda_{\min,V_\|}(\mathbb{E}[X_tX_t^\mathsf{T}])$ and $b(t) := \mathrm{Var}[\|\eta_t\nabla_W\mathcal{L}(W_t)\|_F]$ so that $\tau(t) = \sum_{s=0}^{t-1} a(s)$. Iterating the recursion (C.2) shows that

$$\begin{aligned} f(t) &\leq f(0)\prod_{s=0}^{t-1}(1-a(s))^2 + \sum_{s=0}^{t-1} b(s) \prod_{r=s+1}^{t-1}(1-a(r))^2 \\ &\leq e^{-2\tau(t)}f(0) + \sum_{s=0}^{t-1}\frac{b(s)}{a(s)}e^{2\tau(s+1)-2\tau(t)}(\tau(s+1)-\tau(s)). \end{aligned}$$

Changing variables to $\tau := \tau(t)$ and defining $A(\tau)$, $B(\tau)$, and $F(\tau)$ by $A(\tau(t)) = a(t)$, $B(\tau(t)) = b(t)$, and $F(\tau(t)) = f(t)$, we find by replacing a right Riemann sum by an integral that

$$F(\tau) \lesssim e^{-2\tau}\left[F(0) + \int_0^\tau \frac{B(\sigma)}{A(\sigma)}e^{2\sigma}d\sigma\right]. \tag{C.3}$$

In order for the result of optimization to be independent of the starting point, by (C.3) we must have $\tau \to \infty$ to remove the dependence on $F(0)$; this provides one explanation for the appearance of τ in condition (3.8). Further, (C.3) implies that the fluctuations at an intrinsic time are bounded by an integral against the function $\frac{B(\sigma)}{A(\sigma)}$ which depends only on the ratio of $A(\sigma)$ and $B(\sigma)$.

In the case of minibatch SGD, our proof of Theorem F.1 shows the intrinsic time is $\tau(t) = \sum_{s=0}^{t-1} 2\eta_s \frac{1}{N} \lambda_{\min,V_\parallel}(XX^\mathsf{T})$ and the ratio $\frac{b(t)}{a(t)}$ in (C.3) is by (F.5) bounded uniformly by $\frac{b(t)}{a(t)} \leq C \cdot \frac{\eta_t}{B_t}$ for a constant $C > 0$. Thus, keeping $\frac{b(t)}{a(t)}$ fixed as a function of τ suggests the "linear scaling" $\eta_t \propto B_t$ used empirically in (13) and proposed via an heuristic SDE limit in (27).

D Analysis of Noising Augmentations

In this section, we give a full analysis of the noising augmentations presented in Section 4. Let us briefly recall the notation. As before, we consider overparameterized linear regression with loss

$$\mathcal{L}(W; \mathcal{D}) = \frac{1}{N} \|WX - Y\|_F^2,$$

where the dataset $\mathcal{D}$ of size N consists of data matrices X, Y that each have N columns $x_i \in \mathbb{R}^n, y_i \in \mathbb{R}^p$ with $n > N$. We optimize $\mathcal{L}(W; \mathcal{D})$ by augmented gradient descent or augmented stochastic gradient descent with additive Gaussian noise. This means that at each time t we replace $\mathcal{D} = (X, Y)$ by a random batch $\mathcal{D}_t = (X_t, Y)$ of size B_t, where the columns $x_{i,t}$ of X_t are

$$x_{i,t} = x_i + \sigma_t G_{i,t}, \qquad G_{i,t} \sim \mathcal{N}(0, 1) \text{ i.i.d.}$$

In the case of gradient descent, the batch consists of the entire dataset, and the resulting data matrices are

$$X_t = X + \sigma_t G_t \qquad \text{and} \qquad Y_t = Y.$$

In the case of stochastic gradient descent, the batch consists of B_t datapoints chosen uniformly at random with replacement, and the resulting data matrices are

$$X_t = c_t(XA_t + \sigma_t G_t) \qquad \text{and} \qquad Y_t = c_t Y A_t,$$

where $c_t = \sqrt{N/B_t}$, $A_t \in \mathbb{R}^{N \times B_t}$ has i.i.d. columns with a single non-zero entry equal to 1, and $G_t \in \mathbb{R}^{n \times B_t}$ has i.i.d. Gaussian entries. In both cases, the proxy loss is

$$\overline{\mathcal{L}}_t(W) := \frac{1}{N}\|Y - WX\|_F^2 + \sigma_t^2 \|W\|_F^2,$$

which has ridge minimizer

$$W_t^* = YX^\mathsf{T}(XX^\mathsf{T} + \sigma_t^2 N \cdot \mathrm{Id}_{n \times n})^{-1},$$

and the space $V_\parallel :=$ column span of $\mathbb{E}[X_t X_t^\mathsf{T}]$ is all of $\mathbb{R}^n$. We now separately analyze the cases of GD and SGD in Theorems 4.1 and Theorem 4.2, respectively.

D.1 Proof of Theorem 4.1 for GD

We begin by proving convergence without rates. For this, we seek to show that if $\sigma_t^2, \eta_t \to 0$ with σ_t^2 non-increasing and

$$\sum_{t=0}^{\infty} \eta_t \sigma_t^2 = \infty \qquad \text{and} \qquad \sum_{t=0}^{\infty} \eta_t^2 \sigma_t^2 < \infty, \tag{D.1}$$

then, $W_t \overset{p}{\to} W_{\text{min}}$. We will do this by applying Theorem 3.1, so we check that our assumptions imply the hypotheses of these theorems. For Theorem 3.1, we directly compute

$$\mathbb{E}[Y_t X_t^\mathsf{T}] = YX^\mathsf{T} \qquad \text{and} \qquad \mathbb{E}[X_t X_t^\mathsf{T}] = XX^\mathsf{T} + \sigma_t^2 N \cdot \mathrm{Id}_{n \times n}$$

and

$$\mathbb{E}[X_t X_t^\mathsf{T} X_t] = XX^\mathsf{T}X + \sigma_t^2(N + n + 1)X$$

$$\mathbb{E}[X_t X_t^\mathsf{T} X_t X_t^\mathsf{T}] = XX^\mathsf{T}XX^\mathsf{T} + \sigma_t^2\Big((2N + n + 2)XX^\mathsf{T} + \mathrm{tr}(XX^\mathsf{T})\,\mathrm{Id}_{n \times n}\Big) + \sigma_t^4 N(N + n + 1)\,\mathrm{Id}_{n \times n}.$$

We also find that

$$\begin{aligned}\|\Xi_t^*\|_F &= |\sigma_t^2 - \sigma_{t+1}^2| N \left\| Y X^\mathsf{T} \Big(XX^\mathsf{T} + \sigma_t^2 N \cdot \mathrm{Id}_{n\times n}\Big)^{-1} \Big(XX^\mathsf{T} + \sigma_{t+1}^2 N \cdot \mathrm{Id}_{n\times n}\Big)^{-1} \right\|_F \\ &\leq |\sigma_t^2 - \sigma_{t+1}^2| N \|YX^\mathsf{T}[(XX^\mathsf{T})^+]^2\|_F.\end{aligned}$$

Thus, because σ_t^2 is decreasing, we see that the hypothesis (3.7) of Theorem 3.1 indeed holds. Further, we note that

$$\begin{aligned}&\sum_{t=0}^{\infty} \eta_t^2 \mathbb{E}\left[\|X_tX_t^\mathsf{T} - \mathbb{E}[X_tX_t^\mathsf{T}]\|_F^2 + \|Y_tX_t^\mathsf{T} - \mathbb{E}[Y_tX_t^\mathsf{T}]\|_F^2\right] \\ &\qquad\qquad = \sum_{t=0}^{\infty} \eta_t^2\sigma_t^2 \Big(2(n+1)\|X\|_F^2 + N\|Y\|_F^2 + \sigma_t^2 Nn(n+1)\Big) = O\left(\sum_{t=0}^{\infty} \eta_t^2\sigma_t^2\right),\end{aligned}$$

which by (D.1) implies (B.3). Theorem 3.1 and the fact that $\lim_{t\to\infty} W_t^* = W_{\text{min}}$ therefore yield that $W_t \overset{p}{\to} W_{\text{min}}$.

We now prove convergence with rates, we aim to show that if $\eta_t = \Theta(t^{-x})$ and $\sigma_t^2 = \Theta(t^{-y})$ with $x, y > 0$, $x + y < 1$, and $2x + y > 1$, then for any $\varepsilon > 0$, we have that

$$t^{\min\{y, \frac{1}{2}x\} - \varepsilon} \|W_t - W_{\text{min}}\|_F \overset{p}{\to} 0.$$

We now check the hypotheses for and apply Theorem B.4. For (B.6), notice that $Y_r = \mathrm{Id} - 2\eta_r \frac{X_rX_r^\mathsf{T}}{N}$ satisfies the hypotheses of Theorem A.6 with $a_r = 1 - 2\eta_r\sigma_r^2$ and $b_r^2 = \frac{\eta_r^2\sigma_r^2}{a_r^2}\Big(2(n+1)\|X\|_F^2 + \sigma_r^2 Nn(n+1)\Big)$. Thus, by Theorem A.6 and the fact that $\eta_t = \Theta(t^{-x})$ and $\sigma_t^2 = \Theta(t^{-y})$, we find for some $C_1, C_2 > 0$ that

$$\log \mathbb{E}\left[\left\|\prod_{r=s}^{t}(\mathrm{Id} - 2\eta_r \frac{X_rX_r^\mathsf{T}}{N})\right\|_2^2\right] \leq \sum_{r=s}^{t} b_r^2 + 2\sum_{r=s}^{t} \log(1 - 2\eta_r\sigma_r^2) \leq C_1 - C_2 \int_s^{t+1} r^{-x-y} dr.$$

For (B.7), we find that

$$\|\Xi_t^*\|_F \leq |\sigma_t^2 - \sigma_{t+1}^2| N \|YX^\mathsf{T}[(XX^\mathsf{T})^+]^2\|_F = O(t^{-y-1}).$$

Finally, for (B.8), we find that

$$\begin{aligned}&\eta_t^2 \mathrm{tr}\left[\mathrm{Id} \circ \mathrm{Var}\Big(\mathbb{E}[W_t]X_tX_t^\mathsf{T} - Y_tX_t^\mathsf{T}\Big)\right] = \eta_t^2 \mathbb{E}\left[\|\mathbb{E}[W_t](X_tX_t^\mathsf{T} - \mathbb{E}[X_tX_t^\mathsf{T}]) - (Y_tX_t^\mathsf{T} - \mathbb{E}[Y_tX_t^\mathsf{T}])\|_F^2\right] \\ &\quad \leq 2\eta_t^2 \mathbb{E}\left[\|\mathbb{E}[W_t](X_tX_t^\mathsf{T} - \mathbb{E}[X_tX_t^\mathsf{T}])\|_F^2 + \|Y_tX_t^\mathsf{T} - \mathbb{E}[Y_tX_t^\mathsf{T}]\|_F^2\right] \\ &\quad \leq 2\eta_t^2 \Big(\|\mathbb{E}[W_t]\|_F^2 \mathbb{E}[\|X_tX_t^\mathsf{T} - \mathbb{E}[X_tX_t^\mathsf{T}]\|_F^2] + \|Y_tX_t^\mathsf{T} - \mathbb{E}[Y_tX_t^\mathsf{T}]\|_F^2\Big) \\ &\quad = O(t^{-2x-y}).\end{aligned}$$

Noting finally that $\|W_t^* - W_{\text{min}}\|_F = O(\sigma_t^2) = O(t^{-y})$, we apply Theorem B.4 with $\alpha = x + y$, $\beta_1 = y + 1$, and $\beta_2 = 2x + y$ to obtain the desired estimates. This concludes the proof of Theorem 4.1. □

D.2 Proof of Theorem 4.2 for SGD

We now prove Theorem 4.2 for SGD. As before, we will apply Theorems 3.1 and B.4. To check the hypotheses of these theorems, we will use the following expressions for moments of the augmented data matrices.

Lemma D.1. *We have*

$$\mathbb{E}[Y_tX_t^\mathsf{T}] = YX^\mathsf{T} \qquad \textit{and} \qquad \mathbb{E}[X_tX_t^\mathsf{T}] = XX^\mathsf{T} + \sigma_t^2 N\, \mathrm{Id}_{n\times n}\,. \tag{D.2}$$

Moreover,

$$\begin{aligned}
\mathbb{E}[Y_t X_t^\mathsf{T} X_t Y_t^\mathsf{T}] &= c_t^4 \mathbb{E}[Y A_t A_t^\mathsf{T} X^\mathsf{T} X A_t A_t^\mathsf{T} Y^\mathsf{T} + \sigma_t^2 Y A_t G_t^\mathsf{T} G_t A_t^\mathsf{T} Y^\mathsf{T}] \\
&= \frac{N}{B_t} Y \operatorname{diag}(X^\mathsf{T} X) Y^\mathsf{T} + \frac{B_t - 1}{B_t} Y X^\mathsf{T} X Y^\mathsf{T} + \sigma_t^2 N Y Y^\mathsf{T} \\
\mathbb{E}[Y_t X_t^\mathsf{T} X_t X_t^\mathsf{T}] &= c_t^4 \mathbb{E}[Y A_t A_t^\mathsf{T} X^\mathsf{T} X A_t A_t^\mathsf{T} X^\mathsf{T} + \sigma_t^2 Y A_t G_t^\mathsf{T} G_t A_t^\mathsf{T} X^\mathsf{T} \\
&\quad + \sigma_t^2 Y A_t G_t^\mathsf{T} X A_t G_t^\mathsf{T} + \sigma_t^2 Y A_t A_t^\mathsf{T} X^\mathsf{T} G_t G_t^\mathsf{T}] \\
&= \frac{N}{B_t} Y \operatorname{diag}(X^\mathsf{T} X) X^\mathsf{T} + \frac{B_t - 1}{B_t} Y X^\mathsf{T} X X^\mathsf{T} + \sigma_t^2 (N + \frac{n+1}{B_t/N}) Y X^\mathsf{T} \\
\mathbb{E}[X_t X_t^\mathsf{T} X_t X_t^\mathsf{T}] &= c_t^4 \mathbb{E}[X A_t A_t^\mathsf{T} X^\mathsf{T} X A_t A_t^\mathsf{T} X^\mathsf{T} + \sigma_t^2 G_t G_t^\mathsf{T} X A_t A_t^\mathsf{T} X^\mathsf{T} + \sigma_t^2 X A_t G_t^\mathsf{T} G_t A_t^\mathsf{T} X^\mathsf{T} \\
&\quad + \sigma_t^2 X A_t A_t^\mathsf{T} X^\mathsf{T} G_t G_t^\mathsf{T} + \sigma_t^2 G_t A_t^\mathsf{T} X^\mathsf{T} G_t A_t^\mathsf{T} X^\mathsf{T} + \sigma_t^2 X A_t G_t^\mathsf{T} X A_t G_t^\mathsf{T} \\
&\quad + \sigma_t^2 G_t A_t^\mathsf{T} X^\mathsf{T} X A_t G_t^\mathsf{T} + \sigma_t^4 G_t G_t^\mathsf{T} G_t G_t^\mathsf{T}] \\
&= \frac{N}{B_t} X \operatorname{diag}(X^\mathsf{T} X) X^\mathsf{T} + \frac{B_t - 1}{B_t} X X^\mathsf{T} X X^\mathsf{T} + \sigma_t^2 (2N + \frac{n+2}{B_t/N}) X X^\mathsf{T} \\
&\quad + \sigma_t^2 \frac{N}{B_t} tr(X X^\mathsf{T}) \operatorname{Id}_{n \times n} + \sigma_t^4 N (N + \frac{n+1}{B_t/N}) \operatorname{Id}_{n \times n}.
\end{aligned}$$

Proof. All these formulas are obtained by direct, if slightly tedious, computation. □

With these expressions in hand, we now check the conditions of Theorem 3.1. First, by the Sherman-Morrison-Woodbury matrix inversion formula, we have

$$\begin{aligned}
\|\Xi_t^*\|_F &= |\sigma_t^2 N - \sigma_{t+1}^2 N| \left\| Y X^\mathsf{T} (X X^\mathsf{T} + \sigma_t^2 N \cdot \operatorname{Id}_{n \times n})^{-1} (X X^\mathsf{T} + \sigma_{t+1}^2 N \cdot \operatorname{Id}_{n \times n})^{-1} \right\|_F \\
&\leq N |\sigma_t^2 - \sigma_{t+1}^2| \left\| Y X^\mathsf{T} [(X X^\mathsf{T})^+]^2 \right\|_F .
\end{aligned} \tag{D.3}$$

Because σ_t^2 is non-increasing, this implies (3.7). Next, by Lemma D.1 we have that

$$\sum_{t=0}^{\infty} \eta_t \lambda_{\min, V_\parallel} (\mathbb{E}[X_t X_t^\mathsf{T}]) \geq \sum_{t=0}^{\infty} \eta_t \sigma_t^2 = \infty$$

by the first given condition in (4.7), which verifies (3.8). Finally, by Lemma D.1, we may compute

$$\begin{aligned}
&\mathbb{E}\left[\left\| X_t X_t^\mathsf{T} - \mathbb{E}\left[X_t X_t^\mathsf{T} \right] \right\|_F^2 \right] \\
&\qquad = \frac{1}{B_t} \operatorname{tr}\Big(X (N \operatorname{diag}(X^\mathsf{T} X) - X^\mathsf{T} X) X^\mathsf{T} \Big) + 2 \sigma_t^2 \frac{n+1}{B_t/N} \operatorname{tr}(X X^\mathsf{T}) + \sigma_t^4 \frac{N n (n+1)}{B_t/N}
\end{aligned}$$

and

$$\mathbb{E}\left[\| Y_t X_t^\mathsf{T} - \mathbb{E}[Y_t X_t^\mathsf{T}] \|_F^2 \right] = \frac{1}{B_t} \operatorname{tr}\Big(Y (N \operatorname{diag}(X^\mathsf{T} X) - X^\mathsf{T} X) Y^\mathsf{T} \Big) + \sigma_t^2 N \operatorname{tr}(Y Y^\mathsf{T}).$$

Together, these imply that for some constant $C > 0$, we have that

$$\sum_{t=0}^{\infty} \eta_t^2 \left[\mathbb{E}\left[\left\| X_t X_t^\mathsf{T} - \mathbb{E}\left[X_t X_t^\mathsf{T} \right] \right\|_F^2 \right] + \mathbb{E}\left[\left\| Y_t X_t^\mathsf{T} - \mathbb{E}[Y_t X_t^\mathsf{T}] \right\|_F^2 \right] \right] \leq \sum_{t=0}^{\infty} C \eta_t^2 < \infty$$

by the second given condition in (4.7), which verifies (3.9). Thus the conditions of Theorem 3.1 apply, which shows that $W_t \xrightarrow{p} W_{\text{min}}$, as desired.

For rates of convergence, we check the conditions of Theorem B.4. For (B.6), we will apply Theorem A.6 to bound $\log \mathbb{E} \left\| \prod_{r=s}^{t} \left(\operatorname{Id} - \frac{2\eta_r}{N} X_r X_r^\mathsf{T} \right) \right\|_2^2$. By the second moment expression $\mathbb{E}[X_r X_r^\mathsf{T}] = X X^\mathsf{T} + \sigma_r^2 N \operatorname{Id}_{n \times n}$, we find that

$$\left\| \mathbb{E}\left[\operatorname{Id} - \frac{2\eta_r}{N} X_r X_r^\mathsf{T} \right] \right\|_2 = 1 - 2 \eta_r \sigma_r^2.$$

Moreover, by Lemma D.1, for some constant $C > 0$ we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathrm{Id} - \frac{2\eta_r}{N} X_r X_r^\intercal - \mathbb{E}\left[\mathrm{Id} - \frac{2\eta_r}{N} X_r X_r^\intercal\right]\right\|_2^2\right] \\
&= \frac{4\eta_r^2}{N^2}\mathbb{E}\left[\left\|X_r X_r^\intercal - \mathbb{E}\left[X_r X_r^\intercal\right]\right\|_2^2\right] \\
&\le \frac{4\eta_r^2}{N^2}\left[\frac{1}{B_t}\mathrm{tr}\Big(X(N\operatorname{diag}(X^\intercal X) - X^\intercal X)X^\intercal\Big) + 2\sigma_t^2\frac{n+1}{B_t/N}\mathrm{tr}(XX^\intercal) + \sigma_t^4\frac{Nn(n+1)}{B_t/N}\right] \\
&\le C\eta_r^2.
\end{aligned}$$

Applying Theorem A.6 with $a_r = 1 - 2\eta_r\sigma_r^2$ and $b_r^2 = \frac{C\eta_r^2}{a_r^2}$, we find that

$$\begin{aligned}
\log\mathbb{E}\left\|\prod_{r=s}^{t}\left(\mathrm{Id} - \frac{2\eta_r}{N}X_r X_r^\intercal\right)\right\|_2^2 &\le \sum_{r=s}^{t}\frac{C\eta_r^2}{a_r^2} + 2\log\left(\prod_{r=s}^{t}\left(1 - 2\eta_r\sigma_r^2\right)\right) \\
&\le \sum_{r=s}^{t}\frac{C\eta_r^2}{a_r^2} - 4\sum_{r=s}^{t}\eta_r\sigma_r^2.
\end{aligned}$$

Because $\eta_r = \Theta(r^{-x})$ and $\sigma_r^2 = \Theta(r^{-y})$, we obtain (B.6) with $\alpha = x + y$, $C_1 = \sum_{r=0}^{\infty}\frac{C\eta_r^2}{a_r^2}$, and some $C_2 > 0$, where C_1 is finite because $x > \frac{1}{2}$. Next, (B.7) holds for $\beta_1 = -y - 1$ by (D.3). Finally, it remains to bound

$$\eta_t^2\mathrm{tr}\left[\mathrm{Id}\circ\mathrm{Var}(\mathbb{E}[W_t]X_tX_t^\intercal - Y_tX_t^\intercal\Big)\right]$$

to verify (B.8). Again using D.1 and noting that $W_{\min}X = Y$, we find

$$\begin{aligned}
&\eta_t^2\mathrm{tr}\left[\mathrm{Id}\circ\mathrm{Var}(\mathbb{E}[W_t]X_tX_t^\intercal - Y_tX_t^\intercal)\right] \\
&\quad= \eta_t^2\mathrm{tr}\Big(\frac{1}{B_t}\mathbb{E}[W_t]X(N\operatorname{diag}(X^\intercal X) - X^\intercal X)X^\intercal\mathbb{E}[W_t]^\intercal \\
&\qquad + 2\sigma_t^2\frac{n+1}{B_t/N}\mathbb{E}[W_t]XX^\intercal\mathbb{E}[W_t]^\intercal + (\sigma_t^2\frac{N}{B_t}\mathrm{tr}(XX^\intercal) + \sigma_t^4 N\frac{n+1}{B_t/N})\mathbb{E}[W_t]\mathbb{E}[W_t]^\intercal\Big) \\
&\qquad - 2\eta_t^2\mathrm{tr}\Big(\frac{1}{B_t}Y(N\operatorname{diag}(X^\intercal X) - X^\intercal X)X^\intercal\mathbb{E}[W_t]^\intercal + \sigma_t^2\frac{n+1}{B_t/N}YX^\intercal\mathbb{E}[W_t]^\intercal\Big) \\
&\qquad + \eta_t^2\mathrm{tr}\Big(\frac{1}{B_t}Y(N\operatorname{diag}(X^\intercal X) - X^\intercal X)Y^\intercal + \sigma_t^2 NYY^\intercal\Big) \\
&\quad\le C\eta_t^2(\sigma_t^2 + \|\Delta_t\|_F^2)
\end{aligned}$$

for some $C > 0$ and $\Delta_t := \mathbb{E}[W_t] - W_{\text{min}}$. Define also $\Delta_t' := \mathbb{E}[W_t] - W_t^*$ so that, exactly as in Proposition B.6, we have

$$\Delta_{t+1}' = \Delta_t'\left(\mathrm{Id} - \frac{2\eta_t}{N}\mathbb{E}\left[X_tX_t^\intercal\right]\right) + \frac{2}{N}\Xi_t^*, \qquad \Delta_t' := \mathbb{E}\left[W_t - W_t^*\right].$$

Since $\|\Xi_t^*\|_F = O(t^{-y-1})$ and we already saw that

$$\left\|\mathrm{Id} - \frac{2\eta_t}{N}\mathbb{E}\left[X_tX_t^\intercal\right]\right\|_2 = 1 - 2\eta_t\sigma_t^2,$$

we may use the single sided decay estimates of Lemma A.4 to conclude that $\|\Delta_t'\|_F = O(t^{x-1})$. This implies that

$$\|\Delta_t\|_F \le \|\Delta_t'\|_F + \|W_t^* - W_{\min}\|_F = O(t^{x-1}) + \Theta(t^{-y}) = \Theta(t^{-y})$$

since we assumed that $x + y < 1$. Therefore, we obtain

$$\eta_t^2\mathrm{tr}\left[\mathrm{Id}\circ\mathrm{Var}(\mathbb{E}[W_t]X_tX_t^\intercal - Y_tX_t^\intercal)\right] \le C\eta_t^2(\sigma_t^2 + \Theta(t^{-2y})) = \Theta(t^{-2x-y}),$$

showing that condition (B.8) holds with $\beta_2 = 2x + y$. We have thus verified all of the conditions of Theorem B.4, whose application completes the proof. □

E Analysis of random projection augmentations

In this section, we give a full analysis of the random projection augmentations presented in Section 5. In this setting, we have a preconditioning matrix $\Pi_t = Q_t \widetilde{\Pi}_t Q_t^\mathsf{T} \in \mathbb{R}^{n \times n}$, where $\widetilde{\Pi}_t$ is a projection matrix and Q_t are Haar random orthogonal matrices. Define the normalized trace of the projection matrix by

$$\gamma_t := \frac{\text{tr}(\Pi_t)}{n}. \tag{E.1}$$

We consider the augmentation given by

$$X_t = \Pi_t X \qquad \text{and} \qquad Y_t = Y.$$

In this setting, we first record the values of the lower order moments of the augmented data matrices, with proofs deferred to Section E.1

Lemma E.1. *We have that*

$$\mathbb{E}[X_t] = \gamma_t X \tag{E.2}$$

$$\mathbb{E}[Y_t X_t^\mathsf{T}] = \gamma_t Y X^\mathsf{T} \tag{E.3}$$

$$\mathbb{E}[X_t X_t^\mathsf{T}] = \frac{\gamma_t(\gamma_t + 1/n - 2/n^2)}{1 + 1/n - 2/n^2} X X^\mathsf{T} + \frac{\gamma_t(1-\gamma_t)}{n(1 + 1/n - 2/n^2)} \|X\|_F^2 \,\text{Id} \tag{E.4}$$

$$\mathbb{E}[X_t X_t^\mathsf{T} X_t] = \frac{\gamma_t^2 + \gamma_t(1/n - 2/n^2)}{1 + 1/n - 2/n^2} X X^\mathsf{T} X + \frac{\gamma_t(1-\gamma_t)}{1 + 1/n - 2/n^2} \frac{\|X\|_F^2}{n} X \tag{E.5}$$

$$\mathbb{E}[\|X_t X_t^\mathsf{T}\|_F^2] = \frac{(\gamma_t^2 + \gamma_t(1/n - 2/n^2))\|X X^\mathsf{T}\|_F^2 + \gamma_t(1-\gamma_t)\frac{1}{n}\|X\|_F^4}{1 + 1/n - 2/n^2}. \tag{E.6}$$

We now compute the proxy loss and its optima as follows.

Lemma E.2. *The proxy loss for randomly rotated projections with normalized trace γ_t given by (E.1) is*

$$\overline{\mathcal{L}}_t(W) = \frac{1}{N}\|Y - \gamma_t W X\|_F^2 + \frac{1}{N}\gamma_t(1-\gamma_t)\|X\|_F^2\|W\|_F^2 + \frac{\gamma_t(1-\gamma_t)}{N}\frac{1/n - 2/n^2}{1 + 1/n - 2/n^2}(\|WX\|_F^2 + \frac{1}{n}\|X\|_F^2\|W\|_F^2).$$

It has proxy optima

$$W_t^* = Y X^\mathsf{T}\Big(\frac{\gamma_t + 1/n - 2/n^2}{1 + 1/n - 2/n^2} X X^\mathsf{T} + \frac{1-\gamma_t}{1 + 1/n - 2/n^2}\frac{\|X\|_F^2}{n}\,\text{Id}\Big)^{-1}.$$

Proof. Applying (3.1), we find that

$$\overline{\mathcal{L}}_t(W) = \frac{1}{N}\mathbb{E}[\|Y_t - W X_t\|_F^2] = \frac{1}{N}\mathbb{E}[\|Y_t\|_F^2] - \frac{2}{N}\mathbb{E}[\text{tr}(X_t^\mathsf{T} W^\mathsf{T} Y_t)] + \frac{1}{N}\mathbb{E}[\text{tr}(W^\mathsf{T} W X_t X_t^\mathsf{T})].$$

Applying Lemma E.1, we conclude that

$$\begin{aligned}
\overline{\mathcal{L}}_t(W) &= \frac{1}{N}\|Y\|_F^2 - \frac{2\gamma_t}{N}\text{tr}(X W^\mathsf{T} Y) + \frac{1}{N}\frac{\gamma_t(\gamma_t + 1/n - 2/n^2)}{1 + 1/n - 2/n^2}\|WX\|_F^2 + \frac{1}{N}\frac{\gamma_t(1-\gamma_t)}{n(1 + 1/n - 2/n^2)}\|X\|_F^2\|W\|_F^2 \\
&= \frac{1}{N}\|Y - \gamma_t W X\|_F^2 + \frac{1}{N}\gamma_t(1-\gamma_t)\|X\|_F^2\|W\|_F^2 + \frac{\gamma_t(1-\gamma_t)}{N}\frac{1/n - 2/n^2}{1 + 1/n - 2/n^2}(\|WX\|_F^2 + \frac{1}{n}\|X\|_F^2\|W\|_F^2).
\end{aligned}$$

The formula for W_t^* then results from Lemma E.1 and the general formula (3.4). □

E.1 Proof of Lemma E.1

We apply the Weingarten calculus (7) to compute integrals of polynomial functions of the matrix entries of a Haar orthogonal matrix. Because each matrix entry of the expectations in Lemma E.1 is such a polynomial, this will allow us to compute the relevant expectations. The main result we use is the following on polynomials of degree at most 4 and its corollary.

Proposition E.3 (Corollary 3.4 of (7))**.** *For an $n \times n$ orthogonal matrix Q drawn from the Haar measure, we have that*

$$
\begin{aligned}
\mathbb{E}[Q_{i_1j_1}Q_{i_2j_2}] &= \frac{1}{n}\delta_{i_1i_2}\delta_{j_1j_2} && \text{(E.7)}\\
\mathbb{E}[Q_{i_1j_1}Q_{i_2j_2}Q_{i_3j_3}Q_{i_4j_4}] &= \frac{-1}{n(n-1)(n+2)}\Big(\delta_{i_1i_2}\delta_{i_3i_4}+\delta_{i_1i_3}\delta_{i_2i_4}+\delta_{i_1i_4}\delta_{i_2i_3}\Big)\Big(\delta_{j_1j_2}\delta_{j_3j_4}+\delta_{j_1j_3}\delta_{j_2j_4}+\delta_{j_1j_4}\delta_{j_2j_3}\Big)\\
&\quad+\frac{1}{n(n-1)}\Big(\delta_{i_1i_2}\delta_{i_3i_4}\delta_{j_1j_2}\delta_{j_3j_4}+\delta_{i_1i_3}\delta_{i_2i_4}\delta_{j_1j_3}\delta_{j_2j_4}+\delta_{i_1i_4}\delta_{i_2i_3}\delta_{j_1j_4}\delta_{j_2j_3}\Big), && \text{(E.8)}
\end{aligned}
$$

where δ_{ab} denotes the Kronecker delta function $\delta_{ab} = \mathbf{1}\{a = b\}$.

Corollary E.4. *For matrices $A, B, C \in \mathbb{R}^{n\times n}$ and an $n \times n$ Haar random orthogonal matrix Q, we have*

$$
\begin{aligned}
\mathbb{E}[QAQ^\top] &= \frac{1}{n}tr(A)\cdot\mathrm{Id} && \text{(E.9)}\\
\mathbb{E}[QAQ^\top BQCQ^\top]_{ij} &= -\frac{1}{n(n-1)(n+2)}\Big(tr(A)tr(C)+tr(AC^\top+AC)\Big)\Big(\delta_{ij}tr(B)+B_{ij}+B_{ji}\Big)\\
&\quad+\frac{1}{n(n-1)}\Big(tr(A)tr(C)B_{ij}+tr(AC^\top)B_{ji}+\delta_{ij}tr(AC)tr(B)\Big). && \text{(E.10)}
\end{aligned}
$$

Proof. For (E.9), notice by (E.7) that

$$
\mathbb{E}[QAQ^\top]_{ij} = \sum_{a,b=1}^{n}\mathbb{E}[Q_{ia}A_{ab}Q_{jb}] = \sum_{a,b=1}^{n}\delta_{ab}\delta_{ij}\frac{1}{n}A_{ab} = \delta_{ij}\frac{1}{n}\mathrm{tr}(A).
$$

For (E.10), notice by (E.8) that

$$
\begin{aligned}
\mathbb{E}[QAQ^\top BQCQ^\top]_{ij} &= \sum_{a,b,c,d,e,f=1}^{n}\mathbb{E}[Q_{ia}A_{ab}Q_{cb}B_{cd}Q_{de}C_{ef}Q_{jf}]\\
&= -\frac{1}{n(n-1)(n+2)}\Big(\sum_{a,e=1}^{n}A_{aa}C_{ee}+\sum_{a,b=1}^{n}A_{ab}C_{ab}+\sum_{a,b=1}^{n}A_{ab}C_{ba}\Big)\Big(\sum_{c=1}^{n}\delta_{ij}B_{cc}+B_{ij}+B_{ji}\Big)\\
&\quad+\frac{1}{n(n-1)}\Big(\sum_{a,e=1}^{n}A_{aa}C_{ee}B_{ij}+\sum_{a,b=1}^{n}A_{ab}C_{ab}B_{ji}+\delta_{ij}\sum_{a,b=1}^{n}A_{ab}C_{ba}\sum_{c=1}^{n}B_{cc}\Big),
\end{aligned}
$$

which when simplified gives the desired conclusion. □

We now compute each lower order moment; in each computation, let Q denote a random $n \times n$ orthogonal matrix drawn from the Haar measure. Claims (E.2) and (E.3) follow from Corollary E.4 and the facts that

$$
\mathbb{E}[X_t] = \mathbb{E}[Q\widetilde{\Pi}_tQ^\top]X \qquad \text{and} \qquad \mathbb{E}[Y_tX_t^\top] = Y\mathbb{E}[Q\widetilde{\Pi}_tQ^\top]X.
$$

Claims (E.4) and (E.5) follow from Corollary E.4 and the facts that

$$
\mathbb{E}[X_tX_t^\top] = \mathbb{E}[Q\widetilde{\Pi}_tQ^\top XX^\top Q\widetilde{\Pi}_tQ^\top] \qquad \text{and} \qquad \mathbb{E}[X_tX_t^\top X_t] = \mathbb{E}[Q\widetilde{\Pi}_tQ^\top XX^\top Q\widetilde{\Pi}_t^2Q^\top]X.
$$

Finally, (E.6) follows from Corollary E.4 and the fact that

$$
\mathbb{E}[\|X_tX_t^\top]_F^2] = \mathbb{E}[\mathrm{tr}(Q\widetilde{\Pi}_tQ^\top XX^\top Q\widetilde{\Pi}_t^2Q^\top XX^\top Q\widetilde{\Pi}_tQ^\top)] = \mathbb{E}[\mathrm{tr}(\widetilde{\Pi}_tQ^\top XX^\top Q\widetilde{\Pi}_t^2Q^\top XX^\top Q\widetilde{\Pi}_t)]].
$$

This completes the proof of Lemma E.1.

E.2 Proof of Theorem 5.1

We first show convergence. By Lemma E.1, we find that $V_{\|} = \mathbb{R}^n$. Furthermore, because $\gamma_t \to 1$, we have that $W_\infty^* = \lim_{t\to\infty} W_t^* = W_{\min}$. It therefore suffices to verify the conditions of Theorem 3.1. First, notice that

$$\begin{aligned}\|\Xi_t^*\|_F = \frac{|\gamma_t - \gamma_{t+1}|}{1 + 1/n - 2/n^2} \Bigg\| YX^{\mathsf{T}} \Big(\frac{\gamma_{t+1} + 1/n - 2/n^2}{1 + 1/n - 2/n^2} XX^{\mathsf{T}} + \frac{1 - \gamma_{t+1}}{1 + 1/n - 2/n^2} \frac{\|X\|_F^2}{n} \operatorname{Id}\Big)^{-1} \\ \Big(XX^{\mathsf{T}} - \frac{\|X\|_F^2}{n} \operatorname{Id}\Big)\Big(\frac{\gamma_t + 1/n - 2/n^2}{1 + 1/n - 2/n^2} XX^{\mathsf{T}} + \frac{1 - \gamma_t}{1 + 1/n - 2/n^2} \frac{\|X\|_F^2}{n} \operatorname{Id}\Big)^{-1} \Bigg\|_F \\ \leq \frac{|\gamma_t^{-1} - \gamma_{t+1}^{-1}|}{1 + 1/n - 2/n^2} \|YX^{\mathsf{T}}\|_F \|[XX^{\mathsf{T}}]^+\|_F^2 \|XX^{\mathsf{T}} - \frac{1}{n}\|X\|_F^2 \operatorname{Id}\|_F.\end{aligned}$$

Because γ_t are increasing and $\lim_{t\to\infty} \gamma_t = 1$, we find that (3.7) holds. Now, by Lemma E.1, we have

$$\lambda_{\min}(\mathbb{E}[X_t X_t^{\mathsf{T}}]) \geq \frac{\gamma_t(1 - \gamma_t)}{1 + 1/n - 2/n^2} \frac{\|X\|_F^2}{n},$$

so the first condition in (5.2) implies (3.8). Finally, using Lemma E.1 we may compute

$$\begin{aligned}\mathbb{E}[\|X_t X_t^{\mathsf{T}} - \mathbb{E}[X_t X_t^{\mathsf{T}}]\|_F^2] &= \mathbb{E}[\|X_t X_t\|_F^2] - \|\mathbb{E}[X_t X_t^{\mathsf{T}}]\|_F^2 \\ &= \frac{\gamma_t(1 - \gamma_t)(\gamma_t(1 + \gamma_t)n^4 + (1 + 2\gamma_t)n^3 - (1 + 4\gamma_t)n^2 - 4n + 4)}{(n-1)^2(n+2)^2} \|XX^{\mathsf{T}}\|_F^2 \\ &\quad + \frac{\gamma_t(1 - \gamma_t)((1 - \gamma_t - \gamma_t^2)n^4 + (1 - 2\gamma_t)n^3 + (4\gamma_t - 2)n^2)}{(n-1)^2(n+2)^2} \frac{1}{n} \|X\|_F^4 \\ &\leq 2\gamma_t(1 - \gamma_t)\Big(\|XX^{\mathsf{T}}\|_F^2 + \frac{1}{n}\|X\|_F^4\Big) \end{aligned} \tag{E.11}$$

and also

$$\begin{aligned}\mathbb{E}[\|Y_t X_t^{\mathsf{T}} - \mathbb{E}[Y_t X_t^{\mathsf{T}}]\|_F^2] &= \mathbb{E}[\|Y_t X_t^{\mathsf{T}}\|_F^2] - \|\mathbb{E}[Y_t X_t^{\mathsf{T}}]\|_F^2 \\ &= \frac{\gamma_t(1 - \gamma_t)(1/n - 2/n^2)}{1 + 1/n - 2/n^2} \|YX^{\mathsf{T}}\|_F^2 + \frac{\gamma_t(1 - \gamma_t)}{1 + 1/n - 2/n^2} \frac{1}{n} \|X\|_F^2 \|Y\|_F^2 \\ &\leq \gamma_t(1 - \gamma_t)\Big(\|YX^{\mathsf{T}}\|_F^2 + \frac{1}{n}\|X\|_F^2\|Y\|_F^2\Big). \end{aligned} \tag{E.12}$$

Combining these bounds with the second condition in (5.2) gives (3.9). Having verified (3.7), (3.8), and (3.9), we conclude by Theorem 3.1 that $W_t \xrightarrow{p} W_{\min}$ as desired.

We now obtain rates using Theorem B.4. We aim to show that if $\eta_t = \Theta(t^{-x})$ and $\gamma_t = 1 - \Theta(t^{-y})$ with $x, y > 0$, $x + y < 1$, and $2x + y > 1$, then for any $\varepsilon > 0$ we have that

$$t^{\min\{y, \frac{1}{2}x\} - \varepsilon} \|W_t - W_{\min}\|_F \xrightarrow{p} 0.$$

We now check the hypotheses of Theorem B.4. For (B.6), by Lemma E.1 and (E.11), $Y_r = \operatorname{Id} - \frac{2\eta_r}{N} X_r X_r^{\mathsf{T}}$ satisfies the hypotheses of Theorem A.6 with

$$a_r = 1 - \frac{2\eta_r}{N} \frac{\gamma_t(1 - \gamma_t)}{1 + 1/n - 2/n^2} \frac{\|X\|_F^2}{n} \qquad \text{and} \qquad b_r^2 = \frac{8\eta_r^2 \gamma_t(1 - \gamma_t)}{a_r^2 N^2} \Big(\|XX^{\mathsf{T}}\|_F^2 + \frac{1}{n}\|X\|_F^4\Big).$$

By Theorem A.6 and the fact that $\eta_t = \Theta(t^{-x})$ and $\gamma_t = 1 - \Theta(t^{-y})$, we find for some $C_1, C_2 > 0$ that

$$\begin{aligned}\log \mathbb{E}\left[\left\|\prod_{r=s}^{t}(\operatorname{Id} - \frac{2\eta_r}{N} X_r X_r^{\mathsf{T}})\right\|_2^2\right] &\leq \sum_{r=s}^{t} b_r^2 + 2\sum_{r=s}^{t} \log(1 - \frac{2\eta_r}{N(1 + 1/n - 2/n^2)} \gamma_t(1 - \gamma_t) \frac{\|X\|_F^2}{n}) \\ &\leq C_1 - C_2 \int_s^{t+1} r^{-x-y} dr.\end{aligned}$$

For (B.7), our previous computations show that

$$\|\Xi_t^*\|_F \leq \frac{|\gamma_t^{-1} - \gamma_{t+1}^{-1}|}{1 + 1/n - 2/n^2} \|YX^\intercal\|_F \|[XX^\intercal]^+\|_F^2 \|XX^\intercal - \frac{1}{n}\|X\|_F^2 \operatorname{Id}\|_F = O(t^{-y-1}).$$

Finally, for (B.8), we find by (E.11), (E.12) and the fact that $\|\mathbb{E}[W_t]\|_F = O(1)$ that

$$\begin{aligned}
\eta_t^2 \operatorname{tr}\left[\operatorname{Id} \circ \operatorname{Var}\Big(\mathbb{E}[W_t]X_tX_t^\intercal - Y_tX_t^\intercal\Big)\right] &= \eta_t^2 \mathbb{E}\Big[\|\mathbb{E}[W_t](X_tX_t^\intercal - \mathbb{E}[X_tX_t^\intercal]) - (Y_tX_t^\intercal - \mathbb{E}[Y_tX_t^\intercal])\|_F^2\Big] \\
&\leq 2\eta_t^2 \mathbb{E}\Big[\|\mathbb{E}[W_t](X_tX_t^\intercal - \mathbb{E}[X_tX_t^\intercal])\|_F^2 + \|Y_tX_t^\intercal - \mathbb{E}[Y_tX_t^\intercal]\|_F^2\Big] \\
&\leq 2\eta_t^2 \Big(\|\mathbb{E}[W_t]\|_F^2 \mathbb{E}[\|X_tX_t^\intercal - \mathbb{E}[X_tX_t^\intercal]\|_F^2] + \|Y_tX_t^\intercal - \mathbb{E}[Y_tX_t^\intercal]\|_F^2\Big) \\
&= O(t^{-2x-y}).
\end{aligned}$$

Noting finally that $\|W_t^* - W_{\text{min}}\|_F = O(1 - \gamma_t) = O(t^{-y})$, we apply Theorem B.4 with $\alpha = x + y$, $\beta_1 = y + 1$, and $\beta_2 = 2x + y$ to obtain the desired rate. This concludes the proof of Theorem 5.1.□

E.3 Experimental validation

To validate Theorem 5.1, we ran augmented GD with random projection augmentation on $N = 100$ simulated datapoints. Inputs were i.i.d. Gaussian vectors in dimension $n = 400$, and outputs in dim $p = 1$ were generated by a random linear map with i.i.d Gaussian coefficients drawn from $\mathcal{N}(1, 1)$. The learning rate followed a fixed polynomially decaying schedule $\eta_t = \frac{0.005}{100} \cdot (\text{batch size}) \cdot (1 + \frac{t}{20})^{-0.66}$. Figure E.1 shows MSE and $\|W_{t,\perp}\|_F$ along a single optimization trajectory with different schedules for the dimension ratio $1 - \gamma_t$ used in random projection augmentation. Code to generate this figure is provided in `supplement.zip` in the supplement. It ran in 30 minutes on a standard laptop CPU.

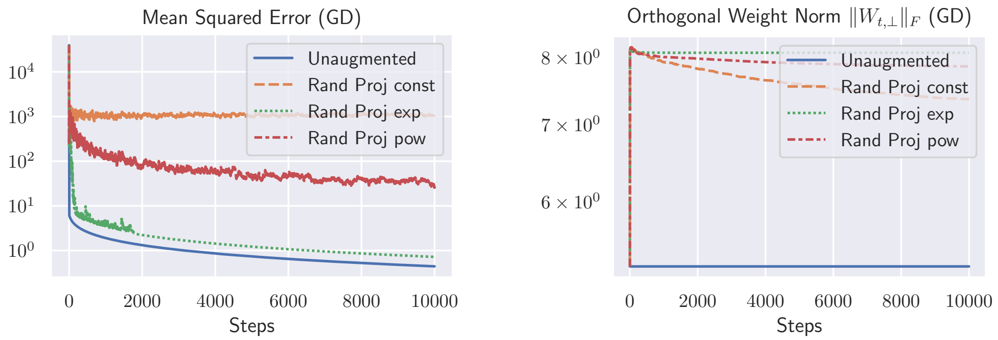


Figure E.1: MSE and $\|W_{t,\perp}\|_F$ for optimization trajectories of GD with random projection augmentation Jointly scheduling learning rate and noise variance to have polynomial decay is necessary for optimization to converge to the minimal norm solution W_{min}. Rand Proj const, Rand Proj exp, and Rand Proj pow have random projection augmentation with $1 - \gamma_t = 0.1, 0.1e^{-0.02t}, 0.1(1 + \frac{t}{50})^{-0.33}$, respectively. Other experimental details are the same as those described in §E.3.

F Analysis of SGD

This section gives an analysis of vanilla SGD using our method. Let us briefly recall the notation. As before, we consider overparameterized linear regression with loss

$$\mathcal{L}(W; \mathcal{D}) = \frac{1}{N} \|WX - Y\|_F^2,$$

where the dataset $\mathcal{D}$ of size N consists of data matrices X, Y that each have N columns $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}^p$ with $n > N$. We optimize $\mathcal{L}(W; \mathcal{D})$ by augmented SGD either with or without additive Gaussian noise. In the former case, this means that at each time t we replace $\mathcal{D} = (X, Y)$ by a random batch $\mathcal{B}_t = (X_t, Y_t)$ given by a prescribed batch size $B_t = |\mathcal{B}_t|$ in which each datapoint in $\mathcal{B}_t$ is chosen uniformly with replacement from $\mathcal{D}$, and the resulting data matrices X_t and Y_t are scaled so that $\overline{\mathcal{L}}_t(W) = \mathcal{L}(W; \mathcal{D})$. Concretely, this means that for the normalizing factor $c_t := \sqrt{N/B_t}$ we have

$$X_t = c_t X A_t \qquad \text{and} \qquad Y_t = c_t Y A_t, \tag{F.1}$$

where $A_t \in \mathbb{R}^{N \times B_t}$ has i.i.d. columns $A_{t,i}$ with a single non-zero entry equal to 1 chosen uniformly at random. In this setting the minimum norm optimum for each t are the same and given by

$$W_t^* = W_{\min} = Y X^{\mathsf{T}} (X X^{\mathsf{T}})^+,$$

which coincides with the minimum norm optimum for the unaugmented loss.

Theorem F.1. *If the learning rate satisfies $\eta_t \to 0$ and*

$$\sum_{t=0}^{\infty} \eta_t = \infty, \tag{F.2}$$

then for any initialization W_0, we have $W_t Q_\parallel \xrightarrow{p} W_{\min}$. If further we have that $\eta_t = \Theta(t^{-x})$ with $0 < x < 1$, then for some $C > 0$ we have

$$e^{Ct^{1-x}} \|W_t Q_\parallel - W_{\min}\|_F \xrightarrow{p} 0.$$

Theorem F.1 recovers the exponential convergence rate for SGD in the overparametrized settting, which has been previously studied through both empirical and theoretical means (22). Because $1 \leq B_t \leq N$ for all t, it does not affect the asymptotic results in Theorem F.1. In practice, however, the number of optimization steps t is often small enough that $\frac{B_t}{N}$ is of order $t^{-\alpha}$ for some $\alpha > 0$, meaning the choice of B_t can affect rates in this non-asymptotic regime. Though we do not attempt to push our generic analysis to this granularity, this is done in (22) to derive optimal batch sizes and learning rates in the overparametrized setting.

Proof of Theorem F.1. In order to apply Theorems B.1 and B.4, we begin by computing the moments of A_t as follows. Recall the notation $\mathrm{diag}(M)$ from Appendix A.1.

Lemma F.2. *For any $Z \in \mathbb{R}^{N \times N}$, we have that*

$$\mathbb{E}[A_t A_t^{\mathsf{T}}] = \frac{B_t}{N} \mathrm{Id}_{N \times N} \qquad \textit{and} \qquad \mathbb{E}[A_t A_t^{\mathsf{T}} Z A_t A_t^{\mathsf{T}}] = \frac{B_t}{N} \mathrm{diag}(Z) + \frac{B_t(B_t - 1)}{N^2} Z.$$

Proof. We have that

$$\mathbb{E}[A_t A_t^{\mathsf{T}}] = \sum_{i=1}^{B_t} \mathbb{E}[A_{i,t} A_{i,t}^{\mathsf{T}}] = \frac{B_t}{N} \mathrm{Id}_{N \times N}\,.$$

Similarly, we find that

$$\begin{aligned} \mathbb{E}[A_t A_t^{\mathsf{T}} Z A_t A_t^{\mathsf{T}}] &= \sum_{i,j=1}^{B_t} \mathbb{E}[A_{i,t} A_{i,t}^{\mathsf{T}} Z A_{j,t} A_{j,t}^{\mathsf{T}}] \\ &= \sum_{i=1}^{B_t} \mathbb{E}[A_{i,t} A_{i,t}^{\mathsf{T}} Z A_{i,t} A_{i,t}^{\mathsf{T}}] + 2 \sum_{1 \leq i < j \leq B_t} \mathbb{E}[A_{i,t} A_{i,t}^{\mathsf{T}} Z A_{j,t} A_{j,t}^{\mathsf{T}}] \\ &= \frac{B_t}{N} \mathrm{diag}(Z) + \frac{B_t(B_t - 1)}{N^2} Z, \end{aligned}$$

which completes the proof. □

Let us first check convergence in mean:

$$\mathbb{E}[W_t] Q_\parallel \to W_{\min}.$$

To see this, note that Lemma F.2 implies

$$\mathbb{E}[Y_t X_t^{\mathsf{T}}] = Y X^{\mathsf{T}} \qquad \mathbb{E}[X_t X_t^{\mathsf{T}}] = X X^{\mathsf{T}},$$

which yields that

$$W_t^* = YX^\mathsf{T}[XX^\mathsf{T}]^+ = W_{\min} \tag{F.3}$$

for all t. We now prove convergence. Since all W_t^* are equal to $W_{\min}$, we find that $\Xi_t^* = 0$. By (B.9) and Lemma F.2 we have

$$\mathbb{E}[W_{t+1}] - W_{\min} = (\mathbb{E}[W_t] - W_{\min})\Big(\operatorname{Id} - \frac{2\eta_t}{N} XX^\mathsf{T}\Big),$$

which implies since $\frac{2\eta_t}{N} < \lambda_{\text{max}}(XX^\mathsf{T})^{-1}$ for large t that for some $C > 0$ we have

$$\begin{aligned}\|\mathbb{E}[W_t]Q_\| - W_{\min}\|_F &\leq \|W_0 Q_\| - W_{\min}\|_F \prod_{s=0}^{t-1} \left\| Q_\| - \frac{2\eta_s}{N} XX^\mathsf{T} \right\|_2 \\ &\leq C\|W_0 Q_\| - W_{\min}\|_F \exp\Big(- \sum_{s=0}^{t-1} \frac{2\eta_s}{N} \lambda_{\text{min},V_\|}(XX^\mathsf{T})\Big). \end{aligned} \tag{F.4}$$

From this we readily conclude using (F.2) the desired convergence in mean $\mathbb{E}[W_t]Q_\| \to W_{\min}$.

Let us now prove that the variance tends to zero. By Proposition B.6, we find that $Z_t = \mathbb{E}[(W_t - \mathbb{E}[W_t])^\mathsf{T}(W_t - \mathbb{E}[W_t])]$ has two-sided decay of type $(\{A_t\}, \{C_t\})$ with

$$A_t = \frac{2\eta_t}{N} X_t X_t^\mathsf{T}, \qquad C_t = \frac{4\eta_t^2}{N^2}\left[\operatorname{Id} \circ \text{Var}((\mathbb{E}[W_t]X_t - Y_t)X_t^\mathsf{T})\right].$$

To understand the resulting rating of convergence, let us first obtain a bound on $\text{tr}(C_t)$. To do this, note that for any matrix A, we have

$$\text{tr}\left(\operatorname{Id} \circ \text{Var}[A]\right) = \text{tr}\Big(\mathbb{E}\left[A^\mathsf{T} A\right] - \mathbb{E}\left[A\right]^\mathsf{T} \mathbb{E}\left[A\right]\Big).$$

Moreover, using the definition (F.1) of the matrix A_t and writing

$$M_t := \mathbb{E}\left[W_t\right] X - Y,$$

we find

$$\left((\mathbb{E}\left[W_t\right] X_t - Y_t)X_t^\mathsf{T}\right)^\mathsf{T} (\mathbb{E}\left[W_t\right] X_t - Y_t)X_t^\mathsf{T} = XA_tA_t^\mathsf{T} M_t^\mathsf{T} M_t A_t A_t^\mathsf{T} X^\mathsf{T}$$

as well as

$$\mathbb{E}\left[\left((\mathbb{E}[W_t]X_t - Y_t)X_t^\mathsf{T}\right)\right]^\mathsf{T} \mathbb{E}\left[(\mathbb{E}[W_t]X_t - Y_t)X_t^\mathsf{T}\right] = X\mathbb{E}\left[A_tA_t^\mathsf{T}\right] M_t^\mathsf{T} M_t \mathbb{E}\left[A_tA_t^\mathsf{T}\right] X^\mathsf{T}.$$

Hence, using the expression from Lemma F.2 for the moments of A_t and recalling the scaling factor $c_t = (N/B_t)^{1/2}$, we find

$$\text{tr}(C_t) = \frac{4\eta_t^2}{B_t} \text{tr}\left(X \left\{ \text{diag}\left(M_t^\mathsf{T} M_t\right) - \frac{1}{N} M_t^\mathsf{T} M_t \right\} X^\mathsf{T} \right).$$

Next, writing

$$\Delta_t := \mathbb{E}[W_t] - W_{\min}$$

and recalling (F.3), we see that

$$M_t = \Delta_t X.$$

Thus, applying the estimates (F.4) about exponential convergence of the mean, we obtain

$$\text{tr}(C_t) \leq \frac{8\eta_t^2}{B_t} \left\|\Delta_t Q_{||}\right\|_2^2 \left\|XX^T\right\|_2^2 \leq C\frac{8\eta_t^2}{B_t} \left\|XX^T\right\|_2^2 \left\|\Delta_0 Q_\|\right\|_F^2 \exp\Big(-\sum_{s=0}^{t-1} \frac{4\eta_s}{N} \lambda_{\text{min},V_\|}(XX^\mathsf{T})\Big). \tag{F.5}$$

Notice now that $Y_r = Q_\| - A_r$ satisfies the conditions of Theorem A.6 with $a_r = 1 - 2\eta_r \frac{1}{N} \lambda_{\text{min},V_\|}(XX^\mathsf{T})$ and $b_r^2 = \frac{4\eta_r^2}{B_r a_r^2 N} \text{tr}\Big(X \,\text{diag}(X^\mathsf{T} X)X - \frac{1}{N} XX^\mathsf{T} XX^\mathsf{T}\Big)$. By Theorem A.6 we then obtain for any $t > s > 0$ that

$$\mathbb{E}\left[\left\| \prod_{r=s+1}^{t} (Q_\| - A_r) \right\|_2^2\right] \leq e^{\sum_{r=s+1}^t b_r^2} \prod_{r=s+1}^{t} \Big(1 - 2\eta_r \frac{1}{N} \lambda_{\text{min},V_\|}(XX^\mathsf{T})\Big)^2. \tag{F.6}$$

By two-sided decay of Z_t, we find by (F.5), (F.6), and (A.9) that

$$\begin{aligned} &\mathbb{E}[\|W_t Q_\| - \mathbb{E}[W_t] Q_\|\|_F^2] = \operatorname{tr}(Q_\| Z_t Q_\|) \\ &\leq e^{-\frac{4}{N}\lambda_{\min, V_\|}(XX^\intercal)\sum_{s=0}^{t-1}\eta_s} \frac{\|XX^\intercal\|_2^2}{N^2}\|\Delta_0 Q_\|\|_F^2 C \sum_{s=0}^{t-1} \frac{8\eta_s^2}{B_s/N} e^{\frac{4\eta_s}{N}\lambda_{\min, V_\|}(XX^\intercal) + \sum_{r=s+1}^{t} b_r^2}. \end{aligned} \tag{F.7}$$

Since $\eta_s \to 0$, we find that $\eta_s \frac{N}{B_s} e^{\frac{4\eta_s}{N}\lambda_{\min, V_\|}(XX^\intercal)}$ is uniformly bounded and that $b_r^2 \leq \frac{4}{N}\lambda_{\min, V_\|}(XX^\intercal)\eta_r$ for sufficiently large r. We therefore find that for some $C' > 0$,

$$\mathbb{E}[\|W_t Q_\| - \mathbb{E}[W_t] Q_\|\|_F^2] \leq C' \sum_{s=0}^{t-1} \eta_s e^{-\frac{4}{N}\lambda_{\min, V_\|}(XX^\intercal)\sum_{r=0}^{s}\eta_r},$$

hence $\lim_{t\to\infty} \mathbb{E}[\|W_t Q_\| - \mathbb{E}[W_t] Q_\|\|_F^2] = 0$ by Lemma A.8. Combined with the fact that $\mathbb{E}[W_t] Q_\| \to W_{\min}$, this implies that $W_t Q_\| \xrightarrow{p} W_{\min}$.

To obtain a rate of convergence, observe that by (F.4) and the fact that $\eta_t = \Theta(t^{-x})$, for some $C_1, C_2 > 0$ we have

$$\|\mathbb{E}[W_t] Q_\| - W_{\min}\|_F \leq C_1 \exp\Big(-C_2 t^{1-x}\Big). \tag{F.8}$$

Similarly, by (F.7) and the fact that $\frac{\eta_s}{B_s/N} < \infty$ uniformly, for some $C_3, C_4, C_5 > 0$ we have

$$\mathbb{E}[\|W_t Q_\| - \mathbb{E}[W_t] Q_\|\|_F^2] \leq C_3 \exp\Big(-C_4 t^{1-x}\Big) t^{1-x}$$

We conclude by Chebyshev's inequality that for any $a > 0$ we have

$$\mathbb{P}\Big(\|W_t Q_\| - W_{\min}\|_F \geq C_1 \exp\Big(-C_2 t^{1-x}\Big) + a \cdot \sqrt{C_3} t^{\frac{1}{2}-\frac{x}{2}} e^{-C_4 t^{1-x}/2}\Big) \leq a^{-2}.$$

Taking $a = t$, we conclude as desired that for some $C > 0$, we have

$$e^{Ct^{1-x}} \|W_t Q_\| - W_{\min}\|_F \xrightarrow{p} 0.$$

This completes the proof of Theorem F.1. □