Proceedings of Machine Learning Research vol 313:1–62, 2026 37th International Conference on Algorithmic Learning Theory

Optimal L2 Regularization in High-dimensional Continual Linear Regression

Gilad Karpel KARPELGILAD@GMAIL.COM
Department of Data and Decisions Sciences, Technion

Edward Moroshko EDWARD.MOROSHKO@GMAIL.COM
School of Engineering, University of Edinburgh

Ran Levinstein RANLEVINSTEIN@GMAIL.COM
Department of Computer Science, Technion

Ron Meir RMEIR@EE.TECHNION.AC.IL
Daniel Soudry DANIEL.SOUDRY@GMAIL.COM
Department of Electrical and Computing Engineering, Technion

Itay Evron ITAY@EVRON.ME
Meta

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

We study generalization in an overparameterized continual linear regression setting, where a model is trained with L2 (isotropic) regularization across a sequence of tasks. We derive a closed-form expression for the expected generalization loss in the high-dimensional regime that holds for arbitrary linear teachers. We demonstrate that isotropic regularization mitigates label noise under both single-teacher and multiple i.i.d. teacher settings, whereas prior work accommodating multiple teachers either did not employ regularization or used memory-demanding methods. Furthermore, we prove that the optimal fixed regularization strength scales nearly linearly with the number of tasks T, specifically as $T/\ln T$. To our knowledge, this is the first such result in theoretical continual learning. Finally, we validate our theoretical findings through experiments on linear regression and neural networks, illustrating how this scaling law affects generalization and offering a practical recipe for the design of continual learning systems.

Keywords: Continual learning, Lifelong learning, Regularization methods, Regularization strength

1. Introduction

In the era of large foundation models (e.g., LLMs), the ability to learn tasks *sequentially* without catastrophic forgetting (McCloskey and Cohen, 1989) or retraining from scratch has become increasingly critical (see surveys in Wang et al., 2024; van de Ven et al., 2024). Continual learning not only saves substantial computational resources but is also essential in settings where previous data cannot be retained, e.g., under privacy or storage constraints.

We focus on continual linear regression, a fundamental analytical setting that has proven useful for illuminating diverse aspects of continual learning systems, such as the effects of task similarity (Lin et al., 2023; Hiratani, 2024; Tsipory et al., 2025), task recurrence (Evron et al., 2022; Kong et al., 2023), overparameterization (Goldfarb et al., 2024), algorithms (Doan et al., 2021; Peng et al., 2023), regularization (Li et al., 2023; Levinstein et al., 2025), and trade-offs such as memory versus generalization (Li et al., 2025). Similar linear settings are also used to study continual *classification* (e.g., Evron et al., 2023; Jung et al., 2025).

© 2026 G. Karpel, E. Moroshko, R. Levinstein, R. Meir, D. Soudry & I. Evron.

We analyze a popular regularization scheme that penalizes changes in the parameter space to prevent forgetting previous expertise. Specifically, we focus on L2 regularization, which penalizes changes isotropically, i.e., $\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w}} \left\{ \mathcal{L}_t(\mathbf{w}) + \lambda d \|\mathbf{w} - \mathbf{w}_{t-1}\|^2 \right\}$. Despite not maintaining an additional regularization weight matrix, L2 regularization has been found to work well in practice (Hsu et al., 2018; Lubana et al., 2021; Smith et al., 2023).

The regularization strength λ plays a central role in methods like ours, governing the balance between stability and plasticity (Mermillod et al., 2013; Kirkpatrick et al., 2017). Theoretically guided scaling laws for the regularization strength can greatly reduce computational cost, alleviating the need for extensive hyperparameter tuning. Only a few theoretical works have focused on the regularization strength. For example, Levinstein et al. (2025) shows that near-optimal convergence rates can be achieved using a fixed strength. However, their analysis pertains only to the training error in the worst case, leaving open the question of whether the chosen strength is truly optimal.

While some prior theoretical work analyzes the *training* loss and forgetting (e.g., Evron et al., 2022, 2026; Kong et al., 2023; Goldfarb et al., 2024; Levinstein et al., 2025), we adopt a complementary and widely studied *statistical* viewpoint, focusing on the *generalization* loss of our continual scheme. Following prior work in a similar spirit (e.g., Lin et al., 2023; Li et al., 2023; Zhao et al., 2024), we study a random design in which labels are generated by an underlying (noisy) linear 'teacher' (i.e., task ground truth). Importantly, our analysis accommodates *multiple teachers*, in contrast to several prior studies (e.g., Ding et al., 2024; Zhao et al., 2024; Levinstein et al., 2025), which assume a single global teacher for all tasks. Allowing multiple teachers provides a more realistic model for practical scenarios where the underlying ground truth (e.g., user preferences or fraud behaviors) naturally evolve over time. Under an i.i.d. feature assumption common in statistical studies (e.g., Lin et al., 2023; Goldfarb and Hand, 2023; Zhao et al., 2025), a single teacher yields generalization affected solely by label noise, while multiple teachers introduce task-level variation, again, making the setting more realistic and challenging.

A closely related paper by Lin et al. (2023), which also considered multiple teachers, studied an *unregularized* scheme where past tasks influence optimization only *implicitly*, through the bias of gradient algorithms (see also Evron et al. 2022). While their analysis enables strong theoretical results, our findings highlight that incorporating a regularizer is crucial, particularly for reducing the noise floor induced by label noise.

Our results provide general insights for *any* setup with multiple teachers, whereas our more concrete findings pertain to teachers drawn i.i.d. from some distribution—allowing simple yet illustrative analysis that captures effects of teacher variance. We place particular emphasis on the regularization strength, proving that its optimal value scales nearly linearly with the number of tasks. This practical rule of thumb alleviates the need for costly hyperparameter tuning with many tasks, as we can now tune the regularization strength using only a few tasks, and then scale it accordingly to longer task horizons. We empirically verify these findings on an `MNIST`-based dataset using both a linear regression model and a ReLU-activated neural network.

Summary of our contributions.

1. We analyze generalization in continual linear regression in a high-dimensional regime, accommodating multiple teachers, label noise, and an arbitrary number of tasks—a setup broader than those in prior work (see Table 1).

2. For both a single teacher and i.i.d. teachers, we identify the optimal regularization strength $\lambda^\star$.

3. We demonstrate that L2 regularization, which does not require maintaining an additional weighting matrix, can close a gap in prior theoretical literature by staying informative in noisy regimes.
4. We validate our theoretical results in a standard regularized continual linear model, using both synthetic random regression data and an `MNIST`-based classification problem.
5. Finally, we demonstrate similar findings with simple neural networks, validating our theory-guided rule of thumb for choosing the regularization strength.

2. Setting: Regularized Continual Linear Regression

As explained in the introduction, we study the illustrative continual linear regression setting using an L2 regularization scheme, described formally below.

Learning tasks in a sequence. In our setting, the learner is exposed to a sequence of T tasks $((\mathbf{X}_i, \mathbf{y}_i))_{i=1}^T$, where $\mathbf{X}_i \in \mathbb{R}^{n\times d}$, $\mathbf{y}_i \in \mathbb{R}^n$.[1] At each iteration t, the learner updates the predictor $\mathbf{w}_t$ by fitting only the current task, *without access to samples from previous tasks*.

To be able to learn continually and mitigate *catastrophic forgetting*, it is common to add a regularization term towards the previous iterate $\mathbf{w}_{t-1}$. While the regularizer is often based on the Fisher information of previous tasks (Kirkpatrick et al., 2017; Benzing, 2022), recent works have observed empirically that L2 (isotropic) regularization achieves similar performance (Hsu et al., 2018; Lubana et al., 2021; Smith et al., 2023). Thus, we follow recent theoretical work (e.g., Li et al., 2023; Levinstein et al., 2025) and focus on the following L2 regularization scheme.

Scheme 1 Continual Linear Regression with L2 Regularization

Input: $\{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^T$, initial iterate $\mathbf{w}_0$, regularization strength λ
For each iteration $t = 1, \dots, T$:
$$\mathbf{w}_t \leftarrow \operatorname{argmin}_{\mathbf{w}} \left\{ \|\mathbf{X}_t\mathbf{w} - \mathbf{y}_t\|^2 + \lambda d \|\mathbf{w} - \mathbf{w}_{t-1}\|^2 \right\}$$
Output: $\mathbf{w}_T$

Note that for $\lambda \to 0$, the algorithm reduces to optimizing the current task alone (as analyzed in Evron et al. (2022)), whereas for large values of λ, the current task has negligible influence. We normalize the regularization strength by the data dimensionality, following a convention commonly used in Bayesian settings, where the prior is often scaled in this way (Lee et al., 2018).

For ease of readability, we keep implicit any dependence of $\mathbf{w}_t$ on other quantities (e.g., λ, d). Furthermore, we collect all notation used throughout the paper in Table 2 (App. A).

2.1. Statistical Setup: High-Dimensional Random Design with Linear Teachers

To study generalization, we adopt a statistical setup with random data, multiple (linear) teachers, and noisy labels. Variants of this setup are commonly used in the theoretical literature on continual learning (e.g., Goldfarb and Hand, 2023; Lin et al., 2023; Zhao et al., 2024; Friedman and Meir, 2025) and in broader machine learning research (e.g., Hsu et al., 2012; Belkin et al., 2020; Bartlett et al., 2021). As shown in Table 1, our setup is comparatively permissive, allowing both label noise and multiple teachers (e.g., i.i.d. teachers).

1. To simplify exposition, we assume all tasks have n samples, though our analysis can extend to varying sample sizes.

Data model. We consider a random design, with independent random feature matrices $\mathbf{X}_1, \ldots, \mathbf{X}_T \in \mathbb{R}^{n \times d}$, where the entries of each matrix $\mathbf{X}_t$ are i.i.d. random variables with mean 0, variance $v_x > 0$, and a finite $(4+\epsilon)^{\text{th}}$ moment for some $\epsilon > 0$.[2] Moreover, we focus on the high-dimensional regime, where

$$n, d \to \infty, \quad \frac{n}{d} \to \alpha \leq 1 \,.$$

For each task $t \in [T] \triangleq \{1, \ldots, T\}$, labels are generated according to a corresponding ground-truth 'teacher' vector $\mathbf{w}_t^\star \in \mathbb{R}^d$ via a noisy linear model. That is,

$$\mathbf{y}_t = \mathbf{X}_t \mathbf{w}_t^\star + \mathbf{z}_t \,,$$

where $\mathbf{z}_t$ is a label noise vector, satisfying $\mathbb{E}[\mathbf{z}_t] = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{z}_t) = v_z \mathbf{I}$ for some noise variance $v_z \geq 0$, sampled independently of other tasks, i.e., of $\mathbf{X}_{t'}, \mathbf{z}_{t'}, \forall t' \neq t$.

Finally, for technical reasons, we assume finite asymptotic teacher correlations. That is,

$$(\mathbf{w}_i^\star - \mathbf{w}_j^\star)^\top (\mathbf{w}_k^\star - \mathbf{w}_l^\star) \xrightarrow[d \to \infty]{} r_{i,j,k,l} \,, \qquad \forall i, j, k, l \in \{0, 1, \ldots, T\} \,.$$

Implications of the data model. High dimensionality allows us to derive explicit expressions using the Marchenko–Pastur theorem (Marchenko and Pastur, 1967). Importantly, accommodating multiple teachers enables us to analyze i.i.d. teachers, thereby characterizing continual scenarios that are more practical than those considered in most prior work (see Section 3.1).

Metrics of interest. Some prior theoretical work in continual learning (e.g., Goldfarb et al., 2024; Evron et al., 2022, 2026) has analyzed the (training) squared error $\|\mathbf{X}_i \mathbf{w} - \mathbf{y}_i\|^2$. However, under our statistical model, it is readily seen that averaging over the random test data and label noise yields

$$\mathbb{E}\left\|\mathbf{X}_i \mathbf{w} - \mathbf{y}_i\right\|^2 = \mathbb{E}\left\|\mathbf{X}_i(\mathbf{w} - \mathbf{w}_i^\star) - \mathbf{z}_i\right\|^2 = v_x n \, \|\mathbf{w} - \mathbf{w}_i^\star\|^2 + v_z d \,, \tag{1}$$

making $\|\mathbf{w} - \mathbf{w}_i^\star\|$ the only quantity influenced by the actual solution $\mathbf{w}$. Thus, we follow other prior work (Lin et al., 2023) and measure *generalization* by the distance to the teacher(s) generating the labels. Our formal definitions are provided below.

Definition 1 (Individual Task Loss) *For any $\mathbf{w} \in \mathbb{R}^d$, the loss of task $i \in [T]$ is defined as the distance of $\mathbf{w}$ from the task's associated teacher. That is,*

$$\mathcal{L}_i(\mathbf{w}) = \|\mathbf{w} - \mathbf{w}_i^\star\|^2 \,.$$

Definition 2 (Average Generalization Loss) *The expected generalization loss at iteration t is defined as the average of the individual losses of* all *tasks encountered so far, evaluated at the iterate $\mathbf{w}_t$. That is,*

$$\mathbb{E}\left[G_t\right] \triangleq \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left[\mathcal{L}_i(\mathbf{w}_t)\right] = \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\,\|\mathbf{w}_t - \mathbf{w}_i^\star\|^2 \,.$$

Note that $\mathbf{w}_t$ is a random variable, as it depends on $\mathbf{X}_i, \mathbf{z}_i$ and the possibly random $\mathbf{w}_i^\star, \forall i \in [t]$.

2. For example, the entries of $\mathbf{X}_1$ can be $\mathcal{N}(0,1)$ and the entries of $\mathbf{X}_2$ can be $\mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$.

3. Generalization Analysis and Optimal Regularization Scaling Laws

In this section, we derive a closed-form expression for the expected generalization loss (of Definition 2). This general expression accommodates diverse teacher setups and serves as the foundation for investigating an important special case in later sections.

Theorem 3 (Generalization under General Teachers) *Recall the high-dimensional statistical setting of Section 2.1, where the number of samples and the dimension $n, d \to \infty$ and $\frac{n}{d} \to \alpha \leq 1$. Moreover, recall that v_x, v_z are the variances of features and labels, respectively. Then, the expected generalization loss (Definition 2) after learning T tasks with Scheme 1 is,*

$$\mathbb{E}[G_T] = \underbrace{v_z c \frac{1-a^T}{1-a}}_{\text{Label noise term}} + \underbrace{\frac{1}{T}\sum_{i=1}^{T} \|\mathbf{w}_T^\star - \mathbf{w}_i^\star\|^2}_{\text{Teacher variability term}} + \underbrace{2\left(\mathbf{w}_T^\star - \bar{\mathbf{w}}^\star\right)^\top \sum_{i=1}^{T} b^{T-i+1}\left(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\right)}_{\text{Interaction term}}$$

$$+ \underbrace{\sum_{i=1}^{T}\sum_{i'=1}^{T} a^{T-\max(i,i')+1} b^{|i-i'|}\left(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\right)^\top \left(\mathbf{w}_{i'-1}^\star - \mathbf{w}_{i'}^\star\right)}_{\text{Accumulated temporal (teacher) correlations}},$$

where we define $\mathbf{w}_0^\star \triangleq \mathbf{w}_0$ (by abuse of notation), the average teacher $\bar{\mathbf{w}}^\star \triangleq \frac{1}{T}\sum_{i=1}^{T} \mathbf{w}_i^\star$, and

$$a \triangleq \frac{1}{2}\left(1 - \alpha + \frac{\lambda(1+\alpha) + v_x(1-\alpha)^2}{\sqrt{\lambda^2 + 2\lambda v_x\,(1+\alpha) + v_x^2\,(1-\alpha)^2}}\right),$$

$$b \triangleq \frac{1}{2}\left(1 - \alpha - \frac{\lambda}{v_x} + \frac{1}{v_x}\sqrt{\lambda^2 + 2\lambda v_x(1+\alpha) + v_x^2(1-\alpha)^2}\right),$$

$$c \triangleq \frac{1}{2v_x}\left(\frac{v_x(1+\alpha) + \lambda}{\sqrt{\lambda^2 + 2\lambda v_x(1+\alpha) + v_x^2(1-\alpha)^2}} - 1\right).$$

To aid interpretation, we first decompose the generalization loss into its constituent terms. We then provide a proof sketch outlining the key steps of the full proof presented in App. C.

Decomposing the generalization loss. The expected generalization in Theorem 3 can be decomposed into four terms, interpreted as follows:

- **Label noise term:** arises from label noise of the underlying model, scaling with label variance.
- **Teacher variability:** captures the intrinsic mismatch between the current teacher and past teachers due to task heterogeneity.
- **Interaction term:** reflects how the deviation of the final teacher from the average teacher interacts with the sequence of teacher updates, discounting older changes.
- **Temporal correlations:** accounts for correlations between successive teacher updates, describing how past changes compound and align (or misalign) over time.

Proof sketch. The iterates of Scheme 1 follow the recursive form:

$$\mathbf{w}_t = \left(\mathbf{X}_t^\top \mathbf{X}_t + \lambda d \mathbf{I}\right)^{-1} \left(\mathbf{X}_t^\top \mathbf{y}_t + \lambda d \mathbf{w}_{t-1}\right) .$$

Denote $\mathbf{P}_t = \lambda d \left(\mathbf{X}_t^\top \mathbf{X}_t + \lambda d \mathbf{I}\right)^{-1}$, $\mathbf{S}_{t:k} \triangleq \mathbf{P}_t \mathbf{P}_{t-1} \cdots \mathbf{P}_k$ and unroll the recursive form, to obtain

$$\mathbf{w}_t - \mathbf{w}_i^\star = \mathbf{S}_{t:1}(\mathbf{w}_0 - \mathbf{w}_1^\star) + \frac{1}{\lambda d} \sum_{k=1}^{t} \mathbf{S}_{t:k} \mathbf{X}_k^\top \mathbf{z}_k + \sum_{k=2}^{t} \mathbf{S}_{t:k} \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right) .$$

To analyze $\mathbb{E}[G_T] = \frac{1}{T} \sum_i \mathbb{E}\|\mathbf{w}_T - \mathbf{w}_i^\star\|^2$, we invoke a Marchenko–Pastur result to approximate the resolvent $\left(\frac{1}{v_x n} \mathbf{X}^\top \mathbf{X} - z \mathbf{I}\right)^{-1}$ with its deterministic equivalent (Couillet and Liao, 2022):

$$\left\| \mathbb{E}\left[\left(\tfrac{1}{v_x n} \mathbf{X}^\top \mathbf{X} - z \mathbf{I}\right)^{-1}\right] - \frac{-\left(1 - \frac{1}{\alpha} - z\right) + \sqrt{\left(1 - \frac{1}{\alpha} - z\right)^2 - \frac{4}{\alpha} z}}{-2\alpha / z} \mathbf{I} \right\| \xrightarrow[n,d \to \infty,\ d/n \to 1/\alpha]{\text{a.s.}} 0 .$$

We use the resolvent and standard identities to obtain closed-form expressions for $\mathbb{E}[\mathbf{P}_t]$, $\mathbb{E}[\mathbf{P}_t^2]$, and $\mathbb{E}[\left(\frac{1}{\lambda d}\right)^2 \mathbf{P}_t \mathbf{X}_t^\top \mathbf{X}_t \mathbf{P}_t]$ as scalar multiples of $\mathbf{I}$, depending only on (λ, v_x, α). These are then substituted into the unrolled form to yield the explicit expression. ■

Sanity check. Lin et al. (2023) studied our Scheme 1 in the limit $\lambda \to 0$, i.e., an equivalent unregularized scheme where each task is solved to convergence via SGD initialized at the previous iterate. Below, we verify that Theorem 3 recovers their Theorem 4.1 as a special case.

Example 1 (No Regularization) *Indeed, in the high-dimensional regime, Theorem 3 generalizes Theorem 4.1 in Lin et al. (2023). Specifically, after adjusting notation, their result becomes,*

$$\mathbb{E}[G_T] = \frac{(1-\alpha)^T}{T} \sum_{i=1}^{T} \|\mathbf{w}_i^\star\|^2 + \frac{1}{T} \sum_{i=1}^{T} (1-\alpha)^{T-i} \alpha \sum_{k=1}^{T} \|\mathbf{w}_k^\star - \mathbf{w}_i^\star\|^2 + \frac{d v_z}{d-n-1} \left(1 - (1-\alpha)^T\right) ,$$

aligning with Theorem 3 in the high-dimensional, unregularized case ($d, n \to \infty$, $\lambda \to 0$, $v_x = 1$). Proof is given in App. C.1.

3.1. Special Case: Multiple i.i.d. Teachers

Thus far, Theorem 3 has accommodated multiple "label-generating" teachers. To obtain more meaningful results, we now assume that teachers are sampled i.i.d.[3] from a common distribution.

Assumption 1 (i.i.d. Teachers) *Considering the statistical setting of Section 2.1, the teachers $\mathbf{w}_1^\star, \ldots, \mathbf{w}_T^\star$ are drawn i.i.d. from a distribution with mean $\mathbf{w}^\star$ and covariance $\mathbf{\Sigma}$ where $\|\mathbf{\Sigma}\|_* \leq C < \infty$, independently of all other random variables (features and labels).*

This assumption is more permissive than most theoretical work in continual learning (Table 1), which typically assumes (i) a single teacher with label noise or (ii) joint realizability across *training* sets (as in highly overparameterized regimes). By contrast, in applications like autonomous systems and e-commerce, underlying dynamics often shift. For instance, varying agents and simulators inevitably introduce teacher variance due to environmental gaps (Zhao et al., 2020). Our setup more accurately captures these practical challenges by explicitly allowing for multiple teachers.

3. More generally, our analysis accommodates teachers drawn independently from several distributions with a common mean and covariance. For ease of exposition, however, we present results only for the i.i.d. setting.

We begin with Corollary 4, which specializes Theorem 3 to the i.i.d. teacher setting. We then use it to study how the regularization strength λ impacts generalization. In this section, all expectations are taken over features, label noise, *and teachers*. For ease of presentation, we assume a zero initialization $\mathbf{w}_0 = 0$. Full proofs are given in App. E.

Corollary 4 (Generalization under i.i.d. Teachers) *Under i.i.d. teachers (Assumption 1), and for all $T > 1$, the expected generalization loss of Scheme 1 becomes,*

$$\mathbb{E}[G_T] = \underbrace{v_z c \frac{1-a^T}{1-a}}_{\textit{Label noise}} + \underbrace{a^T \|\mathbf{w}^\star\|^2}_{\textit{Teacher scale}} + \underbrace{2\mathrm{Tr}(\mathbf{\Sigma})}_{\substack{\textit{Teacher}\\ \textit{variance}}} \left(\frac{1-b}{1-a} + \frac{2b^T - 1}{T} + \frac{a^T(-a+2b-1)}{2(1-a)} \right).$$

The following lemma contrasts regularization calibrated to the task horizon T against any *fixed* strength. Specifically, the generalization loss with optimal $\lambda^\star(T)$ is monotonically decreasing in T (under a single teacher), but non-monotonic for any fixed strength (under i.i.d. teachers).

Lemma 5 (More Tasks Can Be Useful or Harmful—Depending on Teacher Setup) *Consider the i.i.d. teacher setting (Assumption 1) with a label noise $v_z > 0$.*

- ***Single teacher:*** *When $\mathbf{w}_1^\star = \cdots = \mathbf{w}_T^\star$ (i.e., $\mathrm{Tr}(\mathbf{\Sigma}) = 0$), the expected generalization loss with an* optimal *horizon-dependent strength $\lambda^\star(T)$ is monotonically* decreasing *in the task horizon T.*
- ***Multiple i.i.d. teachers:*** *When teachers vary across tasks (i.e., $\mathrm{Tr}(\mathbf{\Sigma}) > 0$), then under any* fixed *strength $\lambda > 0$, the expected generalization loss is monotonically* increasing *in the task horizon T for all $T \geq T'$ for a sufficiently large T'.*

3.1.1. OPTIMAL SCALING OF THE REGULARIZATION STRENGTH: $\lambda^\star \asymp \dfrac{v_x \alpha T}{\log(\text{SIGNAL}/\text{NOISE})}$

Now, we characterize the optimal regularization strength $\lambda^\star$ that minimizes the generalization loss. A theory-guided rule for λ is essential when hyperparameter tuning is prohibitive, as in continual learning. In our setting, we prove a near linear optimal scaling with α, v_x, T leading to mean-square convergence to the "global" teacher, i.e., $\mathbf{w}_T \xrightarrow{T\to\infty} \mathbf{w}^\star$.

Theorem 6 (Optimal Regularization Scales Nearly Linearly with T) *Under i.i.d. teachers (Assumption 1) with non-zero[4] mean teacher $\mathbf{w}^\star$, the optimal fixed regularization strength that minimizes the expected generalization loss after T iterations satisfies $\lambda^\star = \lambda^\star(T) \asymp \frac{T}{\ln T}$. More precisely, for any $\epsilon > 0$, there exists $T_0 = O(1/\epsilon^2)$ such that for all $T \geq T_0$,*

$$(1-\epsilon)\frac{2v_x\alpha T}{\ln\left(\frac{4\alpha T\|\mathbf{w}^\star\|^2 v_x}{v_z + v_x \mathrm{Tr}(\mathbf{\Sigma})(1+\alpha)}\right)} < \lambda^\star < (1+\epsilon)\frac{2v_x\alpha T}{\ln\left(\frac{4\alpha T\|\mathbf{w}^\star\|^2 v_x}{v_z + v_x \mathrm{Tr}(\mathbf{\Sigma})(1+\alpha)}\right)}.$$

Furthermore, in the degenerate noiseless case of $\mathrm{Tr}(\mathbf{\Sigma}) = v_z = 0$, it holds that $\lambda^\star \to 0$.

Interpretation. The denominator can be interpreted as a $\log(\text{signal}/\text{noise})$ factor. The *signal* consists of $\|\mathbf{w}^\star\|^2 v_x$, capturing the strength of shared structure across tasks, which is reinforced over longer task horizons T, while $\frac{n}{d} \to \alpha$ measures how informative each task is. The internal denominator scales with *noise* sources: label noise v_z, feature noise v_x, and teacher variance $\mathrm{Tr}(\mathbf{\Sigma})$.

In Section 6, we contrast our optimal strength with results from prior work.

4. Since $\mathbf{w}_0 = \mathbf{0}$, additionally having $\mathbb{E}[\mathbf{w}_t^\star] = \mathbf{0}$ would directly imply an optimal regularization of $\lambda^\star \to \infty$.

Lemma 5 indicates that for fixed λ, performance deteriorates as T exceeds a certain threshold (resembling overfitting). Consequently, given an unlimited "budget" of tasks, the optimal strategy is to increase regularization as much as possible, as formalized below.

Theorem 7 (Asymptotic Generalization under Strong Regularization) *Under i.i.d. teachers (Assumption 1), as the horizon $T \to \infty$, expected generalization loss decreases monotonically with strength λ. Furthermore, the iterates converge in mean square to the "global" teacher $\mathbf{w}^\star$:*[5]

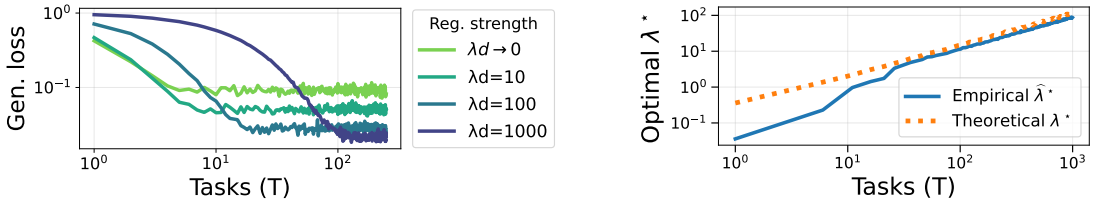
$$\lim_{\lambda\to\infty}\left(\lim_{T\to\infty}\mathbb{E}\left\|\mathbf{w}_T-\mathbf{w}^\star\right\|^2\right)=0\,.$$

4. Experiments on Synthetic Regression Data

Next, we empirically demonstrate how regularization shapes generalization and contrast it with the unregularized case, validating our theoretical analysis.

Noisy teacher setup. For each task $i \in [T]$, we sample a teacher $\mathbf{w}_i^\star \sim \mathcal{N}(\mathbf{w}^\star, \frac{1}{d}\mathbf{I})$, matching Assumption 1. We use $d = 50$ dimensions, $n = 30$ examples per task, and initialization $\mathbf{w}_0 = \mathbf{0}$.

Longer task sequences with stronger regularization overcome label noise. Figure 1(*a*) shows that regularization reduces noise effects, but only when calibrated to the task horizon. We observe that models with weaker regularization benefit from early tasks but soon plateau[6] at a noise floor, reflecting high plasticity but low stability. Conversely, excessive regularization impairs plasticity, effectively preventing learning. Figure 1(*b*) confirms that the optimal regularization strength scales nearly linearly with the horizon—balancing the stability–plasticity tradeoff to lower the noise floor.



(*a*) Regularization can decrease the label noise floor.

(*b*) Optimal strength grows nearly linearly with the number of tasks, as predicted by Theorem 6. We compute $\lambda^\star$ with an exhaustive search.

Figure 1: **Regularization effects on synthetic data.** Interactions between number of tasks (T), regularization strength (λ), and resulting generalization loss (*normalized* by a trivial regressor $\mathbf{w} = \mathbf{0}$). Here, the label noise and feature variances are $v_z = v_x = 1$; Each curve averages over 5 runs.

While the unregularized scheme collapses under label noise, optimal regularization endures. The unregularized scheme solved via gradient descent to convergence corresponds to our Scheme 1 as $\lambda \to 0$. While Evron et al. (2022, 2026) analyze this setup's training loss in jointly realizable—or noiseless—settings, Lin et al. (2023) accommodates label noise and studies the generalization loss. However, as our Example 1 shows, that scheme fails under high label noise due to a noise floor

5. Limits are taken sequentially (first $T \to \infty$, then $\lambda \to \infty$) rather than simultaneously.
6. Lemma 5 predicts that the expected generalization monotonically increases beyond a certain task. While this trend is obscured here by noise under the chosen hyperparameters, it becomes clearer in the figures of Section 5.

$\frac{dv_z}{d-n-1}$ that *persists* as $T \to \infty$ for any number of samples n and dimension d (where $\alpha < 1$). In contrast, we demonstrate next that a properly calibrated regularization in Scheme 1 overcomes even immense label noise v_z, provided a sufficient task horizon T.

Figure 2 compares the generalization loss of the aforementioned unregularized scheme ($\lambda \to 0$) with that of the optimal $\lambda^\star$. Indeed, in the presence of high label noise: (i) the unregularized scheme fails completely due to the high floor noise, *regardless* of the task horizon; and (ii) a properly calibrated $\lambda^\star$ can substantially mitigate label noise after seeing enough tasks.

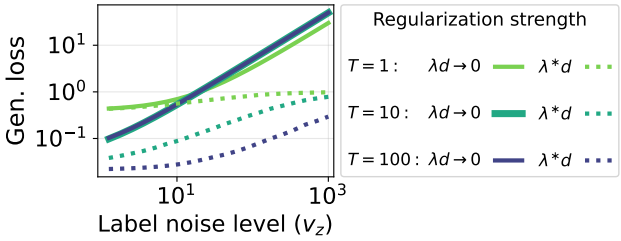


Figure 2: Optimal regularization substantially improves generalization. Here, feature variance is $v_x = 1$; each curve averages over 20 runs.

5. Experiments on MNIST-Based Dataset

While our analysis assumes i.i.d. features and a linear model, we now validate our results in more challenging scenarios where the assumptions of Section 2 do not hold. We use finite-dimensional MNIST data rather than high-dimensional i.i.d. features, solve classification instead of regression, inject bit-flip rather than continuous label noise, and evaluate both linear models and ReLU networks. In all cases, the optimal regularization strength scales as $T/\ln T$, consistent with Theorem 6.

5.1. Constructing a Continual Binary Classification Dataset with Noisy (Flipped) Labels

Samples. We convert the MNIST (multiclass) dataset into a binary classification problem by grouping digits 0–4 into class 0 and digits 5–9 into class 1. To facilitate our teacher models (below), we narrow the data to a random subset of 500 training images and 200 test images.

Teacher models. We consider two types of teachers:

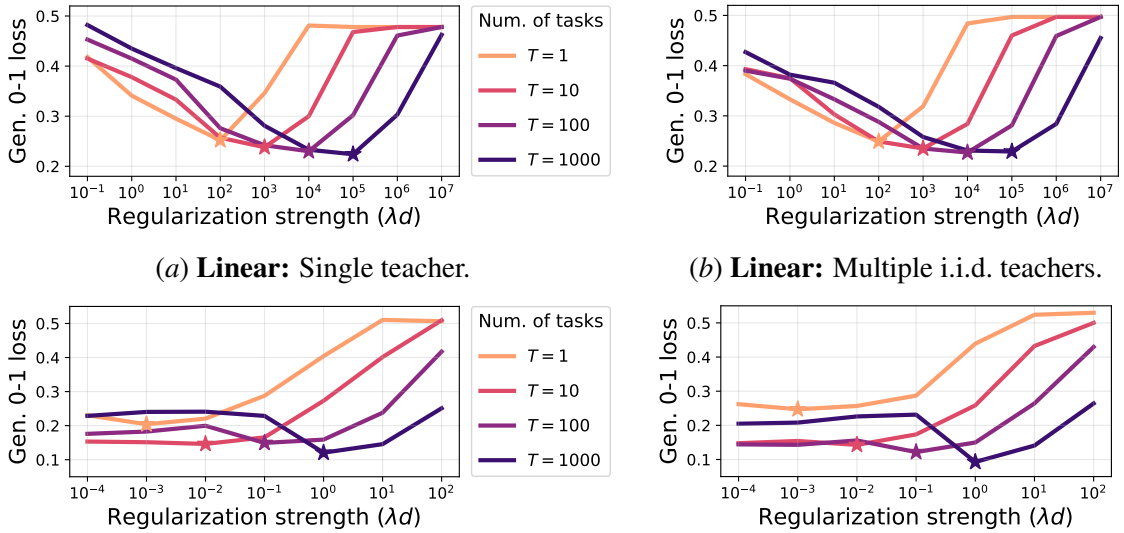
(a) **A single noiseless teacher** that outputs the true labels (tasks still suffer from *label* noise).

(b) **Multiple i.i.d. teachers drawn from a global teacher.** We first train a global teacher with the same architecture as the evaluated model—linear in Section 5.2 and neural in Section 5.3—on the *entire* subsample (train and test). Then, each task draws a teacher by adding Gaussian noise to the global teacher's *weights*. Teachers output binary labels (thresholded at 0.5).

Noisy labels. Each task consists of the *same* sampled images, but differs in labels due to two noise sources. Specifically, each training label is first set by the evaluated teacher (noiseless or noisy), and then corrupted by bit-flip *label* noise—i.e., each label is flipped (i.i.d.) with probability $p = 0.2$.

5.2. Linear Regression Experiments: Validation Beyond i.i.d. Features

We start from a regularized continual linear regression model (Scheme 1), trained with MSE to predict (noisy) *binary* labels generated by either a single noiseless teacher (Figure 3(*a*)) or multiple teachers drawn from a global teacher (Figure 3(*b*)). Predictions are thresholded at 0.5 to yield binary outputs. Empirically, the optimal regularization scales roughly as $T/\ln T$, consistent with Theorem 6 even though the i.i.d. feature assumption does not hold.



(*a*) **Linear:** Single teacher.

(*b*) **Linear:** Multiple i.i.d. teachers.

(*c*) **NN:** Single teacher.

(*d*) **NN:** Multiple i.i.d. teachers.

Figure 3: **Regularization in linear models and neural networks (NN).** We plot the held-out classification error after learning T noisy tasks. Curves are averaged over 5 seeds; stars mark minima.

5.3. Neural Network Experiments: Validation Beyond Linear Models

We also consider a two-layer ReLU neural network (784→500→1) with a sigmoid output, trained continually with L2 regularization. Again, noisy binary labels are generated by either a single noiseless teacher or multiple teachers sampled from a global (linear) teacher (Figures 3(*c*) and 3(*d*)). We optimize MSE for consistency with Scheme 1, thresholding predictions at 0.5. Even in this neural setting, optimal regularization follows the $T/\ln T$ scaling, matching the linear theory in Theorem 6.

Take-Home Message: Practical Recipe for Regularization Strength λ

We observe that the optimal regularization scaling $\lambda^\star(T) \asymp T$ persists across architectures, potentially up to log factors.[a] To determine the leading constants, first tune λ on a short task horizon and then scale it for longer horizons.
For example, for $T = 1000$, tune on $T = 10$ and set $\lambda(1000) = \frac{1000}{10}\,\widehat{\lambda^\star}(10) = 100\,\widehat{\lambda^\star}(10)$.

a. While Theorem 6 suggests a $T/\ln T$, other unknown logarithmic terms—e.g., $\|\mathbf{w}^*\|^2$—may dominate the denominator in practice. Consequently, we recommend focusing on the more significant linear scaling in T.

6. Related Work and Discussion

We extend the literature on generalization in statistical continual learning settings (e.g., Lin et al., 2023; Li et al., 2023; Ding et al., 2024; Zhao et al., 2025) by incorporating both *multiple* teachers and L2 regularization. While optimization-focused studies typically provide worst-case upper bounds (e.g., Evron et al., 2022, 2026; Levinstein et al., 2025), our statistical approach enables an *exact* average-case generalization analysis. This allows us to establish, to the best of our knowledge, the first characterization of optimal regularization strength, spanning both classical single-teacher and general i.i.d. multi-teacher frameworks. A comparative overview is provided in Table 1.

Optimal regularization. Much of the research on continual learning has focused on designing regularization matrices that perform well empirically (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018). Such work typically emphasizes the regularization *directions* or relative magnitudes of coefficients, while leaving the overall strength λ to hyperparameter tuning. In contrast, L2 regularization, which also exhibits strong empirical performance (Hsu et al., 2018; Lubana et al., 2021; Smith et al., 2023), offers a convenient analytical proxy for studying regularization strength, as it depends only on a single tunable parameter, λ.

Levinstein et al. (2025) studied L2 regularization under random task orderings with arbitrary fixed features and a single noiseless teacher. They showed that a fixed[7] $\lambda \asymp \ln T$ yields near-optimal worst-case convergence guarantees, though they do not rule out the possibility that other fixed strengths could perform better (see their Appendix A). Interestingly, in a similar single-teacher setting with no label noise, our analysis dictates that $\lambda^\star \to 0$ is the optimal strength for generalization (Theorem 6). We find that $\lambda^\star > 0$ emerges only in the presence of non-zero label noise or teachers that vary between tasks. Indeed, in our average-case setting, while regularization slows down convergence, it decreases the asymptotic error caused by label noise or teacher variation. Finally, under a *cyclic* task ordering with N cycles over a sequence of T tasks, Cai and Diakonikolas (2025) suggest a fixed $\lambda \asymp T\sqrt{\ln N}$. While their cyclic ordering and our arbitrary one are not directly comparable, their linear dependence on T appears mostly consistent with our findings.

7. As in our work, their strength is static during training and determined a priori by the horizon T. In contrast, they also employ a dynamic strength—increasing at every iteration—to achieve an optimal worst-case rate.

Table 1: **Comparison to related work on continual linear regression.** We highlight papers most closely related to our work—primarily ones that also adopt a statistical perspective. The second column shows the studied regularization method, and the number of regularization parameters it requires.

Paper	**Regularization** (# reg. parameters)	**Multiple teachers?**	**Label noise?**	**Features**
Lin et al. (2023)	None (N/A)	✓	✓	Gaussian i.i.d. entries
Li et al. (2023)	L2 (scalar)	✗	✓	Fixed matrices with commutable covariance (2 tasks only)
Zhao et al. (2024)	Aligned with covariance (matrix)	✗	✓	Fixed matrices with commutable covariance
Ding et al. (2024)	Implicit from step budget (⋆) (N/A)	✗	✓	Random matrices with commutable covariance (⋆⋆)
Zhu et al. (2025)	Based on all directions (matrix)	✓	✓	Any
Levinstein et al. (2025)	L2 (scalar)	✗	✗	Any fixed matrices (in random task ordering)
Zhao et al. (2025) (Concurrent)	L2 (scalar)	✗	✓	Random matrices with commutable covariance
Ours	L2 (scalar)	✓	✓	Zero-mean i.i.d. entries

(⋆) A finite number of steps induces an implicit bias towards previous solutions (Jung et al., 2025).
(⋆⋆) Their main proof implicitly assumes commutable covariance matrices in Eq. (13).

Regularization weighting matrices and covariance structure. Zhao et al. (2024); Zhu et al. (2025) build a weighting matrix directly from the empirical covariance, leading to a worst-case requirement of d^2 regularization parameters. However, this might be considered as "too much" memory for realizable linear regression, as with so many parameters we can avoid forgetting entirely (Evron et al. 2023, Proposition 5.5). In contrast, we use a practical single regularization parameter.

It is also important to distinguish between the observed empirical covariance $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ and the population covariance that characterizes the sample-generating distribution. For example, we explore L2 regularization, i.e., a weighting matrix of $\lambda\mathbf{I}$ assuming a feature *population* covariance of $v_x\mathbf{I}$. Alternatively, some prior work assumes a design with *fixed* features, but require that the *empirical* covariance matrices commute (e.g., Li et al., 2023, 2025; Zhao et al., 2024).

Beyond isotropic covariance. We assumed a $\mathbf{\Sigma} = v_x\mathbf{I}$ feature covariance. However, our analysis can extend to a known, spectrally bounded covariance matrix $c\mathbf{I} \preceq \mathbf{\Sigma} \preceq C\mathbf{I}$, by naturally modifying the learning algorithm to include a whitening step. That is, apply the same learning algorithm, $\mathbf{w}_t = \operatorname{argmin}\left\{\left\|\tilde{\mathbf{X}}_t\mathbf{w} - \mathbf{y}_t\right\|^2 + \lambda d\left\|\mathbf{w} - \mathbf{w}_{t-1}\right\|^2\right\}$, with whitened matrices $\tilde{\mathbf{X}}_t = \mathbf{X}_t\mathbf{\Sigma}^{-1/2}$.

Connections to prior-based methods. Our work analyzes L2 regularization in continual learning, where an additional $\lambda\left\|\mathbf{w} - \mathbf{w}_{t-1}\right\|^2$ term acts as a dynamic *prior* to stabilize updates. This formulation is fundamentally a proximal method (Cai and Diakonikolas, 2025) and shares heritage with online and meta-learning frameworks that utilize static or dynamic priors, such as FOBOS (Duchi and Singer, 2009), FTRL and FTRL-Proximal (McMahan, 2011), and ARUBA (Khodak et al., 2019). Although those methods may provide stronger *regret* guarantees than our proximal method, they depart from standard regularization methods in continual learning (e.g., Kirkpatrick et al., 2017), especially ones involving isotropic regularization (e.g., Smith et al., 2023). While exploring the connections between meta-learning and non-isotropic weight matrices remains a promising direction for future work, our current setting isolates the effects of regularization *strength*, providing direct practical implications for common training recipes.

Task typicality and ordering. While we do not focus on aspects of task ordering, our Theorem 3 nevertheless provides insights into how the teacher ordering influences generalization. For example, the theorem decomposes the loss and reveals a *teacher variability* term that is minimized when the teacher of the *final* task is the one closest to the mean of teachers $\bar{\mathbf{w}}^\star$. In other words, learning should end with the most *typical* task, matching recent findings by Li and Hiratani (2025).

Label noise vs. teacher variance. The teacher variance and label noise are naturally related. In our setting (Section 2.1), the labels of task $t \in [T]$ are generated as $\mathbf{y}_t = \mathbf{X}_t\mathbf{w}_t^\star + \mathbf{z}_t$ for a teacher $\mathbf{w}_t^\star$ and random label noise $\mathbf{z}_t$. If the teacher itself exhibits noise as in Section 3.1, labels can be rewritten as $\mathbf{y}_t = \mathbf{X}_t\big(\mathbf{w}^\star + \boldsymbol{\xi}_t\big) + \mathbf{z}_t = \mathbf{X}_t\mathbf{w}^\star + \tilde{\mathbf{z}}_t$, where $\tilde{\mathbf{z}}_t = \mathbf{X}_t\boldsymbol{\xi}_t + \mathbf{z}_t$. However, these two sources are not equivalent: the label noise $\mathbf{z}_t$ is independent of the design matrix $\mathbf{X}_t$, while the effective noise $\tilde{\mathbf{z}}_t$ inherently *depends* on it, leading to non-trivial theoretical and practical implications.

Theoretical vs. empirical scaling We demonstrate that the qualitative prediction of Theorem 6—namely, that $\lambda^\star(T) \asymp T/\ln T$—is consistent across both our linear regression and neural network experiments (Figure 3). However, the *quantitative* prediction—i.e., the exact value accounting for problem-dependent factors—does not align perfectly with the empirical optimum, as illustrated in Figure 4 of Section B.1. To isolate the source of this mismatch, we show that second-order feature

correlations are largely responsible: by introducing a preprocessing whitening step (see earlier discussion), the observed optimal regularization aligns with the theoretical prediction (see Figure 5 in Section B.2). Finally, we validate a nearly linear dependence of the empirically optimal regularization on the aspect ratio α (Figure 6 in Section B.3), in agreement with the scaling suggested by Theorem 6.

Future directions. In this work, we analyzed continual learning in a simplified setting with a linear model and isotropic regularization, assuming i.i.d. features and i.i.d. teachers. Despite these simplifications, we extracted a meaningful scaling law that appears to hold in practical neural network settings: the optimal regularization strength increases nearly linearly with the number of tasks T. Future research could relax these assumptions by considering non-linear models, more expressive non-isotropic regularization, or non-i.i.d. environments (e.g., Markovian task sequences). We hope our analysis provides a foundation for more robust algorithms and a deeper theoretical understanding of continual learning.

Acknowledgments

The research of DS was funded by the European Union (ERC, A-B-C-Deep, 101039436). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (ERCEA). Neither the European Union nor the granting authority can be held responsible for them.

The research of RM was partially supported by the Israel Science Foundation (grant 1693/22) and by the Skillman chair.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018. (cited on p. 11)

João Carlos Alves Barata and Mahir Saleh Hussein. The moore–penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics*, 42(1):146–165, 2012. (cited on p. 30)

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 118(37):e2100030118, 2021. (cited on p. 3)

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020. (cited on p. 3)

Frederik Benzing. Unifying regularisation methods for continual learning. *AISTATS*, 2022. (cited on p. 3)

Xufeng Cai and Jelena Diakonikolas. Last iterate convergence of incremental methods and applications in continual learning. In *The Thirteenth International Conference on Learning Representations*, 2025. (cited on p. 11, 12)

Romain Couillet and Zhenyu Liao. *Random matrix methods for machine learning*. Cambridge University Press, 2022. (cited on p. 6, 22, 23)

Meng Ding, Kaiyi Ji, Di Wang, and Jinhui Xu. Understanding forgetting in continual learning with linear regression. In *Forty-first International Conference on Machine Learning*, 2024. (cited on p. 2, 10)

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 1072–1080, 2021. (cited on p. 1)

John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(99):2899–2934, 2009. (cited on p. 12)

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory (COLT)*, pages 4028–4079. PMLR, 2022. (cited on p. 1, 2, 3, 4, 8, 10)

Itay Evron, Edward Moroshko, Gon Buzaglo, Maroun Khriesh, Badea Marjieh, Nathan Srebro, and Daniel Soudry. Continual learning in linear classification on separable data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 9440–9484. PMLR, 23–29 Jul 2023. (cited on p. 1, 12)

Itay Evron, Ran Levinstein, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Nathan Srebro. Better rates for random task orderings in continual linear models. In *37th International Conference on Algorithmic Learning Theory*, 2026. (cited on p. 2, 4, 8, 10)

Lior Friedman and Ron Meir. Data-dependent and oracle bounds on forgetting in continual learning. In *Proceedings of The 4th Conference on Lifelong Learning Agents*, 2025. (cited on p. 3)

Daniel Goldfarb and Paul Hand. Analysis of catastrophic forgetting for random orthogonal transformation tasks in the overparameterized regime. In *International Conference on Artificial Intelligence and Statistics*, pages 2975–2993. PMLR, 2023. (cited on p. 2, 3)

Daniel Goldfarb, Itay Evron, Nir Weinberger, Daniel Soudry, and Paul Hand. The joint effect of task similarity and overparameterization on catastrophic forgetting - an analytical model. In *The Twelfth International Conference on Learning Representations*, 2024. (cited on p. 1, 2, 4)

Naoki Hiratani. Disentangling and mitigating the impact of task similarity for continual learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. (cited on p. 1)

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1. JMLR Workshop and Conference Proceedings, 2012. (cited on p. 3)

Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*, 2018. (cited on p. 2, 3, 11)

Hyunji Jung, Hanseul Cho, and Chulhee Yun. Convergence and implicit bias of gradient descent on continual linear classification. In *The Thirteenth International Conference on Learning Representations*, 2025. (cited on p. 1)

Mikhail Khodak, Maria-Florina Balcan, and Ameet Talwalkar. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. (cited on p. 12)

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwińska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. (cited on p. 2, 3, 11, 12)

Mark Kong, William Swartworth, Halyun Jeong, Deanna Needell, and Rachel Ward. Nearly optimal bounds for cyclic forgetting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. (cited on p. 1, 2)

Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. (cited on p. 3)

Ran Levinstein, Amit Attia, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Itay Evron. Optimal rates in continual linear regression via increasing regularization. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025. (cited on p. 1, 2, 3, 10, 11)

Haoran Li, Jingfeng Wu, and Vladimir Braverman. Fixed design analysis of regularization-based continual learning. In Sarath Chandar, Razvan Pascanu, Hanie Sedghi, and Doina Precup, editors, *Proceedings of The 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pages 513–533. PMLR, 22–25 Aug 2023. (cited on p. 1, 2, 3, 10, 12)

Haoran Li, Jingfeng Wu, and Vladimir Braverman. Memory-statistics tradeoff in continual learning with structural regularization. *arXiv preprint arXiv:2504.04039*, 2025. (cited on p. 1, 12)

Ziyan Li and Naoki Hiratani. Optimal task order for continual learning of multiple tasks. In *Forty-second International Conference on Machine Learning*, 2025. (cited on p. 12)

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21078–21100. PMLR, 23–29 Jul 2023. (cited on p. 1, 2, 3, 4, 6, 8, 10, 30)

Ekdeep Singh Lubana, Puja Trivedi, Danai Koutra, and Robert P. Dick. How do quadratic regularizers prevent catastrophic forgetting: The role of interpolation. In *ICML Workshop on Theory and Foundations of Continual Learning*, 2021. (cited on p. 2, 3, 11)

Volodomyr Marchenko and Leonid Pastur. Distribution of eigenvalues for some ensembles of random matrices. *Matematicheskii Sbornik*, 72(4):507–536, 1967. In Russian. (cited on p. 4)

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. (cited on p. 1)

Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 525–533. JMLR Workshop and Conference Proceedings, 2011. (cited on p. 12)

Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects, 2013. (cited on p. 2)

Liangzu Peng, Paris Giampouras, and René Vidal. The ideal continual learner: An agent that never forgets. In *International Conference on Machine Learning*, 2023. (cited on p. 1)

James Seale Smith, Junjiao Tian, Yen-Chang Hsu, and Zsolt Kira. A closer look at rehearsal-free continual learning. In *CVPR Workshop on Continual Learning in Computer Vision*, 2023. (cited on p. 2, 3, 11, 12)

Matan Tsipory, Ran Levinstein, Itay Evron, Mark Kong, Deanna Needell, and Daniel Soudry. Are greedy task orderings better than random in continual linear regression? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. (cited on p. 1)

Gido M van de Ven, Nicholas Soures, and Dhireesha Kudithipudi. Continual learning and catastrophic forgetting. *arXiv preprint arXiv:2403.05175*, 2024. (cited on p. 1)

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. (cited on p. 1)

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017. (cited on p. 11)

Wenshuai Zhao, Yilu Qu, Jing Liu, Hongyi Li, Jian Luo, Peng Yang, and Yu-Huan Zeng. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. *arXiv preprint arXiv:2003.09555*, 2020. (cited on p. 6)

Xuyang Zhao, Huiyuan Wang, Weiran Huang, and Wei Lin. A statistical theory of regularization-based continual learning. In *Forty-first International Conference on Machine Learning*, 2024. (cited on p. 2, 3, 12)

Yihan Zhao, Wenqing Su, and Ying Yang. High-dimensional asymptotics of generalization performance in continual ridge regression. *arXiv preprint arXiv:2508.15494*, 2025. (cited on p. 2, 10)

Fei Zhu, Yujing Liu, Wenzhuo Liu, and Zhaoxiang Zhang. Global convergence of continual learning on non-iid data. *arXiv preprint arXiv:2503.18511*, 2025. (cited on p. 12)

Appendix Contents

A Table of Notations **18**

B Additional Continual Linear Regression Experiments **19**

B.1 Theory vs. Experiments: Qualitative, but not Quantitative, Alignment 19

B.2 A Whitening Step Reconciles Experiments with Theory 20

B.3 Aspect-Ratio Scaling Predicted by Theory Extends Beyond i.i.d. Features 20

C Proving the Main Result (Theorem 3) **21**

C.1 Proof of Example 1 30

D Auxiliary Lemmas **33**

E Proofs for Section 3.1 (Special Case: Multiple i.i.d. Teachers) **41**

E.1 Proof of Corollary 4: Generalization under i.i.d. Teachers 41

E.2 Generalization with an Infinite Task Horizon 43

E.3 Proof of Theorem 7: Asymptotic Generalization under Strong Regularization . . . 44

E.4 Proof of Lemma 5: More Tasks Can Be Useful or Harmful 48

E.5 Auxiliary Lemma: Optimal Regularization is Near Linear in T 52

E.6 Proof of Theorem 6: Key Result on Optimal Regularization Scaling 58

Appendix A. Table of Notations

Table 2: Notation table.

Notation / Definition	**Short description**
d	Dimension (number of input features)
n	Number of data items (samples) per task
$\alpha \triangleq \dfrac{n}{d}$	Aspect ratio (sample-to-dimension ratio)
T	Number of iterations (tasks)
v_z	Label noise variance
v_x	Feature noise variance
λ	Regularization strength
$\mathbf{X}$	Data set (feature matrix)
$\mathbf{y}$	Labels
$\mathbf{z}$	Label noise vector
$\mathbf{w}$	Predictor (model parameters)
$\mathbf{w}^\star$	Teacher (ground-truth parameter)
$\bar{\mathbf{w}}^\star \triangleq \frac{1}{T}\sum_{i=1}^{T} \mathbf{w}_i^\star$	Average teacher
G_t	Generalization error at iteration t
$\mathbf{\Sigma}$	Teacher covariance matrix
$\mathbf{P} \triangleq \lambda d\Big(\mathbf{X}^\top\mathbf{X} + \lambda d\,\mathbf{I}\Big)^{-1}$	Regularized inverse covariance matrix
$\tilde{\lambda} \triangleq \dfrac{\lambda}{v_x}$	Scaled regularization strength
$\boldsymbol{\xi}$	Random variable modeling teacher noise
$D \triangleq \sqrt{\tilde{\lambda}^2 + 2\tilde{\lambda}(1+\alpha) + (1-\alpha)^2}$	Discriminant term appearing in closed-form expressions
$a \triangleq \frac{1}{2}\left(1 - \alpha + \dfrac{\tilde{\lambda}(1+\alpha) + (1-\alpha)}{D}\right)$	Memory contraction coefficient
$b \triangleq \frac{1}{2}\Big(1 - \alpha - \tilde{\lambda} + D\Big)$	Weight assigned to task-to-task drift
$c \triangleq \frac{1}{2}\Big((1+\alpha) + \tilde{\lambda} - D\Big)$	Noise amplification coefficient
$f \triangleq v_z\, c$	Effective label-noise contribution

Appendix B. Additional Continual Linear Regression Experiments

In this appendix, we experiment with the regularized continual linear regression model (Scheme 1). Compared to the previous setting of Section 5.1, we introduce additive Gaussian label noise with variance 0.2, and construct each task from a *different* subsample of images. We report the generalization error corresponding to the mean squared error (MSE). Furthermore, the teacher is evaluated on a random subset of 500 sampled data points.

B.1. Theory vs. Experiments: Qualitative, but not Quantitative, Alignment

We train the model on the *normalized*[8] `MNIST` dataset. We visualize the empirical generalization error[9] as a heat map over the regularization strength and number of tasks. For comparison, we also plot a curve of the optimal regularization strength predicted by Theorem 6.

The experiment demonstrates that in the absence of whitening (see next section), there is a mismatch between the empirically optimal regularization and the theoretical prediction.

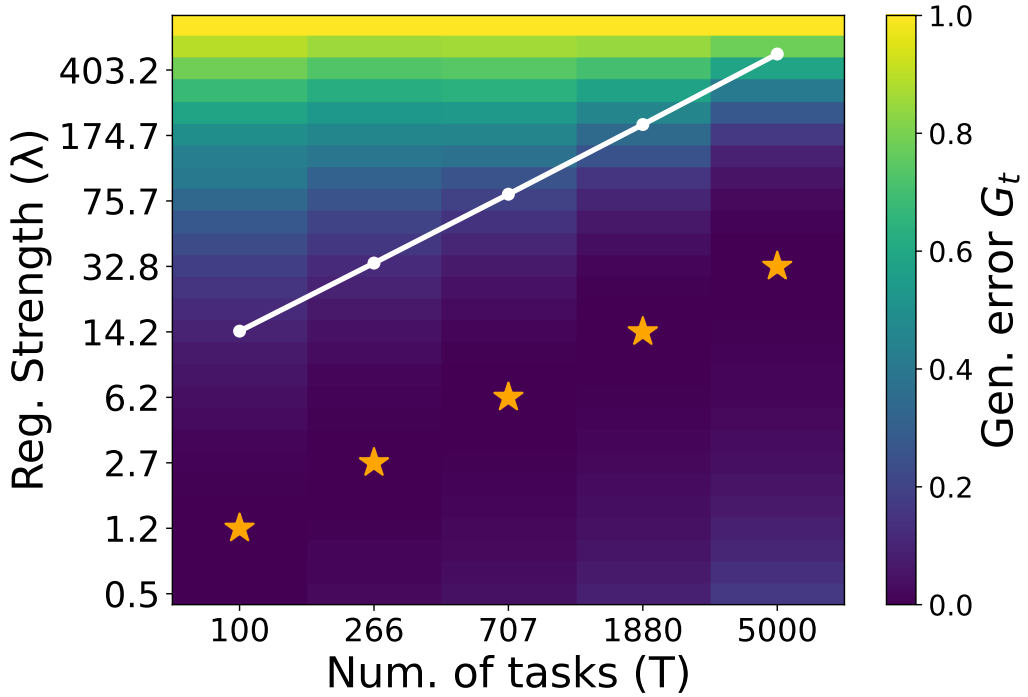


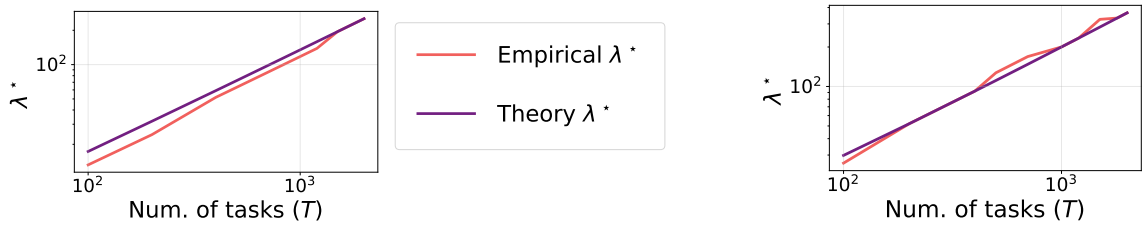
Figure 4: **Regularization strength vs. Task horizon.** Stars mark the empirical optimal regularization strength obtained after training on a noisy sequence of T tasks and averaging over 5 random seeds. We also plot the analytical optima, predicted by our Theorem 6; shown as a white curve. We observe a seemingly-multiplicative mismatch between the empirical and analytical optima.

8. By normalization we mean per-feature centering and scaling to unit variance, i.e., $x_j \leftarrow (x_j - \hat{\mathbb{E}}[x_j])/\sqrt{\widehat{\mathrm{Var}}(x_j)}$; unlike whitening, this preprocessing does not remove cross-feature correlations.
9. The empirical generalization error is normalized by the maximum empirical generalization error across all regularization strengths for the specific task horizon T. That is, each column in the heatmap is scaled between 0 and 1.

B.2. A Whitening Step Reconciles Experiments with Theory

Here, prior to training, the input images are whitened using the empirical feature covariance computed over the entire dataset of MNIST. The empirical optimal strength is obtained by searching over a (one dimensional) grid centered at the theoretically predicted optimum, and compared to the theoretical optimal regularization predicted by Theorem 6.

We observe that the optimal regularization strength matches the prediction of Theorem 6 almost perfectly. This implies that the mismatch in the previous section stems largely from second-order feature correlations (while our theory assumes a isotropic covariance matrix).



(a) Single teacher. (b) Multiple i.i.d. teachers.

Figure 5: **Empirical optimal regularization strength vs. Task horizon.** We plot the empirical optimal strength after learning a noisy sequence with T tasks—averaged over 5 seeds—and compare it to the optimum predicted by our analysis.

B.3. Aspect-Ratio Scaling Predicted by Theory Extends Beyond i.i.d. Features

We evaluate a linear model trained on the original *non-whitened* MNIST-based data, but vary the aspect ratio $\frac{n}{d} \to \alpha$. Here, we focus on the single noiseless teacher. Still, the optimal regularization exhibits a near linear dependence on α, in agreement with the scaling predicted by Theorem 6.

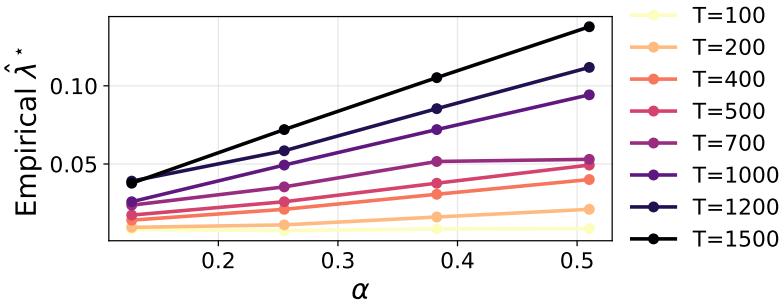


Figure 6: **Effect of aspect ratio on optimal regularization strength.** We plot the empirical optimal regularizer after learning a noisy sequence with T tasks. Each curve reports the average empirical and theoretical values over 5 random seeds.

Appendix C. Proving the Main Result (Theorem 3)

Recall Theorem 3. Recall the high-dimensional statistical setting of Section 2.1, where the number of samples and the dimension $n, d \to \infty$ and $\frac{n}{d} \to \alpha \leq 1$. Moreover, recall that v_x, v_z are the variances of features and labels, respectively. Then, the expected generalization loss (Definition 2) after learning T tasks with Scheme 1 is,

$$\mathbb{E}[G_T] = \underbrace{v_z c \frac{1-a^T}{1-a}}_{\text{Label noise term}} + \underbrace{\frac{1}{T}\sum_{i=1}^{T} \|\mathbf{w}_T^\star - \mathbf{w}_i^\star\|^2}_{\text{Teacher variability term}} + \underbrace{2\left(\mathbf{w}_T^\star - \bar{\mathbf{w}}^\star\right)^\top \sum_{i=1}^{T} b^{T-i+1}\left(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\right)}_{\text{Interaction term}}$$

$$+ \underbrace{\sum_{i=1}^{T}\sum_{i'=1}^{T} a^{T-\max(i,i')+1}\, b^{|i-i'|}\left(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\right)^\top \left(\mathbf{w}_{i'-1}^\star - \mathbf{w}_{i'}^\star\right)}_{\text{Accumulated temporal (teacher) correlations}},$$

where we define $\mathbf{w}_0^\star \triangleq \mathbf{w}_0$ (by abuse of notation), the average teacher $\bar{\mathbf{w}}^\star \triangleq \frac{1}{T}\sum_{i=1}^{T} \mathbf{w}_i^\star$, and

$$a \triangleq \frac{1}{2}\left(1 - \alpha + \frac{\lambda(1+\alpha) + v_x(1-\alpha)^2}{\sqrt{\lambda^2 + 2\lambda v_x\,(1+\alpha) + v_x^2\,(1-\alpha)^2}}\right),$$

$$b \triangleq \frac{1}{2}\left(1 - \alpha - \frac{\lambda}{v_x} + \frac{1}{v_x}\sqrt{\lambda^2 + 2\lambda v_x(1+\alpha) + v_x^2(1-\alpha)^2}\right),$$

$$c \triangleq \frac{1}{2v_x}\left(\frac{v_x(1+\alpha) + \lambda}{\sqrt{\lambda^2 + 2\lambda v_x(1+\alpha) + v_x^2(1-\alpha)^2}} - 1\right).$$

We begin by proving the following Lemma, which provides a closed-form expression for the predictor update at each iteration t.

Lemma C.1 *At each iteration t, the iterative update under Scheme 1 is given by*

$$\mathbf{w}_t = \left(\mathbf{X}_t^\top \mathbf{X}_t + \lambda d \mathbf{I}\right)^{-1}\left(\mathbf{X}_t^\top \mathbf{y}_t + \lambda d \mathbf{w}_{t-1}\right).$$

Proof. Define the convex objective

$$J_t(\mathbf{w}) = \|\mathbf{X}_t \mathbf{w} - \mathbf{y}_t\|^2 + \lambda d\, \|\mathbf{w} - \mathbf{w}_{t-1}\|^2.$$

Its gradient is

$$\begin{aligned}\nabla_{\mathbf{w}} J_t(\mathbf{w}) &= 2\,\mathbf{X}_t^\top\left(\mathbf{X}_t \mathbf{w} - \mathbf{y}_t\right) + 2\lambda d\left(\mathbf{w} - \mathbf{w}_{t-1}\right) \\ &= 2\left[\left(\mathbf{X}_t^\top \mathbf{X}_t + \lambda d \mathbf{I}\right)\mathbf{w} - \left(\mathbf{X}_t^\top \mathbf{y}_t + \lambda d \mathbf{w}_{t-1}\right)\right].\end{aligned}$$

At an optimizer $\mathbf{w}_t$ we must have $\nabla_{\mathbf{w}} J_t(\mathbf{w}_t) = \mathbf{0}$, hence the normal equations

$$\left(\mathbf{X}_t^\top \mathbf{X}_t + \lambda d \mathbf{I}\right)\mathbf{w}_t = \mathbf{X}_t^\top \mathbf{y}_t + \lambda d \mathbf{w}_{t-1}.$$

If $\lambda > 0$, then $\mathbf{X}_t^\top \mathbf{X}_t + \lambda d \mathbf{I}$ is symmetric positive definite and invertible, so

$$\mathbf{w}_t = \left(\mathbf{X}_t^\top \mathbf{X}_t + \lambda d \mathbf{I}\right)^{-1} \left(\mathbf{X}_t^\top \mathbf{y}_t + \lambda d \mathbf{w}_{t-1}\right).$$

■

Definition 2 (Matrix Equivalent (Couillet and Liao, 2022, Notation 1)) *For $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times n}$ (deterministic or random matrices), we say that $\mathbf{X} \leftrightarrow \mathbf{Y}$ if, for all $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ of unit norms, we have*

$$\frac{1}{n} \mathrm{Tr}\left(\mathbf{A}\left(\mathbf{X} - \mathbf{Y}\right)\right) \to 0, \quad \mathbf{a}^\top \left(\mathbf{X} - \mathbf{Y}\right) \mathbf{b} \to 0, \quad \left\|\mathbb{E}\left[\mathbf{X} - \mathbf{Y}\right]\right\| \to 0$$

where, for random variables, the convergence is almost surely.

Proof for Theorem 3. We now turn to prove the theorem. First, we aim to derive a formula independent of the predictors at each iteration.

According to Scheme 1, at each iteration t the predictor is updated as

$$\mathbf{w}_t \leftarrow \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \|\mathbf{X}_t \mathbf{w} - \mathbf{y}_t\|^2 + \lambda d \|\mathbf{w} - \mathbf{w}_{t-1}\|^2 \right\}$$

From Lemma C.1, the solution to this problem is given by:

$$\mathbf{w}_t = \left(\mathbf{X}_t^\top \mathbf{X}_t + \lambda d \mathbf{I}\right)^{-1} \left(\mathbf{X}_t^\top \mathbf{y}_t + \lambda d \mathbf{w}_{t-1}\right)$$

For some $1 \leq k, i \leq T$ we have

$$\begin{aligned}
\mathbf{w}_k - \mathbf{w}_i^\star &= \left(\mathbf{X}_k^\top \mathbf{X}_k + \lambda d \mathbf{I}\right)^{-1} \left(\mathbf{X}_k^\top \left(\mathbf{X}_k \mathbf{w}_k^\star + \mathbf{z}_k\right) + \lambda d \mathbf{w}_{k-1}\right) - \mathbf{w}_i^\star \\
&= \left(\mathbf{X}_k^\top \mathbf{X}_k + \lambda d \mathbf{I}\right)^{-1} \left(\mathbf{X}_k^\top \left(\mathbf{X}_k \mathbf{w}_k^\star + \mathbf{z}_k\right) + \lambda d \mathbf{w}_{k-1}\right) \\
&\qquad - \left(\mathbf{X}_k^\top \mathbf{X}_k + \lambda d \mathbf{I}\right)^{-1} \left(\mathbf{X}_k^\top \mathbf{X}_k + \lambda d \mathbf{I}\right) \mathbf{w}_i^\star \\
&= \left(\mathbf{X}_k^\top \mathbf{X}_k + \lambda d \mathbf{I}\right)^{-1} \left[\mathbf{X}_k^\top \left(\mathbf{X}_k \mathbf{w}_k^\star + \mathbf{z}_k\right) + \lambda d \mathbf{w}_{k-1} - \left(\mathbf{X}_k^\top \mathbf{X}_k + \lambda d \mathbf{I}\right) \mathbf{w}_i^\star\right] \\
&= \left(\mathbf{X}_k^\top \mathbf{X}_k + \lambda d \mathbf{I}\right)^{-1} \left[\mathbf{X}_k^\top \mathbf{z}_k + \lambda d \mathbf{w}_{k-1} - \lambda d \mathbf{w}_i^\star + \mathbf{X}_k^\top \mathbf{X}_k \left(\mathbf{w}_k^\star - \mathbf{w}_i^\star\right)\right] \\
&= \left(\mathbf{X}_k^\top \mathbf{X}_k + \lambda d \mathbf{I}\right)^{-1} \left[\mathbf{X}_k^\top \mathbf{z}_k + \lambda d \left(\mathbf{w}_{k-1} - \mathbf{w}_i^\star\right) + \mathbf{X}_k^\top \mathbf{X}_k \left(\mathbf{w}_k^\star - \mathbf{w}_i^\star\right)\right]. \qquad (2)
\end{aligned}$$

Denote

$$\mathbf{P}_t = \lambda d \left(\mathbf{X}_t^\top \mathbf{X}_t + \lambda d \mathbf{I}\right)^{-1}.$$

Next, we derive the deterministic equivalent for $\mathbf{P}_t$; see Definition 2. Let $\mathbf{Z} \in \mathbb{R}^{n \times d}$ with entries Z_{ij} that are i.i.d. with mean 0 and unit variance, and having a finite moment of order $4 + \epsilon$ for some $\epsilon > 0$. Define the resolvent

$$Q(z) = \left(\tfrac{1}{n} \mathbf{Z}^\top \mathbf{Z} - z \mathbf{I}_d\right)^{-1}.$$

Then, by Theorem 2.6 from Couillet and Liao (2022): as $n, d \to \infty$ with $d/n \to \frac{1}{\alpha} \in (0, \infty)$, $Q(z)$ admits a deterministic equivalent

$$Q(z) \leftrightarrow \bar{Q}(z) \;=\; m_{\frac{1}{\alpha}}(z)\,\mathbf{I}_d, \tag{3}$$

where $m_{\frac{1}{\alpha}}(z)$ satisfies the following closed-form expression:

$$m_{\frac{1}{\alpha}}(z) = \frac{-\left(1-\frac{1}{\alpha}-z\right)+\sqrt{\left(1-\frac{1}{\alpha}-z\right)^2-\frac{4}{\alpha}z}}{-\frac{2}{\alpha}z}. \tag{4}$$

Since $X_{i,j}$ has mean 0 and variance v_x, we have $\mathbf{X} \overset{d}{=} \sqrt{v_x}\,\mathbf{Z}$. We aim to derive closed-form expressions for $\mathbb{E}[\mathbf{P}]$, $\mathbb{E}[\mathbf{P}^2]$ and $\mathbb{E}\left[\frac{1}{(\lambda d)^2}\mathbf{P}\mathbf{X}^\top\mathbf{X}\mathbf{P}\right]$, which will be used in our subsequent derivations. In our setting (Section 2.1), $\mathbf{P}$ denotes the random resolvent $\mathbf{P} \triangleq \lambda d(\mathbf{X}^\top\mathbf{X}+\lambda d\,\mathbf{I})^{-1}$. Expectations are taken with respect to the limiting spectral distribution of $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ (Marchenko–Pastur law, Eq. (3)), under which $\mathbb{E}[\mathbf{P}]$ and $\mathbb{E}[\mathbf{P}^2]$ exist and are scalar multiples of the identity, as we show below.

Using the explicit form of $m_{\frac{1}{\alpha}}(z)$ in Eq. (4) we obtain,

$$\begin{aligned}
\mathbb{E}[\mathbf{P}] &= \mathbb{E}\left[\lambda d\left(v_x\mathbf{Z}^\top\mathbf{Z}-(-\lambda d)\,\mathbf{I}\right)^{-1}\right] = \frac{\lambda d}{n v_x}\mathbb{E}\left[\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}-\left(-\frac{\lambda d}{n v_x}\right)\mathbf{I}\right)^{-1}\right] \\
&= \frac{\lambda d}{n v_x} m_{\frac{1}{\alpha}}\left(-\frac{\lambda d}{n v_x}\right)\mathbf{I} \\
&= \frac{-\left(1-\frac{1}{\alpha}+\frac{\lambda d}{n v_x}\right)+\sqrt{\left(1-\frac{1}{\alpha}+\frac{\lambda d}{n v_x}\right)^2+\frac{4}{\alpha}\frac{\lambda d}{n v_x}}}{\frac{2}{\alpha}}\mathbf{I} \\
&= \frac{1}{2}\left(1-\alpha-\frac{\lambda}{v_x}+\sqrt{\left(\frac{\lambda}{v_x}\right)^2+\frac{2\lambda}{v_x}(\alpha+1)+(1-\alpha)^2}\right)\mathbf{I}
\end{aligned} \tag{5}$$

and,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{P}^2\right] &= \mathbb{E}\left[(\lambda d)^2\left(\lambda d\mathbf{I}+v_x\mathbf{Z}^\top\mathbf{Z}\right)^{-2}\right] = \mathbb{E}\left[-\frac{(\lambda d)^2}{d}\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(-(-\lambda d)\,\mathbf{I}+v_x\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\right] \\
&= -\frac{\lambda^2 d}{n v_x}\mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(-\left(-\frac{\lambda d}{n v_x}\right)\mathbf{I}+\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\right] = \left(\frac{\lambda d}{n v_x}\right)^2 m'_{\frac{1}{\alpha}}\left(-\frac{\lambda d}{n v_x}\right)\mathbf{I} \\
&= \frac{\alpha}{2}\,\frac{\left(1+\frac{1}{\alpha}\right)\frac{\lambda d}{n v_x}+\left(1-\frac{1}{\alpha}\right)\left(\left(1-\frac{1}{\alpha}\right)-\sqrt{\left(\frac{\lambda d}{n v_x}\right)^2+2\left(1+\frac{1}{\alpha}\right)\frac{\lambda d}{n v_x}+\left(1-\frac{1}{\alpha}\right)^2}\right)}{\sqrt{\left(\frac{\lambda d}{n v_x}\right)^2+2\left(1+\frac{1}{\alpha}\right)\frac{\lambda d}{n v_x}+\left(1-\frac{1}{\alpha}\right)^2}}\,\mathbf{I} \\
&= \frac{1}{2}\left(1-\alpha+\frac{(\alpha+1)\frac{\lambda}{v_x}+(\alpha-1)^2}{\sqrt{\left(\frac{\lambda}{v_x}\right)^2+\frac{2\lambda}{v_x}(\alpha+1)+(\alpha-1)^2}}\right)\mathbf{I}\,.
\end{aligned} \tag{6}$$

The term $\mathbb{E}\left[\frac{1}{(\lambda d)^2}\mathbf{P}\mathbf{X}^\top\mathbf{X}\mathbf{P}\right]$ is a linear combination of $\mathbb{E}[\mathbf{P}]$ and $\mathbb{E}[\mathbf{P}^2]$,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{(\lambda d)^2}\mathbf{P}\mathbf{X}^\top\mathbf{X}\mathbf{P}\right] &= \mathbb{E}\left[\left(\lambda d\mathbf{I}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\lambda d\mathbf{I}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\right] \\
&= \mathbb{E}\left[\left(\lambda d\mathbf{I}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\left(\lambda \mathbf{d}\mathbf{I}+\mathbf{X}^\top\mathbf{X}\right)-\lambda d\mathbf{I}\right)\left(\lambda d\mathbf{I}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\right] \\
&= \mathbb{E}\left[\left(\lambda d\mathbf{I}+\mathbf{X}^\top\mathbf{X}\right)^{-1}-\lambda d\left(\lambda d\mathbf{I}+\mathbf{X}^\top\mathbf{X}\right)^{-2}\right] \\
&= \frac{1}{\lambda d}\left[\mathbb{E}\left[\mathbf{P}\right]-\mathbb{E}\left[\mathbf{P}^2\right]\right]\mathbf{I} \\
&= \frac{1}{2dv_x}\left(\frac{v_x\left(\alpha+1\right)+\lambda}{\sqrt{\lambda^2+2\lambda v_x\left(\alpha+1\right)+v_x^2\left(1-\alpha\right)^2}}-1\right)\mathbf{I}.
\end{aligned}$$

We continue with our main result. From Eq. (2) we obtain,

$$\begin{aligned}
&\mathbf{w}_k-\mathbf{w}_i^\star \\
&= \lambda d\left(\mathbf{X}_t^\top\mathbf{X}_t+\lambda d\mathbf{I}\right)^{-1}(\mathbf{w}_{k-1}-\mathbf{w}_i^\star)+\left(\mathbf{X}_t^\top\mathbf{X}_t+\lambda d\mathbf{I}\right)^{-1}\left(\mathbf{X}_k^\top\mathbf{z}_k+\mathbf{X}_k^\top\mathbf{X}_k\left(\mathbf{w}_k^\star-\mathbf{w}_i^\star\right)\right) \\
&= \mathbf{P}_k\left(\mathbf{w}_{k-1}-\mathbf{w}_i^\star\right)+\frac{1}{\lambda d}\mathbf{P}_k\left(\mathbf{X}_k^\top\mathbf{z}_k+\mathbf{X}_k^\top\mathbf{X}_k\left(\mathbf{w}_k^\star-\mathbf{w}_i^\star\right)\right). \qquad (7)
\end{aligned}$$

Setting $i=k$ in Eq. (7) we then obtain,

$$\mathbf{w}_k-\mathbf{w}_k^\star=\mathbf{P}_k\left(\mathbf{w}_{k-1}-\mathbf{w}_k^\star\right)+\frac{1}{\lambda d}\mathbf{P}_k\mathbf{X}_k^\top\mathbf{z}_k. \qquad (8)$$

For all $1\leq i,t\leq T$ using Eq. (8) we have,

$$\mathbf{w}_t-\mathbf{w}_i^\star=\mathbf{P}_t\left(\mathbf{w}_{t-1}-\mathbf{w}_t^\star\right)+\frac{1}{\lambda d}\mathbf{P}_t\mathbf{X}_t^\top\mathbf{z}_t+\left(\mathbf{w}_t^\star-\mathbf{w}_i^\star\right). \qquad (9)$$

From Eq. (9),

$$\begin{aligned}
\mathbf{w}_t-\mathbf{w}_i^\star &= \mathbf{P}_t\left(\mathbf{P}_{t-1}\left(\mathbf{w}_{t-2}-\mathbf{w}_{t-1}^\star\right)+\frac{1}{\lambda d}\mathbf{P}_{t-1}\mathbf{X}_{t-1}^\top\mathbf{z}_{t-1}+\left(\mathbf{w}_{t-1}^\star-\mathbf{w}_t^\star\right)\right) \\
&\quad+\frac{1}{\lambda d}\mathbf{P}_t\mathbf{X}_t^\top\mathbf{z}_t+\left(\mathbf{w}_t^\star-\mathbf{w}_i^\star\right) \\
&\quad\cdots \\
\mathbf{w}_t-\mathbf{w}_i^\star &= \left[\prod_{m=t}^{1}\mathbf{P}_m\right]\left(\mathbf{w}_0-\mathbf{w}_1^\star\right)+\frac{1}{\lambda d}\sum_{k=1}^{t}\left[\prod_{m=t}^{k}\mathbf{P}_m\right]\mathbf{X}_k^\top\mathbf{z}_k+ \\
&\quad\sum_{k=2}^{t}\left[\prod_{m=t}^{k}\mathbf{P}_m\right]\left(\mathbf{w}_{k-1}^\star-\mathbf{w}_k^\star\right)+\left(\mathbf{w}_t^\star-\mathbf{w}_i^\star\right). \qquad (10)
\end{aligned}$$

We proceed to the calculation of the main result $\mathbb{E}\left[G_T\right] = \frac{1}{T}\sum_{i=1}^{T}\mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}_i^\star\|^2\right]$. For convenience, we define $\mathbf{S}_{i:j} \triangleq \prod_{m=i}^{j}\mathbf{P}_m = \mathbf{P}_i \dots \mathbf{P}_j$ for $i \geq j$ and $\mathbf{S}_{i:j} = \mathbf{I}$ for $i < j$. We use the standard notation: for $a, b \in \mathbb{R}$ we write $\max(a,b) = a \vee b$ and $\min(a,b) = a \wedge b$.

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}_i^\star\|^2\right] \\
&= \mathbb{E}\left\|\mathbf{S}_{t:1}\left(\mathbf{w}_0 - \mathbf{w}_1^\star\right) + \frac{1}{\lambda d}\sum_{k=1}^{t}\mathbf{S}_{t:k}\mathbf{X}_k^\top\mathbf{z}_k + \sum_{k=2}^{t}\mathbf{S}_{t:k}\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)\right\|^2 \\
&= \underbrace{\mathbb{E}\left\|\mathbf{S}_{t:1}\left(\mathbf{w}_0 - \mathbf{w}_1^\star\right) + \frac{1}{\lambda d}\sum_{k=1}^{t}\mathbf{S}_{t:k}\mathbf{X}_k^\top\mathbf{z}_k\right\|^2}_{\text{term 1}} \\
&\quad + \underbrace{2\mathbb{E}\left(\sum_{k=2}^{t}\mathbf{S}_{t:k}\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)\right)^\top\left(\mathbf{S}_{t:1}\left(\mathbf{w}_0 - \mathbf{w}_1^\star\right) + \frac{1}{\lambda d}\sum_{k=1}^{t}\mathbf{S}_{t:k}\mathbf{X}_k^\top\mathbf{z}_k\right)}_{\text{term 2}} \\
&\quad + \underbrace{\mathbb{E}\left\|\sum_{k=2}^{t}\mathbf{S}_{t:k}\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)\right\|^2}_{\text{term 3}}. \qquad (11)
\end{aligned}
$$

In the following pages, we derive the terms above, employing the following properties.

[1] For any $1 \leq k \leq t \leq T$,

$$\begin{aligned}
&\mathbb{E}\left[\prod_{m=k}^{t}\mathbf{P}_m\prod_{m=t}^{k}\mathbf{P}_m\right] = \mathbb{E}\left[\mathbf{P}_k\cdots\mathbf{P}_{t-1}\mathbf{P}_t\mathbf{P}_t\mathbf{P}_{t-1}\cdots\mathbf{P}_k\right] \\
&= \mathbb{E}_{\mathbf{P}_k,\ldots,\mathbf{P}_{t-1}}\left[\mathbb{E}_{\mathbf{P}_t}\left[\mathbf{P}_k\cdots\mathbf{P}_{t-1}\mathbf{P}_t\mathbf{P}_t\mathbf{P}_{t-1}\cdots\mathbf{P}_k\middle|\mathbf{P}_k,\mathbf{P}_{k+1},\ldots,\mathbf{P}_{t-1}\right]\right] \\
&[3] = \mathbb{E}\left[\mathbf{P}_k\cdots\mathbf{P}_{t-1}\mathbb{E}\left[\mathbf{P}_t^2\right]\mathbf{P}_{t-1}\cdots\mathbf{P}_k\right] = \mathbb{E}\left[\mathbf{P}_k\cdots\mathbf{P}_{t-1}\mathbb{E}\left[\mathbf{P}^2\right]\mathbf{P}_{t-1}\cdots\mathbf{P}_k\right] \\
&[4] = \mathbb{E}\left[\mathbf{P}^2\right]\mathbb{E}\left[\mathbf{P}_k\cdots\mathbf{P}_{t-1}\mathbf{P}_{t-1}\cdots\mathbf{P}_k\right] = \ldots = \left(\mathbb{E}\left[\mathbf{P}^2\right]\right)^{t-k+1}.
\end{aligned}$$

[2] For any $1 \leq k' < k \leq t \leq T$,

$$\begin{aligned}
&\mathbb{E}\left[\prod_{m=k}^{t}\mathbf{P}_m\prod_{m=t}^{k}\mathbf{P}_m\prod_{m=k-1}^{k'}\mathbf{P}_m\right] = \mathbb{E}\left[\mathbf{P}_k\cdots\mathbf{P}_t\mathbf{P}_t\cdots\mathbf{P}_k\mathbf{P}_{k-1}\cdots\mathbf{P}_{k'}\right] \\
&= \mathbb{E}_{\mathbf{P}_{k'},\ldots,\mathbf{P}_{t-1}}\left[\mathbb{E}_{\mathbf{P}_t}\left[\mathbf{P}_k\cdots\mathbf{P}_t\mathbf{P}_t\cdots\mathbf{P}_k\mathbf{P}_{k-1}\cdots\mathbf{P}_{k'}\middle|\mathbf{P}_{k'},\ldots,\mathbf{P}_{t-1}\right]\right] \\
&[3] = \mathbb{E}_{\mathbf{P}_{k'},\ldots,\mathbf{P}_{t-1}}\left[\mathbb{E}_{\mathbf{P}_t}\left[\mathbf{P}_{k'}\cdots\mathbf{P}_{t-1}\mathbb{E}\left[\mathbf{P}_t^2\right]\mathbf{P}_{t-1}\cdots\mathbf{P}_k\cdots\mathbf{P}_{k'}\middle|\mathbf{P}_{k'},\ldots,\mathbf{P}_{t-1}\right]\right] \\
&[4] = \mathbb{E}\left[\mathbf{P}^2\right]\mathbb{E}\left[\mathbf{P}_{k'}\cdots\mathbf{P}_{t-1}\mathbf{P}_{t-1}\cdots\mathbf{P}_k\ldots\mathbf{P}_{k'}\right] = \cdots = \mathbb{E}\left[\mathbf{P}_{k'}\cdots\mathbf{P}_{k-1}\right]\mathbb{E}\left[\mathbf{P}^2\right]^{t-k+1} \\
&= \mathbb{E}_{\mathbf{P}_{k'},\ldots,\mathbf{P}_{k-2}}\left[\mathbb{E}_{\mathbf{P}_{k-1}}\left[\mathbf{P}_{k'}\cdots\mathbf{P}_{k-1}\middle|\mathbf{P}_{k'},\ldots,\mathbf{P}_{k-2}\right]\right]\mathbb{E}\left[\mathbf{P}^2\right]^{t-k+1} \\
&[3] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{P}_{k'}\cdots\mathbf{P}_{k-2}\mathbb{E}\left[\mathbf{P}_{k-1}\right]\right]\right]\mathbb{E}\left[\mathbf{P}^2\right]^{t-k+1} \\
&[4] = \mathbb{E}\left[\mathbf{P}\right]\mathbb{E}\left[\mathbf{P}_{k'}\cdots\mathbf{P}_{k-2}\right]\mathbb{E}\left[\mathbf{P}^2\right]^{t-k+1} = \ldots = \mathbb{E}\left[\mathbf{P}\right]^{k-k'}\mathbb{E}\left[\mathbf{P}^2\right]^{t-k+1}.
\end{aligned}$$

Note that when $k = k'$, it reduces to case [1].

[3] The data matrices are assumed to be independent of the past (see Section 2.1).

[4] $\mathbb{E}[\mathbf{P}], \mathbb{E}[\mathbf{P}^2]$ **are both multiples of I, and thus commutative.** By Eqs. (5) and (6), which give closed-form expressions for $\mathbb{E}[\mathbf{P}]$ and $\mathbb{E}[\mathbf{P}^2]$, i.e., both are scalar multiples of the identity matrix, then commutativity is automatically implied.

Term 1:

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{S}_{t:1}\left(\mathbf{w}_0-\mathbf{w}_1^\star\right)+\frac{1}{\lambda d}\sum_{k=1}^{t}\mathbf{S}_{t:k}\mathbf{X}_k^\top\mathbf{z}_k\right\|^2\\
&=\mathbb{E}\left(\left(\mathbf{w}_0-\mathbf{w}_1^\star\right)^\top\mathbf{S}_{t:1}^\top+\frac{1}{\lambda d}\sum_{k=1}^{t}\mathbf{z}_k^\top\mathbf{X}_k\mathbf{S}_{t:k}^\top\right)\left(\mathbf{S}_{t:1}(\mathbf{w}_0-\mathbf{w}_1^\star)+\frac{1}{\lambda d}\sum_{k=1}^{t}\mathbf{S}_{t:k}\mathbf{X}_k^\top\mathbf{z}_k\right)\\
[*]&=\mathbb{E}\left(\left(\mathbf{w}_0-\mathbf{w}_1^\star\right)^\top\mathbf{S}_{t:1}^\top\mathbf{S}_{t:1}\left(\mathbf{w}_0-\mathbf{w}_1^\star\right)+\frac{1}{(\lambda d)^2}\sum_{k=1}^{t}\left(\mathbf{z}_k\right)^\top\mathbf{X}_k\mathbf{S}_{t:k}^\top\sum_{k'=1}^{t}\mathbf{S}_{t:k'}\mathbf{X}_{k'}^\top\mathbf{z}_{k'}\right)\\
[**]&=\left(\mathbb{E}\left[\mathbf{P}^2\right]\right)^t\left\|\mathbf{w}_0-\mathbf{w}_1^\star\right\|^2+v_z d\mathbb{E}\left[\frac{1}{(\lambda d)^2}\mathbf{P}_k\left(\mathbf{X}_k\right)^\top\mathbf{X}_k\mathbf{P}_k\right]\frac{1-\left(\mathbb{E}\left[\mathbf{P}^2\right]\right)^t}{1-\mathbb{E}\left[\mathbf{P}^2\right]},
\end{aligned}
$$

where [*] follows by Section 2.1 as the noise variables are sampled independently across tasks with $\mathbb{E}\left[\mathbf{z}_t\right]=\mathbf{0}$, and [**] follows as the left inner term (in the third line) is,

$$
\left(\mathbf{w}_0-\mathbf{w}_1^\star\right)^\top\mathbb{E}\left[\mathbf{S}_{t:1}^\top\mathbf{S}_{t:1}\right]\left(\mathbf{w}_0-\mathbf{w}_1^\star\right)\overset{[1,4]}{=}\left(\mathbb{E}\left[\mathbf{P}^2\right]\right)^t\left\|\mathbf{w}_0-\mathbf{w}_1^\star\right\|^2,
$$

and the right inner term is,

$$
\begin{aligned}
&\frac{1}{(\lambda d)^2}\mathbb{E}\left[\sum_{k=1}^{t}\sum_{k'=1}^{t}\mathbf{z}_k^\top\mathbf{X}_k\mathbf{S}_{t:k}^\top\mathbf{S}_{t:k}\mathbf{X}_{k'}^\top\mathbf{z}_{k'}\right]=\frac{1}{(\lambda d)^2}\mathbb{E}\left[\sum_{k=1}^{t}\mathbf{z}_k^\top\mathbf{X}_k\mathbf{S}_{t:k}^\top\mathbf{S}_{t:k}\mathbf{X}_k^\top\mathbf{z}_k\right]\\
&=\frac{1}{(\lambda d)^2}\mathbb{E}\left[\sum_{k=1}^{t}\mathrm{Tr}\left[\mathbf{z}_k^\top\mathbf{X}_k\mathbf{S}_{t:k}^\top\mathbf{S}_{t:k}\mathbf{X}_k^\top\mathbf{z}_k\right]\right]=\frac{v_z}{(\lambda d)^2}\sum_{k=1}^{t}\mathbb{E}\mathrm{Tr}\left[\mathbf{X}_k^\top\mathbf{X}_k\mathbf{S}_{t:k}^\top\mathbf{S}_{t:k}\right]\\
&=\frac{v_z}{(\lambda d)^2}\sum_{k=1}^{t}\mathbb{E}\mathrm{Tr}\left[\mathbf{X}_k^\top\mathbf{X}_k\left(\mathbf{S}_{t:k+1}\mathbf{P}_k\right)^\top\left(\mathbf{S}_{t:k+1}\mathbf{P}_k\right)\right]\\
&=\frac{v_z}{(\lambda d)^2}\sum_{k=1}^{t}\mathbb{E}\mathrm{Tr}\left[\mathbf{P}_k\mathbf{X}_k^\top\mathbf{X}_k\mathbf{P}_k^\top\mathbf{S}_{t:k+1}^\top\mathbf{S}_{t:k+1}\right]\\
&=v_z\sum_{k=1}^{t}\mathrm{Tr}\left[\mathbb{E}\left[\frac{1}{(\lambda d)^2}\mathbf{P}_k\mathbf{X}_k^\top\mathbf{X}_k\mathbf{P}_k\right]\mathbb{E}\left[\mathbf{S}_{t:k+1}^\top\mathbf{S}_{t:k+1}\right]\right]\\
[1]&=v_z\sum_{k=1}^{t}\mathrm{Tr}\left[\mathbb{E}\left[\frac{1}{(\lambda d)^2}\mathbf{P}_k\mathbf{X}_k^\top\mathbf{X}_k\mathbf{P}_k\right]\left(\mathbb{E}\left[\mathbf{P}^2\right]\right)^{t-k}\right]\\
[4]&=v_z\sum_{k=1}^{t}d\mathbb{E}\left[\frac{1}{(\lambda d)^2}\mathbf{P}_k\mathbf{X}_k^\top\mathbf{X}_k\mathbf{P}_k\right]\left(\mathbb{E}\left[\mathbf{P}^2\right]\right)^{t-k}\\
&=v_z d\mathbb{E}\left[\frac{1}{(\lambda d)^2}\mathbf{P}_k\mathbf{X}_k^\top\mathbf{X}_k\mathbf{P}_k\right]\sum_{k=1}^{t}\left(\mathbb{E}\left[\mathbf{P}^2\right]\right)^{t-k}\\
&=v_z d\mathbb{E}\left[\frac{1}{(\lambda d)^2}\mathbf{P}_k\mathbf{X}_k^\top\mathbf{X}_k\mathbf{P}_k\right]\frac{1-\left(\mathbb{E}\left[\mathbf{P}^2\right]\right)^t}{1-\mathbb{E}\left[\mathbf{P}^2\right]},
\end{aligned}
$$

where [1], [4] are explained above.

Term 2:

$$
\begin{aligned}
&2\mathbb{E}\left(\sum_{k=2}^{t}\mathbf{S}_{t:k}\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)+\left(\mathbf{w}_{t}^{\star}-\mathbf{w}_{i}^{\star}\right)\right)^{\top}\left(\mathbf{S}_{t:1}\left(\mathbf{w}_{0}-\mathbf{w}_{1}^{\star}\right)+\frac{1}{\lambda}\sum_{k=1}^{t}\mathbf{S}_{t:k}\mathbf{X}_{k}^{\top}\mathbf{z}_{k}\right)\\
&=2\mathbb{E}\Bigg[\left(\sum_{k=2}^{t}\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)^{\top}\mathbf{S}_{t:k}^{\top}\right)\left(\mathbf{S}_{t:1}\left(\mathbf{w}_{0}-\mathbf{w}_{1}^{\star}\right)+\frac{1}{\lambda}\sum_{k=1}^{t}\mathbf{S}_{t:k}\mathbf{X}_{k}^{\top}\mathbf{z}_{k}\right)\\
&\qquad+\left(\mathbf{w}_{t}^{\star}-\mathbf{w}_{i}^{\star}\right)^{\top}\left(\mathbf{S}_{t:1}\left(\mathbf{w}_{0}-\mathbf{w}_{1}^{\star}\right)+\frac{1}{\lambda}\sum_{k=1}^{t}\mathbf{S}_{t:k}\mathbf{X}_{k}^{\top}\mathbf{z}_{k}\right)\Bigg]\\
&=2\mathbb{E}\Bigg[\sum_{k=2}^{t}\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)^{\top}\mathbf{S}_{t:k}^{\top}\mathbf{S}_{t:k}\mathbf{S}_{k-1:1}\left(\mathbf{w}_{0}-\mathbf{w}_{1}^{\star}\right)+\left(\mathbf{w}_{t}^{\star}-\mathbf{w}_{i}^{\star}\right)^{\top}\mathbf{S}_{t:1}\left(\mathbf{w}_{0}-\mathbf{w}_{1}^{\star}\right)\Bigg]\\
&[2]=2\Bigg(\sum_{k=2}^{t}\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)^{\top}\mathbb{E}\left[\mathbf{P}^{2}\right]^{t-k+1}\mathbb{E}\left[\mathbf{P}\right]^{k-1}\left(\mathbf{w}_{0}-\mathbf{w}_{1}^{\star}\right)\\
&\qquad+\left(\mathbf{w}_{t}^{\star}-\mathbf{w}_{i}^{\star}\right)^{\top}\mathbb{E}\left[\mathbf{P}\right]^{t}\left(\mathbf{w}_{0}-\mathbf{w}_{1}^{\star}\right)\Bigg).
\end{aligned}
$$

Term 3:

$$
\begin{aligned}
&\mathbb{E}\left\|\sum_{k=2}^{t}\mathbf{S}_{t:k}\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)+\left(\mathbf{w}_{t}^{\star}-\mathbf{w}_{i}^{\star}\right)\right\|^{2}\\
&=\mathbb{E}\Bigg[\left(\sum_{k=2}^{t}\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)^{\top}\mathbf{S}_{t:k}^{\top}\right)\left(\sum_{k=2}^{t}\mathbf{S}_{t:k}\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)\right)\\
&\qquad+2\left(\mathbf{w}_{t}^{\star}-\mathbf{w}_{i}^{\star}\right)^{\top}\left(\sum_{k=2}^{t}\mathbb{E}\left[\mathbf{S}_{t:k}\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)\right]\right)+\left\|\mathbf{w}_{t}^{\star}-\mathbf{w}_{i}^{\star}\right\|^{2}\Bigg]\\
&=\sum_{k=2}^{t}\sum_{k'=2}^{t}\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)^{\top}\mathbb{E}\left[\mathbf{S}_{t:k}^{\top}\mathbf{S}_{t:k'}\right]\left(\mathbf{w}_{k'-1}^{\star}-\mathbf{w}_{k'}^{\star}\right)\\
&\qquad+2\left(\mathbf{w}_{t}^{\star}-\mathbf{w}_{i}^{\star}\right)^{\top}\sum_{k=2}^{t}\mathbb{E}\left[\mathbf{S}_{t:k}\right]\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)+\left\|\mathbf{w}_{t}^{\star}-\mathbf{w}_{i}^{\star}\right\|^{2}\\
&=\sum_{k=2}^{t}\sum_{k'=2}^{t}\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)^{\top}\mathbb{E}\left[\mathbf{S}_{t:k\vee k'}^{\top}\mathbf{S}_{t:k\vee k'}\mathbf{S}_{((k\vee k')-1):k\wedge k'}\right]\left(\mathbf{w}_{k'-1}^{\star}-\mathbf{w}_{k'}^{\star}\right)\\
&\qquad+2\left(\mathbf{w}_{t}^{\star}-\mathbf{w}_{i}^{\star}\right)^{\top}\sum_{k=2}^{t}\mathbb{E}\left[\mathbf{S}_{t:k}\right]\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)+\left\|\mathbf{w}_{t}^{\star}-\mathbf{w}_{i}^{\star}\right\|^{2}\\
&[2]=\sum_{k=2}^{t}\sum_{k'=2}^{t}\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)^{\top}\mathbb{E}\left[\mathbf{P}^{2}\right]^{t-(k\vee k')+1}\mathbb{E}\left[\mathbf{P}\right]^{|k-k'|}\left(\mathbf{w}_{k'-1}^{\star}-\mathbf{w}_{k'}^{\star}\right)\\
&\qquad+2\left(\mathbf{w}_{t}^{\star}-\mathbf{w}_{i}^{\star}\right)^{\top}\sum_{k=2}^{t}\mathbb{E}\left[\mathbf{P}\right]^{t-k+1}\left(\mathbf{w}_{k-1}^{\star}-\mathbf{w}_{k}^{\star}\right)+\left\|\mathbf{w}_{t}^{\star}-\mathbf{w}_{i}^{\star}\right\|^{2},
\end{aligned}
$$

where [2][10] is explained above.

To finalize the result, we sum all the terms we obtained. First, we define

$$a = \left\|\mathbb{E}\left[\mathbf{P}^2\right]\right\|_2, \; b = \left\|\mathbb{E}\left[\mathbf{P}\right]\right\|_2, \; c = \frac{d}{(\lambda d)^2}\left\|\mathbb{E}\left[\mathbf{P}\mathbf{X}^\top\mathbf{X}\mathbf{P}\right]\right\|_2.$$

Then, we plug all the terms into Eq. (11), and obtain,

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{w}_t - \mathbf{w}_i^\star\right\|^2\right] &= a^t\left\|\mathbf{w}_0 - \mathbf{w}_1^\star\right\|^2 + v_z c\frac{1-a^t}{1-a} \\
&+ 2\sum_{k=2}^{t} a^{t-k+1}b^{k-1}\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top\left(\mathbf{w}_0 - \mathbf{w}_1^\star\right) + 2b^t\left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)^\top\left(\mathbf{w}_0 - \mathbf{w}_1^\star\right) \\
&+ \sum_{k=2}^{t}\sum_{k'=2}^{t} a^{t-\max(k,k')+1}b^{|k-k'|}\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top\left(\mathbf{w}_{k'-1}^\star - \mathbf{w}_{k'}^\star\right) \\
&+ 2\sum_{k=2}^{t} b^{t-k+1}\left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)^\top\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \left\|\mathbf{w}_t^\star - \mathbf{w}_i^\star\right\|^2.
\end{aligned} \tag{12}$$

Finally, the theorem follows by substituting the above expression into

$$\mathbb{E}\left[G_T\right] = \frac{1}{T}\sum_{i=1}^{T}\mathbb{E}\left[\left\|\mathbf{w}_T - \mathbf{w}_i^\star\right\|^2\right].$$

■

10. In general, for $k \neq k'$ one obtains a two-case identity: if $k > k'$ then $\mathbb{E}[\mathbf{S}_{t:k}^\top\mathbf{S}_{t:k'}] = \mathbb{E}[\mathbf{P}^2]^{\,t-k+1}\,\mathbb{E}[\mathbf{P}]^{\,k-k'}$, while if $k < k'$ then $\mathbb{E}[\mathbf{S}_{t:k}^\top\mathbf{S}_{t:k'}] = \mathbb{E}[\mathbf{P}]^{\,k'-k}\,\mathbb{E}[\mathbf{P}^2]^{\,t-k'+1}$, where the factor $\mathbb{E}[\mathbf{P}]^{|k-k'|}$ appears on the side of the longer product. In our setting $\mathbb{E}[\mathbf{P}] = b\mathbf{I}$ and $\mathbb{E}[\mathbf{P}^2] = a\mathbf{I}$ (hence they commute), so both cases collapse to $\mathbb{E}[\mathbf{P}^2]^{\,t-(k\vee k')+1}\mathbb{E}[\mathbf{P}]^{\,|k-k'|}$.

C.1. Proof of Example 1

Recall Example 1. Indeed, in the high-dimensional regime, Theorem 3 generalizes Theorem 4.1 in Lin et al. (2023). Specifically, after adjusting notation, their result becomes,

$$\mathbb{E}[G_T] = \tfrac{(1-\alpha)^T}{T}\sum_{i=1}^{T}\|\mathbf{w}_i^\star\|^2 + \tfrac{1}{T}\sum_{i=1}^{T}(1-\alpha)^{T-i}\,\alpha\sum_{k=1}^{T}\|\mathbf{w}_k^\star - \mathbf{w}_i^\star\|^2 + \tfrac{dv_z}{d-n-1}\left(1-(1-\alpha)^T\right),$$

aligning with Theorem 3 in the high-dimensional, unregularized case ($d, n \to \infty$, $\lambda \to 0$, $v_x = 1$).

Proof. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$. From Barata and Hussein (2012, Theorem 4.3), we know that $B(\lambda) \triangleq (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top \to \mathbf{X}^+$ as $\lambda \to 0$, where $\mathbf{X}^+$ is the Moore–Penrose pseudoinverse.

$$\lim_{\lambda\to 0^+} a = \lim_{\lambda\to 0^+}\frac{1}{2}\left(1-\alpha+\frac{\frac{\lambda}{v_x}(1+\alpha)+(1-\alpha)^2}{\sqrt{\left(\frac{\lambda}{v_x}\right)^2+2\frac{\lambda}{v_x}(1+\alpha)+(1-\alpha)^2}}\right) = 1-\alpha$$

$$\lim_{\lambda\to 0^+} b = \lim_{\lambda\to 0^+}\frac{1}{2}\left(1-\alpha-\frac{\lambda}{v_x}+\sqrt{\left(\frac{\lambda}{v_x}\right)^2+2\frac{\lambda}{v_x}(1+\alpha)+(1-\alpha)^2}\right) = 1-\alpha$$

Substituting the values to our result in Eq. (12), with $\mathbf{w}_0 = 0$,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}_i^\star\|^2\right] &= \underbrace{v_z\frac{1}{2v_x}\left(\frac{v_x(1+\alpha)+\lambda}{\sqrt{\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2}}-1\right)\frac{1-(1-\alpha)^t}{1-(1-\alpha)}}_{\text{term 1}} \\
&+ \underbrace{(1-\alpha)^t\|\mathbf{w}_1^\star\|^2 + 2(1-\alpha)^t\sum_{k=2}^{t}\left(\mathbf{w}_k^\star-\mathbf{w}_{k-1}^\star\right)^\top\mathbf{w}_1^\star + 2(1-\alpha)^t\left(\mathbf{w}_i^\star-\mathbf{w}_t^\star\right)^\top\mathbf{w}_1^\star}_{\text{term 2}} \\
&+ \underbrace{\sum_{k=2}^{t}\sum_{k'=2}^{t}(1-\alpha)^{t-\min(k,k')+1}\left(\mathbf{w}_{k-1}^\star-\mathbf{w}_k^\star\right)^\top\left(\mathbf{w}_{k'-1}^\star-\mathbf{w}_{k'}^\star\right)}_{\text{term 3}} \\
&\underbrace{+2\sum_{k=2}^{t}(1-\alpha)^{t-k+1}\left(\mathbf{w}_t^\star-\mathbf{w}_i^\star\right)^\top\left(\mathbf{w}_{k-1}^\star-\mathbf{w}_k^\star\right)+\|\mathbf{w}_t^\star-\mathbf{w}_i^\star\|^2}_{\text{term 3}}.
\end{aligned}$$

First term is:

$$\begin{aligned}
&\lim_{\lambda\to 0^+} v_z\frac{1}{2v_x}\left(\frac{v_x(1+\alpha)+\lambda}{\sqrt{\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2}}-1\right)\frac{1-(1-\alpha)^t}{1-(1-\alpha)} \\
&\overset{[1]}{=} v_z\frac{1}{2}\left(\frac{1+\alpha}{1-\alpha}-1\right)\frac{1-(1-\alpha)^t}{1-(1-\alpha)} = v_z\frac{\alpha}{1-\alpha}\frac{1-(1-\alpha)^t}{\alpha} = v_z\frac{1-(1-\alpha)^t}{1-\alpha},
\end{aligned}$$

where step [1] follows from setting $v_x = 1, \lambda \to 0$ according to the notation in Lin et al. (2023).

Second term is:

$$
\begin{aligned}
&(1-\alpha)^t \|\mathbf{w}_1^\star\|^2 + 2(1-\alpha)^t \sum_{k=2}^{t} \left(\mathbf{w}_k^\star - \mathbf{w}_{k-1}^\star\right)^\top \mathbf{w}_1^\star + 2(1-\alpha)^t \left(\mathbf{w}_i^\star - \mathbf{w}_t^\star\right)^\top \mathbf{w}_1^\star \\
&= (1-\alpha)^t \|\mathbf{w}_1^\star\|^2 + 2(1-\alpha)^t \left[(\cancel{\mathbf{w}_t^\star} - \mathbf{w}_1^\star)^\top \mathbf{w}_1^\star + \left((\mathbf{w}_i^\star)^\top \mathbf{w}_1^\star - \cancel{(\mathbf{w}_t^\star)^\top \mathbf{w}_1^\star} \right) \right] \\
&= (1-\alpha)^t \|\mathbf{w}_1^\star\|^2 + 2(1-\alpha)^t \left(-\|\mathbf{w}_1^\star\|^2 + (\mathbf{w}_i^\star)^\top \mathbf{w}_1^\star \right) \\
&= -(1-\alpha)^t \|\mathbf{w}_1^\star\|^2 + 2(1-\alpha)^t (\mathbf{w}_i^\star)^\top \mathbf{w}_1^\star - (1-\alpha)^t \|\mathbf{w}_i^\star\|^2 + (1-\alpha)^t \|\mathbf{w}_i^\star\|^2 \\
&= -(1-\alpha)^t \|\mathbf{w}_i^\star - \mathbf{w}_1^\star\|^2 + (1-\alpha)^t \|\mathbf{w}_i^\star\|^2 \ .
\end{aligned}
$$

Finally, the third term is:

$$
\begin{aligned}
&\sum_{k=2}^{t} \sum_{k'=2}^{t} (1-\alpha)^{t-\min(k,k')+1} \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top \left(\mathbf{w}_{k'-1}^\star - \mathbf{w}_{k'}^\star\right) \\
&\quad + 2 \sum_{k=2}^{t} (1-\alpha)^{t-k+1} \left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)^\top \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \|\mathbf{w}_t^\star - \mathbf{w}_i^\star\|^2 \\
&= \sum_{s=2}^{t} (1-\alpha)^{t-s+1} \left[2 \left(\mathbf{w}_{s-1}^\star - \mathbf{w}_s^\star\right)^\top \sum_{r=s}^{t} \left(\mathbf{w}_{r-1}^\star - \mathbf{w}_r^\star\right) - \left\|\mathbf{w}_{s-1}^\star - \mathbf{w}_s^\star\right\|^2 \right] \\
&\quad + 2 \sum_{k=2}^{t} (1-\alpha)^{t-k+1} \left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)^\top \left(\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_t^\star\right) - \left(\mathbf{w}_k^\star - \mathbf{w}_t^\star\right) \right) + \|\mathbf{w}_t^\star - \mathbf{w}_i^\star\|^2 \\
&= 2 \sum_{s=2}^{t} (1-\alpha)^{t-s+1} \left[\left\|\mathbf{w}_{s-1}^\star - \mathbf{w}_t^\star\right\|^2 - \left\|\mathbf{w}_{s-1}^\star - \mathbf{w}_t^\star\right\|^2 + 2 \left(\mathbf{w}_{s-1}^\star - \mathbf{w}_s^\star\right)^\top \left(\mathbf{w}_{s-1}^\star - \mathbf{w}_t^\star\right) \right. \\
&\qquad\qquad \left. - \left\|\mathbf{w}_{s-1}^\star - \mathbf{w}_s^\star\right\|^2 \right] + \|\mathbf{w}_t^\star - \mathbf{w}_i^\star\|^2 \\
&\quad + 2(1-\alpha)^{t-1} \left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)^\top \left(\mathbf{w}_1^\star - \mathbf{w}_t^\star\right) + 2\alpha \sum_{k=3}^{t} (1-\alpha)^{t-k+1} \left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)^\top \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_t^\star\right) \\
&= \left(\sum_{s=2}^{t} (1-\alpha)^{t-s+1} \left[\left\|\mathbf{w}_{s-1}^\star - \mathbf{w}_t^\star\right\|^2 - \|\mathbf{w}_t^\star - \mathbf{w}_s^\star\|^2 \right] \right) + \|\mathbf{w}_t^\star - \mathbf{w}_i^\star\|^2 \\
&\quad + 2(1-\alpha)^{t-1} \left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)^\top \left(\mathbf{w}_1^\star - \mathbf{w}_t^\star\right) + 2\alpha \sum_{k=3}^{t} (1-\alpha)^{t-k+1} \left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)^\top \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_t^\star\right) \\
&= (1-\alpha)^{t-1} \|\mathbf{w}_1^\star - \mathbf{w}_t^\star\|^2 + \alpha \sum_{k=3}^{t} (1-\alpha)^{t-k+1} \left\|\mathbf{w}_{k-1}^\star - \mathbf{w}_t^\star\right\|^2 + \|\mathbf{w}_t^\star - \mathbf{w}_i^\star\|^2 \\
&\quad + 2(1-\alpha)^{t-1} \left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)^\top \left(\mathbf{w}_1^\star - \mathbf{w}_t^\star\right) + 2\alpha \sum_{k=3}^{t} (1-\alpha)^{t-k+1} \left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)^\top \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_t^\star\right)
\end{aligned}
$$

$$
\begin{aligned}
&= (1-\alpha)^{t-1}\left(\|\mathbf{w}_1^\star - \mathbf{w}_t^\star\|^2 + 2\left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)^\top\left(\mathbf{w}_1^\star - \mathbf{w}_t^\star\right)\right) + \|\mathbf{w}_t^\star - \mathbf{w}_i^\star\|^2 \\
&\quad + \alpha\sum_{k=3}^{t}(1-\alpha)^{t-k+1}\left\|\mathbf{w}_{k-1}^\star - \mathbf{w}_i^\star\right\|^2 - \alpha\sum_{k=3}^{t}(1-\alpha)^{t-k+1}\|\mathbf{w}_t^\star - \mathbf{w}_i^\star\|^2 \\
&= (1-\alpha)^{t-1}\left(\|\mathbf{w}_1^\star - \mathbf{w}_t^\star\|^2 + 2\left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)^\top\left(\mathbf{w}_1^\star - \mathbf{w}_t^\star\right)\right) + \|\mathbf{w}_t^\star - \mathbf{w}_i^\star\|^2 \\
&\quad + \alpha\sum_{k=2}^{t-1}(1-\alpha)^{t-k}\|\mathbf{w}_k^\star - \mathbf{w}_i^\star\|^2 - (1-\alpha)\left(1-(1-\alpha)^{t-2}\right)\|\mathbf{w}_t^\star - \mathbf{w}_i^\star\|^2 \\
&= (1-\alpha)^{t-1}\left(\|\mathbf{w}_1^\star - \mathbf{w}_t^\star\|^2 + 2\left(\mathbf{w}_t^\star - \mathbf{w}_i^\star\right)^\top\left(\mathbf{w}_1^\star - \mathbf{w}_t^\star\right) + \|\mathbf{w}_t^\star - \mathbf{w}_i^\star\|^2\right) \\
&\quad + \alpha\sum_{k=2}^{t}(1-\alpha)^{t-k}\|\mathbf{w}_k^\star - \mathbf{w}_i^\star\|^2 \\
&= (1-\alpha)^{t}\|\mathbf{w}_i^\star - \mathbf{w}_1^\star\|^2 + \alpha\sum_{k=1}^{t}(1-\alpha)^{t-k}\|\mathbf{w}_k^\star - \mathbf{w}_i^\star\|^2 .
\end{aligned}
$$

Summing all the terms:

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}_i^\star\|^2\right] &= \sum_{k=1}^{t}(1-\alpha)^{t-k}\alpha\|\mathbf{w}_k^\star - \mathbf{w}_i^\star\|^2 + \cancel{(1-\alpha)^t\|\mathbf{w}_i^\star - \mathbf{w}_1^\star\|^2} \\
&\quad \cancel{-(1-\alpha)^t\|\mathbf{w}_i^\star - \mathbf{w}_1^\star\|^2} + (1-\alpha)^t\|\mathbf{w}_i^\star\| + \frac{v_z}{1-\alpha}\left(1-(1-\alpha)^t\right) \\
&= \sum_{k=1}^{t}(1-\alpha)^{t-k}\alpha\|\mathbf{w}_k^\star - \mathbf{w}_i^\star\|^2 + (1-\alpha)^t\|\mathbf{w}_i^\star\|^2 + \frac{v_z}{1-\alpha}\left(1-(1-\alpha)^t\right) .
\end{aligned}
$$

Overall,

$$
\begin{aligned}
&\mathbb{E}\left[G_T\right] = \frac{1}{T}\sum_{i=1}^{T}\mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}_i^\star\|^2\right] \\
&= \frac{1}{T}\sum_{i=1}^{T}(1-\alpha)^{T-i}\alpha\sum_{k=1}^{T}\|\mathbf{w}_k^\star - \mathbf{w}_i^\star\|^2 + \frac{1}{T}(1-\alpha)^T\sum_{i=1}^{T}\|\mathbf{w}_i^\star\|^2 + \frac{v_z}{1-\alpha}\left(1-(1-\alpha)^T\right) \\
&\overset{[1]}{=} \frac{(1-\alpha)^T}{T}\sum_{i=1}^{T}\|\mathbf{w}_i^\star\|^2 + \frac{1}{T}\sum_{i=1}^{T}(1-\alpha)^{T-i}\alpha\sum_{k=1}^{T}\|\mathbf{w}_k^\star - \mathbf{w}_i^\star\|^2 + \frac{dv_z}{d-n-1}\left(1-(1-\alpha)^T\right)
\end{aligned}
$$

Where [1] follows by our Setting 2.1 where $n, d \to \infty$. ■

Appendix D. Auxiliary Lemmas

We begin with a few auxiliary Lemmas that will be used in the proofs.

First, we introduce a few auxiliary variables.

$$\tilde{\lambda} \triangleq \frac{\lambda}{v_x}, \quad N \triangleq (1+\alpha)\,\tilde{\lambda} + (1-\alpha)^2\,,$$
$$D \triangleq \sqrt{\tilde{\lambda}^2 + 2\tilde{\lambda}\,(1+\alpha) + (1-\alpha)^2} = \sqrt{\left(1+\alpha+\tilde{\lambda}\right)^2 - 4\alpha}.$$

Lemma D.1 . *For all $\lambda > 0$ we have $b > a$.*

Proof. Recalling the notations from Table 2, we have that,

$$\begin{aligned} b - a &= \frac{1}{2}\left(1-\alpha-\tilde{\lambda}+D-\left(1-\alpha+\frac{\tilde{\lambda}\,(1+\alpha)+(1-\alpha)^2}{D}\right)\right) \\ &= \frac{-\tilde{\lambda}D + D^2 - \tilde{\lambda}\,(1+\alpha) - (1-\alpha)^2}{2D} = \frac{-\tilde{\lambda}D + \tilde{\lambda}^2 + \tilde{\lambda}\,(1+\alpha)}{2D} \\ &= \frac{\tilde{\lambda}\left(\left(\tilde{\lambda}+\alpha+1\right)-D\right)}{2D} = \frac{2\alpha\tilde{\lambda}}{D\left(\tilde{\lambda}+\alpha+1+D\right)} > 0. \end{aligned}$$

Note that $\tilde{\lambda} > 0 \iff \lambda > 0$, so $b > a$ for all $\lambda > 0$. ■

Lemma D.2 *For all $\lambda > 0$ we have $0 < a < 1, 0 < b < 1$.*

Proof. Recalling the notations from Table 2, we have that,

$$a = \frac{1}{2}\left(1-\alpha+\frac{\tilde{\lambda}(1+\alpha)+(1-\alpha)^2}{D}\right) \overset{1\geq\alpha}{>} 0, \tag{13}$$

and

$$\begin{aligned} b = \frac{1}{2}\left(1-\alpha-\tilde{\lambda}+D\right) &= \frac{1}{2}\left(1-\alpha-\tilde{\lambda}+\sqrt{\left(1+\alpha+\tilde{\lambda}\right)^2-4\alpha}\right) \\ &\overset{[\alpha>0]}{<} \frac{1}{2}\left(1-\alpha-\tilde{\lambda}+\sqrt{\left(1+\alpha+\tilde{\lambda}\right)^2}\right) = 1. \end{aligned}$$

Note that $\tilde{\lambda} > 0 \iff \lambda > 0$, using Lemma D.1 we have $0 < a < 1, 0 < b < 1$ for all $\lambda > 0$. ■

Lemma D.3 *The derivative of a with respect to λ takes the compact form $\frac{\mathrm{d}}{\mathrm{d}\lambda}a = \frac{2\tilde{\lambda}\alpha}{D^3 v_x}$ and is positive for all $\lambda > 0$.*

Proof. Recalling the notation from Table 2, we have by definition

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}a = \frac{1}{2v_x}\left(\frac{1+\alpha}{D} - \frac{N}{D^3}\left(\tilde{\lambda}+\alpha+1\right)\right) = \frac{(1+\alpha)\,D^2 - N\left(\tilde{\lambda}+\alpha+1\right)}{2v_x D^3},$$

we focus on the numerator:

$$\begin{aligned}
&(1+\alpha)\,D^2 - N\left(\tilde{\lambda}+\alpha+1\right)\\
&= (1+\alpha)\left(\tilde{\lambda}^2 + 2\tilde{\lambda}\,(1+\alpha) + (1-\alpha)^2\right) - \left((1+\alpha)\,\tilde{\lambda} + (1-\alpha)^2\right)\left(\tilde{\lambda}+(1+\alpha)\right)\\
&= \cancel{(1+\alpha)\,\tilde{\lambda}^2} + 2\tilde{\lambda}\,(1+\alpha)^2 + \cancel{(1+\alpha)\,(1-\alpha)^2}\\
&\quad - \left[\cancel{(1+\alpha)\,\tilde{\lambda}^2} + (1+\alpha)^2\,\tilde{\lambda} + (1-\alpha)^2\,\tilde{\lambda} + \cancel{(1+\alpha)\,(1-\alpha)^2}\right]\\
&= \left[(1+\alpha)^2 - (1-\alpha)^2\right]\tilde{\lambda} = 4\tilde{\lambda}\alpha.
\end{aligned} \tag{14}$$

Substituting back, we obtain $\frac{\mathrm{d}}{\mathrm{d}\lambda}a = \frac{2\tilde{\lambda}\alpha}{D^3 v_x}$, which is always positive as each factor in the product in positive. ■

Lemma D.4 *The derivative of b with respect to λ takes the compact form $\frac{\mathrm{d}}{\mathrm{d}\lambda}b = \frac{1}{2v_x}\left(\frac{-D+\tilde{\lambda}+\alpha+1}{D}\right)$ and is positive for all $\lambda > 0$.*

Proof. Recalling the notation from Table 2, we have by definition

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}b = \frac{1}{2v_x}\Big(-1 + \frac{\tilde{\lambda}+\alpha+1}{D}\Big),$$

since

$$\left(\tilde{\lambda}+\alpha+1\right)^2 = \tilde{\lambda}^2 + 2\tilde{\lambda}\,(1+\alpha) + (1+\alpha)^2 > \tilde{\lambda}^2 + 2\tilde{\lambda}\,(1+\alpha) + (1-\alpha)^2 = D^2, \tag{15}$$

then,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}b = \frac{1}{2v_x}\underbrace{\Big(-1 + \frac{\tilde{\lambda}+\alpha+1}{D}\Big)}_{>0} > 0.$$

■

Lemma D.5 $\tilde{\lambda}+\alpha+1+D > 2$ *for all* $\tilde{\lambda} > 0$.

Proof. Recalling the notations from Table 2, we have that,

$$\tilde{\lambda}+\alpha+1+D = \tilde{\lambda}+\alpha+1+\sqrt{\left(1-\alpha+\tilde{\lambda}\right)^2 + 4\tilde{\lambda}\alpha} \geq \tilde{\lambda}+\alpha+1+1-\alpha+\tilde{\lambda} = 2\tilde{\lambda}+2 > 2.$$

■

Lemma D.6 $1-b \geq \frac{\alpha}{1+\alpha+\tilde{\lambda}}$ *for all* $\tilde{\lambda} > 0$.

Proof. Recalling the notations from Table 2, we have that,

$$1-b = \frac{1+\alpha+\tilde{\lambda}-\sqrt{\left(1+\alpha+\tilde{\lambda}\right)^2-4\alpha}}{2} = \frac{\cancel{\left(1+\alpha+\tilde{\lambda}\right)^2}-\cancel{\left(1+\alpha+\tilde{\lambda}\right)^2}+4\alpha}{2\left(1+\alpha+\tilde{\lambda}+\sqrt{\left(1+\alpha+\tilde{\lambda}\right)^2-4\alpha}\right)}$$

$$= \frac{2\alpha}{\left(1+\alpha+\tilde{\lambda}+\sqrt{\left(1+\alpha+\tilde{\lambda}\right)^2-4\alpha}\right)} \geq \frac{\alpha}{1+\alpha+\tilde{\lambda}}.$$

■

Lemma D.7 $a, b \leq 1-\frac{\alpha}{1+\alpha+\tilde{\lambda}}$ *for all* $\tilde{\lambda} > 0$. *(a sharper bound on* a, b*)*

Proof. Lemmas D.1 and D.6 imply that $1-a \geq 1-b \geq \frac{\alpha}{1+\alpha+\tilde{\lambda}}$ so $a, b \leq 1-\frac{\alpha}{1+\alpha+\tilde{\lambda}}$. ■

Lemma D.8 *If* $\lim_{x\to 0^+} h(x) = a$ *for some* $a \in \mathbb{R}$ *and* $\lim_{x\to\infty} h(x) = b$ *for some* $b \in \mathbb{R}$*, and* h *is continuous for all* $x > 0$*, then* h *is bounded on* $(0, \infty)$.

Proof. Since $\lim_{x\to 0^+} h(x) = a$, there exists $\delta > 0$ such that

$$0 < x < \delta \;\Rightarrow\; |h(x) - a| < 1.$$

Hence for $0 < x < \delta$ we have $|h(x)| \leq |a| + 1 \triangleq M_1$.

Similarly, since $\lim_{x\to\infty} h(x) = b$, there exists $R > 0$ such that

$$x > R \;\Rightarrow\; |h(x) - b| < 1,$$

so for $x > R$ we have $|h(x)| \leq |b| + 1 =: M_3$.

On the closed, bounded interval $[\delta, R]$, the function h is continuous; by Weierstrass, h attains a maximum of its absolute value there. Thus there exists $M_2 > 0$ such that

$$|h(x)| \leq M_2 \quad \text{for all } x \in [\delta, R].$$

Combining the three regions, for all $x > 0$ we have

$$|h(x)| \leq M \triangleq \max\{M_1, M_2, M_3\},$$

so h is bounded on $(0, \infty)$. ■

Lemma D.9 *If $\alpha < 1$ then $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}a$, $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}b$ are bounded for all $\tilde{\lambda} \geq 0$.*

Proof. Recall the notations from Table 2. First, we show that,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}2a &= (1+\alpha)\,D - \frac{\left(\tilde{\lambda}+(1+\alpha)\right)\left(\tilde{\lambda}\,(1+\alpha)+(1-\alpha)^2\right)}{D^3} \\ &= \frac{(1+\alpha)\,D^2 - \left(\tilde{\lambda}+(1+\alpha)\right)\left(\tilde{\lambda}\,(1+\alpha)+(1-\alpha)^2\right)}{D^3} \\ &= \frac{\tilde{\lambda}\,(1+\alpha)^2 - \tilde{\lambda}\,(1-\alpha)^2}{D^3} = \frac{4\alpha\tilde{\lambda}}{D^3}\,. \end{aligned} \tag{16}$$

Furthermore,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}2b &= \frac{-D+\left(\tilde{\lambda}+\alpha+1\right)}{D} = \frac{-D^2+\left(\tilde{\lambda}+\alpha+1\right)^2}{D\left(\tilde{\lambda}+\alpha+1+D\right)} \\ &= \frac{-(1-\alpha)^2+(1+\alpha)^2}{D\left(\tilde{\lambda}+\alpha+1+D\right)} = \frac{4\alpha}{D\left(\tilde{\lambda}+\alpha+1+D\right)}\,. \end{aligned} \tag{17}$$

Finally, invoking Lemma D.5,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}a &= \frac{2\alpha\tilde{\lambda}}{D^3} \overset{D\geq 1-\alpha, D\geq\tilde{\lambda}}{\leq} \frac{2\alpha}{(1-\alpha)^2} \\ \frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}b &= \frac{2\alpha}{D\left(\tilde{\lambda}+\alpha+1+D\right)} \overset{D\geq 1-\alpha}{\leq} \frac{\alpha}{1-\alpha}\,. \end{aligned}$$

■

Lemma D.10 *If $\alpha = 1$ then $\lim_{\tilde{\lambda}\to 0^+} a = 0$, $\lim_{\tilde{\lambda}\to 0^+} b = 0$.*

Proof. By the notations from Table 2, if $\alpha = 1$ then,

$$a = \frac{1}{2}\left(\frac{2\tilde{\lambda}}{\sqrt{\tilde{\lambda}^2+4\tilde{\lambda}}}\right) = \frac{\cancel{\tilde{\lambda}}}{\cancel{\tilde{\lambda}}\sqrt{1+\frac{4}{\tilde{\lambda}}}} = \frac{1}{\sqrt{1+\frac{4}{\tilde{\lambda}}}}\,, \tag{18}$$

and

$$b = \frac{1}{2}\left(-\tilde{\lambda}+\sqrt{\tilde{\lambda}^2+4\tilde{\lambda}}\right)\,. \tag{19}$$

So, $\lim_{\tilde{\lambda}\to 0^+} a = 0$ and $\lim_{\tilde{\lambda}\to 0^+} b = 0$. ■

Lemma D.11 *If $\alpha = 1$ then $\forall n_1 \geq 2, n_2 \geq 3$, we have $a^{n_1}\frac{\mathrm{d}a}{\mathrm{d}\tilde{\lambda}}, b^{n_2}\frac{\mathrm{d}b}{\mathrm{d}\tilde{\lambda}}$ are bounded for all $\tilde{\lambda} > 0$.*

Proof. By Lemma D.2, it is sufficient to show this is true for $n_1 = 2, n_2 = 3$.

We use the closed-from expression for $\frac{\mathrm{d}a}{\mathrm{d}\tilde{\lambda}}, \frac{\mathrm{d}b}{\mathrm{d}\tilde{\lambda}}$ given in Lemmas D.3 and D.4.

Using Eqs. (16) to (19),

$$\lim_{\tilde{\lambda}\to 0^+} a^2 \frac{\mathrm{d}a}{\mathrm{d}\tilde{\lambda}} = \lim_{\tilde{\lambda}\to 0^+} \frac{2}{\left(1+\frac{4}{\tilde{\lambda}}\right)} \frac{\tilde{\lambda}}{\left(\tilde{\lambda}^2+4\tilde{\lambda}\right)^{3/2}} = 0$$

$$\lim_{\tilde{\lambda}\to \infty} a^2 \frac{\mathrm{d}a}{\mathrm{d}\tilde{\lambda}} = \lim_{\tilde{\lambda}\to \infty} a^2 \frac{\tilde{\lambda}}{2\left(\tilde{\lambda}^2+4\tilde{\lambda}\right)^{3/2}} \overset{[1]}{=} 0\,,$$

where [1] follows since a by Lemma D.2 is bounded for all $\tilde{\lambda} > 0$. Furthermore,

$$\lim_{\tilde{\lambda}\to 0^+} b^3 \frac{\mathrm{d}b}{\mathrm{d}\tilde{\lambda}} = \lim_{\tilde{\lambda}\to 0^+} \frac{\left(-\tilde{\lambda}+\sqrt{\tilde{\lambda}^2+4\tilde{\lambda}}\right)^3}{\sqrt{\tilde{\lambda}^2+4\tilde{\lambda}}\left(\tilde{\lambda}+2+\sqrt{\tilde{\lambda}^2+4\tilde{\lambda}}\right)} = 0$$

$$\lim_{\tilde{\lambda}\to \infty} b^3 \frac{\mathrm{d}b}{\mathrm{d}\tilde{\lambda}} = \lim_{\tilde{\lambda}\to \infty} b^3 \frac{2}{\sqrt{\tilde{\lambda}^2+4\tilde{\lambda}}\left(\tilde{\lambda}+2+\sqrt{\tilde{\lambda}^2+4\tilde{\lambda}}\right)} \overset{[2]}{=} 0,$$

where [2] follows since b by Lemma D.2 is bounded for all $\tilde{\lambda} > 0$.

Both of terms are bounded by Lemma D.8 as each one of the terms is continuous for all $\tilde{\lambda} > 0$. ■

For the next few Lemmas we define $f \triangleq v_z \frac{1}{2v_x}\left(\frac{(1+\alpha)+\tilde{\lambda}}{\sqrt{\tilde{\lambda}^2+2\tilde{\lambda}(1+\alpha)+(1-\alpha)^2}} - 1\right) = v_z c$.

Lemma D.12 *$f > 0$ for all $\tilde{\lambda} > 0$.*

Proof. It is sufficient to show that $(1+\alpha+\tilde{\lambda})^2 > \tilde{\lambda}^2 + 2\tilde{\lambda}(1+\alpha) + (1-\alpha)^2$ which is true if and only if $(1+\alpha)^2 > (1-\alpha)^2 \iff 4\alpha > 0 \iff \alpha > 0$. ■

Lemma D.13 *If $\alpha = 1$ then for all $n_1 \geq 4, n_2 \geq 2$ we have $a^{n_1}\frac{\mathrm{d}f}{\mathrm{d}\tilde{\lambda}}, a^{n_2} f$ is bounded for all $\tilde{\lambda} > 0$.*

Proof. By Lemma D.2 it is sufficient to show this is true for $n_1 = 4, n_2 = 2$.

Note that using Eq. (18), we get,

$$a^4 \frac{\mathrm{d}f}{\mathrm{d}\tilde{\lambda}} = \left(\frac{\tilde{\lambda}}{\tilde{\lambda}+4}\right)^2 \left(-\frac{2v_z}{v_x\left(\tilde{\lambda}^2+4\tilde{\lambda}\right)^{3/2}}\right) = -\frac{2v_z}{v_x}\frac{\sqrt{\tilde{\lambda}}}{(\tilde{\lambda}+4)^{7/2}}. \tag{20}$$

Furthermore,

$$a^2 f = v_x \frac{1}{2v_z}\left(\frac{\tilde{\lambda}+2}{\tilde{\lambda}\sqrt{1+\frac{4}{\tilde{\lambda}}}} - 1\right)\left(\frac{\tilde{\lambda}}{\tilde{\lambda}+4}\right). \tag{21}$$

By Eqs. (20) and (21), we have $a^4\frac{\mathrm{d}f}{\mathrm{d}\tilde{\lambda}} \to 0$, $a^2 f \to 0$ as $\tilde{\lambda} \to 0^+$, but note that also $a^4\frac{\mathrm{d}f}{\mathrm{d}\tilde{\lambda}} \to 0$, $a^2 f \to 0$ as $\tilde{\lambda} \to \infty$ by the same equations. Since both terms are continuous for all $\tilde{\lambda} > 0$ they are bounded in this region by Lemma D.8. ■

Lemma D.14 *If $\alpha = 1$ then $\frac{1-b}{1-a} = \frac{1}{1+a}$.*

Proof. Recall the notations from Table 2,

$$a = \frac{1}{2}\left(1 - \alpha + \frac{(1+\alpha)\tilde{\lambda} + (1-\alpha)^2}{D}\right) = \frac{\tilde{\lambda}}{D}, \qquad b = \frac{1}{2}\left(1 - \alpha - \tilde{\lambda} + D\right) = \frac{D - \tilde{\lambda}}{2}.$$

Thus

$$\frac{1-b}{1-a} = \frac{1 - \frac{D-\tilde{\lambda}}{2}}{1 - \frac{\tilde{\lambda}}{D}} = \frac{D(2 - D + \tilde{\lambda})}{2(D - \tilde{\lambda})} = \frac{D}{D + \tilde{\lambda}} = \frac{1}{1 + \frac{\tilde{\lambda}}{D}} = \frac{1}{1+a},$$

where the penultimate equality uses $(2 - D + \tilde{\lambda})(D + \tilde{\lambda}) = 2(D - \tilde{\lambda}) \iff D^2 = \tilde{\lambda}^2 + 4\tilde{\lambda}$, true by the definition of D. ■

Lemma D.15 *If $\alpha = 1$ then $\frac{1-b}{1-a}, \frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\frac{1-b}{1-a}$ are bounded for all $\tilde{\lambda} > 0$.*

Proof. Since $a \to 0$, $b \to 0$ as $\tilde{\lambda} \to 0^+$ by Lemma D.10 then $\lim_{\tilde{\lambda}\to 0^+} \frac{1-b}{1-a} = 1$, furthermore, we know from Lemma D.19 that $\lim_{\tilde{\lambda}\to\infty} \frac{1-b}{1-a} = \frac{1}{2}$, invoking Lemma D.8 $\frac{1-b}{1-a}$ is bounded for all $\tilde{\lambda} > 0$.

When $\alpha = 1$ we obtain the following equality $\frac{1-b}{1-a} = \frac{1}{1+a}$ by Lemma D.14, then by Lemma D.2 this quantity is bounded. ■

Lemma D.16 *Let h be a function analytic at ∞, where we define $h(x) = g\left(\frac{1}{x}\right)$, consider the Taylor series around 0 of $g(x) = P_k(x) + R(x)$ where $R(x) = O\left(x^k\right)$ and $P_k(x)$ is a polynomial of degree at most k then $\frac{\mathrm{d}}{\mathrm{d}x}h(x) = -\frac{\mathrm{d}}{\mathrm{d}x}P_k\left(\frac{1}{x}\right)\frac{1}{x^2} + O\left(\frac{1}{x^{k+2}}\right)$.*

Proof. From the assumptions g is analytic around 0 then for some $\rho > 0$,

$$h\left(\frac{1}{x}\right) = g(x) = \sum_{n=0}^{k} \frac{g^{(n)}(0)}{n!}x^n + \sum_{n=k+1}^{\infty} \frac{g^{(n)}(0)}{n!}x^n \quad \forall |x| < \rho$$

A power series can be differentiated term by term within its radius of convergence, as the series of derivatives converges uniformly on any closed interval $|x| \le r < \rho$, differentiating,

$$\frac{\mathrm{d}}{\mathrm{d}x}g(x) = \frac{\mathrm{d}}{\mathrm{d}x}\sum_{n=0}^{k} \frac{g^{(n)}(0)}{n!}x^n + \sum_{n=k+1}^{\infty} \frac{g^{(n)}(0)}{(n-1)!}x^{n-1}, \quad \forall |x| \le r.$$

For any $|x| < r$ write $|x| = r\theta$ where $\theta \in [0,1]$

$$\left|\frac{\mathrm{d}}{\mathrm{d}x}g(x) - \frac{\mathrm{d}}{\mathrm{d}x}\sum_{n=0}^{k} \frac{g^{(n)}(0)}{n!}x^n\right| = \sum_{n=k+1}^{\infty} \frac{g^{(n)}(0)}{(n-1)!}|x|^{n-1} \overset{r\theta=|x|}{=} \sum_{n=k+1}^{\infty} \frac{g^{(n)}(0)}{(n-1)!}(r\theta)^{n-1}$$

$$\le \theta^k \underbrace{\sum_{n=k+1}^{\infty} \frac{g^{(n)}(0)}{(n-1)!}r^{n-1}}_{\text{convergent sum}} = C\theta^k = \frac{C}{r^k}|x|^k.$$

Overall,

$$\frac{\mathrm{d}}{\mathrm{d}x} g(x) = \frac{\mathrm{d}}{\mathrm{d}x} \sum_{n=0}^{k} \frac{g^{(n)}(0)}{n!} x^n + O\left(x^k\right) \quad \forall |x| \le r$$

So,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x} h(x) &= g'\left(\frac{1}{x}\right) \cdot \left(-\frac{1}{x^2}\right) = \left[P_k'\left(\frac{1}{x}\right) + O\left(\frac{1}{x^k}\right)\right] \cdot \left(-\frac{1}{x^2}\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}x} P_k\left(\frac{1}{x}\right) + O\left(\frac{1}{x^{k+2}}\right) \quad \forall |x| \ge \frac{1}{r}\end{aligned}$$

■

Lemma D.17 *Let f be a function analytic at ∞, so that as $x \to \infty$ it admits the expansion of the form $f(x) = C + O(x^{-n})$ for some constant C and $n \in \mathbb{N}$, if in addition f is continuous for all $x \ge \delta$ for some $\delta \in \mathbb{R}$ then f is bounded for all $x \ge \delta$.*

Proof. There exists some $R > 0, M > 0$ which for all $x \in [R, \infty)$ we have $|f(x) - C| \le MR^{-n}$. On the compact interval $x \in [\delta, R]$ f is continuous thus bounded, denote $\sup_{x\in[\delta,R]} |f(x)| = B < \infty$. Therefore $|f(x)| \le \max\{B,\ |C| + MR^{-n}\}$ for all $x \ge \delta$, proving boundedness. ■

Lemma D.18 *We derive the following Laurent expansion for $\tilde{\lambda} \to \infty$.*

Recall the notations from Table 2, using the standard Taylor expansion around $u = 0$.

$$\begin{aligned}
a &= \frac{1}{2}\left(1 - \alpha + \frac{(1+\alpha) + \frac{(1-\alpha)^2}{\tilde{\lambda}}}{\sqrt{1 + \frac{2(1+\alpha)}{\tilde{\lambda}} + \frac{(1-\alpha)^2}{\tilde{\lambda}^2}}}\right) = \\
\left[\substack{(1+u)^{-1/2} \\ =1-\frac{1}{2}u+O\left(u^2\right)}\right] &= \frac{1}{2}\left(1 - \alpha + \left((1+\alpha) + \frac{(1-\alpha)^2}{\tilde{\lambda}}\right)\left(1 - \frac{1+\alpha}{\tilde{\lambda}} + O\left(\tilde{\lambda}^{-2}\right)\right)\right) \\
&= \frac{1}{2}\left(2 - \frac{4\alpha}{\tilde{\lambda}} + O\left(\tilde{\lambda}^{-2}\right)\right) = 1 - \frac{2\alpha}{\tilde{\lambda}} + O\left(\tilde{\lambda}^{-2}\right), \\
a^t &= \exp(t \ln(a)) = \exp\left(-\frac{2\alpha t}{\tilde{\lambda}} + O\left(t\tilde{\lambda}^{-2}\right)\right) = 1 - \frac{2\alpha t}{\tilde{\lambda}} + O\left(t^2\tilde{\lambda}^{-2}\right), \\
b &= \frac{1}{2}\left(1 - \alpha - \tilde{\lambda} + \tilde{\lambda}\sqrt{1 + \frac{2(1+\alpha)}{\tilde{\lambda}} + \frac{(1-\alpha)^2}{\tilde{\lambda}^2}}\right), \\
\left[\substack{(1+u)^{1/2} \\ =1+\frac{1}{2}u-\frac{1}{8}u^2+O\left(u^3\right)}\right] &= \frac{1}{2}\left(1 - \alpha - \tilde{\lambda} + \tilde{\lambda}\left(1 + \frac{(1+\alpha)}{\tilde{\lambda}} - \frac{2\alpha}{\tilde{\lambda}^2}\right) + O\left(\tilde{\lambda}^{-2}\right)\right) \\
&= \frac{1}{2}\left(2 - \frac{2\alpha}{\tilde{\lambda}} + O\left(\tilde{\lambda}^{-2}\right)\right) = 1 - \frac{\alpha}{\tilde{\lambda}} + O\left(\tilde{\lambda}^{-2}\right) \\
b^t &= \exp(t \ln(b)) = \exp\left(-\frac{\alpha t}{\tilde{\lambda}} + O\left(t\tilde{\lambda}^{-2}\right)\right) = 1 - \frac{\alpha t}{\tilde{\lambda}} + O\left(t^2\tilde{\lambda}^{-2}\right).
\end{aligned}$$

Lemma D.19 *The Laurent expansion of* $\frac{1-b}{1-a} = \frac{1}{2} + O\left(\tilde{\lambda}^{-1}\right)$.

Proof. Using Lemma D.18, we have

$$\frac{1-b}{1-a} = \frac{\frac{\alpha}{\tilde{\lambda}} + O\left(\tilde{\lambda}^{-2}\right)}{\frac{2\alpha}{\tilde{\lambda}} + O\left(\tilde{\lambda}^{-2}\right)} = \frac{1}{2} + O\left(\tilde{\lambda}^{-1}\right) .$$

■

Lemma D.20 *The Laurent expansion of* $f = O\left(\tilde{\lambda}^{-2}\right)$.

Proof. Recall the notations from Table 2, namely, $f \triangleq \frac{v_z}{2v_x}\left(\frac{1+\alpha+\tilde{\lambda}}{D} - 1\right)$, then,

$$\frac{2v_x}{v_z} f = \frac{\tilde{\lambda}+\alpha+1}{\sqrt{\left(\tilde{\lambda}+\alpha+1\right)^2 - 4\alpha}} - 1 = \frac{1}{\sqrt{1 - \frac{4\alpha}{\left(\tilde{\lambda}+\alpha+1\right)^2}}} - 1$$

$$\left[(1-u)^{-\frac{1}{2}} = 1 + \frac{1}{2}u + O\left(u^2\right)\right] = 1 + \frac{2\alpha}{\left(\tilde{\lambda}+\alpha+1\right)^2} + O\left(\tilde{\lambda}^{-4}\right) - 1 = O\left(\tilde{\lambda}^{-2}\right) ,$$

so $f = \frac{v_z}{2v_x} O\left(\tilde{\lambda}^{-2}\right) = O\left(\tilde{\lambda}^{-2}\right)$. ■

Lemma D.21 *The Laurent expansion of* $D = \tilde{\lambda} + \alpha + 1 - \frac{2\alpha}{\tilde{\lambda}+\alpha+1} + O\left(\tilde{\lambda}^{-3}\right)$.

Proof. Recall the notations from Table 2, namely, $D \triangleq \sqrt{\left(\tilde{\lambda}+\alpha+1\right)^2 - 4\alpha}$, then,

$$D = \sqrt{\left(\tilde{\lambda}+\alpha+1\right)^2 - 4\alpha} = \left(\tilde{\lambda}+\alpha+1\right)\sqrt{1 - \frac{4\alpha}{\left(\tilde{\lambda}+\alpha+1\right)^2}}$$

$$\left[\sqrt{1-u} = 1 - \frac{1}{2}u + O\left(u^2\right)\right] = \left(\tilde{\lambda}+\alpha+1\right)\left(1 - \frac{2\alpha}{\left(\tilde{\lambda}+\alpha+1\right)^2} + O\left(\tilde{\lambda}^{-4}\right)\right)$$

$$= \tilde{\lambda} + \alpha + 1 - \frac{2\alpha}{\tilde{\lambda}+\alpha+1} + O\left(\tilde{\lambda}^{-3}\right) .$$

■

Appendix E. Proofs for Section 3.1 (Special Case: Multiple i.i.d. Teachers)

E.1. Proof of Corollary 4: Generalization under i.i.d. Teachers

In this subsection, we derive a closed-form expression for $\mathbb{E}\big[\|\mathbf{w}_T - \mathbf{w}_i^\star\|^2\big]$ for any $1 \le i \le T$, with $T > 1$, under the i.i.d. teacher model introduced in Assumption 1. This result serves as the key intermediate step toward obtaining a closed-form expression for the average generalization error. Under Assumption 1, each task-specific teacher can be written as

$$\mathbf{w}_i^\star = \mathbf{w}^\star + \boldsymbol{\xi}_i, \qquad i \in [T],$$

where the random vectors $\boldsymbol{\xi}_i$ are i.i.d. with mean $\mathbf{0}$ and covariance $\mathrm{Cov}(\boldsymbol{\xi}_i) = \boldsymbol{\Sigma}$. Substituting this representation into the general expression derived in Eq. (12), and using the initialization $\mathbf{w}_0 = \mathbf{0}$, yields the following explicit decomposition of the squared error,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{w}_T - \mathbf{w}_i^\star\|^2\right] &= a^T \|\mathbf{w}^\star + \boldsymbol{\xi}_1\|^2 + 2b^T \left(\boldsymbol{\xi}_i - \boldsymbol{\xi}_T\right)^\top \left(\mathbf{w}^\star + \boldsymbol{\xi}_1\right) \\
&\quad + v_z \frac{1}{2v_x}\left(\frac{v_x\left(1+\alpha\right)+\lambda}{\sqrt{\lambda^2 + 2\lambda v_x(1+\alpha) + v_x^2\left(1-\alpha\right)^2}} - 1\right)\frac{1-a^T}{1-a} \\
&\quad + 2\sum_{k=2}^{T} a^{T-k+1} b^{k-1} \left(\boldsymbol{\xi}_k - \boldsymbol{\xi}_{k-1}\right)^\top \left(\mathbf{w}^\star + \boldsymbol{\xi}_1\right) \\
&\quad + \sum_{k=2}^{T}\sum_{k'=2}^{T} a^{T-\max(k,k')+1} b^{\max(k,k')-\min(k,k')} \left(\boldsymbol{\xi}_{k-1} - \boldsymbol{\xi}_k\right)^\top \left(\boldsymbol{\xi}_{k'-1} - \boldsymbol{\xi}_{k'}\right) \\
&\quad + 2\sum_{k=2}^{T} b^{T-k+1} \left(\boldsymbol{\xi}_T - \boldsymbol{\xi}_i\right)^\top \left(\boldsymbol{\xi}_{k-1} - \boldsymbol{\xi}_k\right) + \|\boldsymbol{\xi}_T - \boldsymbol{\xi}_i\|^2 .
\end{aligned}$$

Additionally, for all $i \in [T]$ we have,

$$\mathbb{E}\left[\boldsymbol{\xi}_i^\top \boldsymbol{\xi}_i\right] = \mathbb{E}\sum_{j=1}^{d} \xi_{i,j}^2 = \sum_{j=1}^{d} \mathbb{E}\xi_{i,j}^2 = \sum_{j=1}^{d} \mathrm{Var}\left[\xi_{i,j}\right] = \mathrm{Tr}\left(\boldsymbol{\Sigma}\right) . \tag{22}$$

where the equality follows from independence across coordinates and the definition of the covariance matrix. Using the identity in Eq. (22), and taking expectations with respect to the random variables $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_T$, we evaluate the resulting expression by separating the analysis into three cases, depending on the relative position of the index i.

In the special case, when $i = T$:

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{w}_T - \mathbf{w}_i^\star\|^2\right] \\
&= \left(a^T - 2a^{T-1}b\right)\mathrm{Tr}\left(\boldsymbol{\Sigma}\right) + a^T\|\mathbf{w}^\star\|^2 + cv_z\frac{1-a^T}{1-a} + 2\mathrm{Tr}\left(\boldsymbol{\Sigma}\right)\left(\sum_{k=2}^{T} a^{T-k+1} - b\sum_{k=3}^{T} a^{T-k+1}\right) \\
&= \left(a^T - 2a^{T-1}b\right)\mathrm{Tr}\left(\boldsymbol{\Sigma}\right) + a^T\|\mathbf{w}^\star\|^2 \\
&\qquad + cv_z\frac{1-a^T}{1-a} + 2\mathrm{Tr}\left(\boldsymbol{\Sigma}\right)\left(\frac{a\left(1-a^{T-1}\right)}{1-a} - b\cdot\frac{a\left(1-a^{T-2}\right)}{1-a}\right) .
\end{aligned} \tag{23}$$

When $1 < i < T$:

$$\mathbb{E}\left[\|\mathbf{w}_T - \mathbf{w}_i^\star\|^2\right] = \left(a^T - 2a^{T-1}b\right)\mathrm{Tr}\left(\mathbf{\Sigma}\right) + a^T\left\|\mathbf{w}^\star\right\|^2 + cv_z\frac{1-a^T}{1-a}$$
$$+ 2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{a\left(1-a^{T-1}\right)}{1-a} - b\cdot\frac{a\left(1-a^{T-2}\right)}{1-a} + 1\right) + 2\left(b^{T-i+1} - b^{T-i} - b\right)\mathrm{Tr}\left(\mathbf{\Sigma}\right). \quad (24)$$

When $i = 1$:

$$\mathbb{E}\left[\|\mathbf{w}_T - \mathbf{w}_i^\star\|^2\right] = \left(a^T - 2a^{T-1}b\right)\mathrm{Tr}\left(\mathbf{\Sigma}\right) + a^T\|\mathbf{w}^\star\|^2 + 2b^T\mathrm{Tr}\left(\mathbf{\Sigma}\right) + cv_z\frac{1-a^T}{1-a}$$
$$+ 2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{a\left(1-a^{T-1}\right)}{1-a} - b\cdot\frac{a\left(1-a^{T-2}\right)}{1-a} + 1\right) + 2\left(b^T - b^{T-1} - b\right)\mathrm{Tr}\left(\mathbf{\Sigma}\right). \quad (25)$$

Recall Corollary 4. Under the i.i.d. teacher setting (Assumption 1), and for all $T > 1$, the expected generalization loss of Scheme 1 becomes,

$$\mathbb{E}[G_T] = \underbrace{v_z c\frac{1-a^T}{1-a}}_{\text{Label noise}} + \underbrace{a^T\left\|\mathbf{w}^\star\right\|^2}_{\text{Teacher scale}} + \underbrace{2\mathrm{Tr}(\mathbf{\Sigma})}_{\substack{\text{Teacher}\\ \text{variance}}}\left(\frac{1-b}{1-a} + \frac{2b^T-1}{T} + \frac{a^T\left(-a+2b-1\right)}{2\left(1-a\right)}\right).$$

Proof. Using Eqs. (23) to (25) when summing over all cases,

$$\mathbb{E}\left[G_T\right] = \frac{1}{T}\sum_{i=1}^{T}\mathbb{E}\left[\|\mathbf{w}_T - \mathbf{w}_i^\star\|^2\right]$$
$$= \left(a^T - 2a^{T-1}b\right)\mathrm{Tr}\left(\mathbf{\Sigma}\right) + a^T\left\|\mathbf{w}^\star\right\|^2 + 2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{a\left(1-a^{T-1}\right)}{1-a} - b\cdot\frac{a\left(1-a^{T-2}\right)}{1-a}\right) +$$
$$cv_z\frac{1-a^T}{1-a} + \frac{2}{T}\left(2b^T + T - 1 - Tb\right)\mathrm{Tr}\left(\mathbf{\Sigma}\right) \quad (26)$$
$$= cv_z\frac{1-a^T}{1-a} + a^T\left\|\mathbf{w}^\star\right\|^2 +$$
$$2\mathrm{Tr}(\mathbf{\Sigma})\left(1 - b + \frac{2b^T-1}{T} + \frac{a}{1-a}\left(1 - a^{T-1} - b\left(1-a^{T-2}\right)\right) + \frac{a^{T-1}\left(a-2b\right)}{2}\right)$$
$$= cv_z\frac{1-a^T}{1-a} + a^T\left\|\mathbf{w}^\star\right\|^2 +$$
$$2\mathrm{Tr}(\mathbf{\Sigma})\left(1 - b + \frac{2b^T-1}{T} + \frac{a}{1-a}\left(1 - b + a^{T-2}\left(b-a\right)\right) + \frac{a^{T-1}\left(a-2b\right)}{2}\right)$$
$$= cv_z\frac{1-a^T}{1-a} + a^T\left\|\mathbf{w}^\star\right\|^2 +$$
$$2\mathrm{Tr}(\mathbf{\Sigma})\left(\frac{(1-b)(1-a)}{1-a} + \frac{a(1-b)}{1-a} + \frac{2b^T-1}{T} + \frac{a^{T-1}}{1-a}\left(b-a\right) + \frac{a^{T-1}\left(a-2b\right)}{2}\right)$$
$$= cv_z\frac{1-a^T}{1-a} + a^T\left\|\mathbf{w}^\star\right\|^2 + 2\mathrm{Tr}(\mathbf{\Sigma})\left(\frac{1-b}{1-a} + \frac{2b^T-1}{T} + \frac{a^T\left(-a+2b-1\right)}{2\left(1-a\right)}\right).$$

■

E.2. Generalization with an Infinite Task Horizon

We simplify Corollary 4 further, by considering the limit $T \to \infty$.

Corollary 1 *Under i.i.d. teachers (Assumption 1) we have,*

$$\lim_{T\to\infty} \mathbb{E}\left[G_T\right] = 2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{1-b}{1-a}\right) + \frac{v_z}{2v_x}\left(\frac{v_x\left(1+\alpha\right)+\lambda}{\sqrt{\lambda^2+2\lambda v_x\left(1+\alpha\right)+v_x^2\left(1-\alpha\right)^2}}-1\right)\frac{1}{1-a}\,.$$

Proof. This is readily proved by taking $T \to \infty$ in Corollary 4,

$$\begin{aligned}
&\lim_{T\to\infty} \mathbb{E}\left[G_T\right] \\
&= \lim_{T\to\infty}\left(v_z c\frac{1-a^T}{1-a} + a^T\left\|\mathbf{w}^\star\right\|^2 + 2\mathrm{Tr}(\mathbf{\Sigma})\left(\frac{1-b}{1-a}+\frac{2b^T-1}{T}+\frac{a^T\left(-a+2b-1\right)}{2(1-a)}\right)\right) \\
&= 2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{1-b}{1-a}\right) + \frac{v_z}{2v_x}\left(\frac{v_x\left(1+\alpha\right)+\lambda}{\sqrt{\lambda^2+2\lambda v_x\left(1+\alpha\right)+v_x^2\left(1-\alpha\right)^2}}-1\right)\frac{1}{1-a}\,.
\end{aligned}$$

We note in passing that, for any *fixed* $i > 1$, the same limit holds for $\mathbb{E}\left[\left\|\mathbf{w}_T - \mathbf{w}_i^\star\right\|^2\right]$ by taking $T \to \infty$ in Eq. (24). That is,

$$\begin{aligned}
&\lim_{T\to\infty} \mathbb{E}\left[\left\|\mathbf{w}_T - \mathbf{w}_i^\star\right\|^2\right] \\
&\quad = 2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{1-b}{1-a}\right) + \frac{v_z}{2v_x}\left(\frac{v_x\left(1+\alpha\right)+\lambda}{\sqrt{\lambda^2+2\lambda v_x\left(1+\alpha\right)+v_x^2\left(1-\alpha\right)^2}}-1\right)\frac{1}{1-a}\,. \qquad (27)
\end{aligned}$$

■

The term depending on $\mathrm{Tr}\left(\mathbf{\Sigma}\right)$ is an irreducible teacher-variance term, stemming from our definition of generalization, i.e., $\mathbb{E}\left[G_T\right] \triangleq \frac{1}{T}\sum_{i=1}^{T}\mathbb{E}\left\|\mathbf{w}_T - \mathbf{w}_i^\star\right\|^2$ (Definition 2), similar to the irreducible label-noise term in Eq. (1). The term here can be avoided by analyzing $\mathbb{E}\left\|\mathbf{w}_T - \mathbf{w}^\star\right\|^2$ instead.

E.3. Proof of Theorem 7: Asymptotic Generalization under Strong Regularization

Recall Theorem 7. Under the i.i.d. teacher setting (Assumption 1), when the task horizon is $T \to \infty$, the expected generalization loss is monotonically decreasing in the regularization strength λ. Furthermore, the iterates converge in mean square to the "global" teacher $\mathbf{w}^\star$, i.e.,

$$\lim_{\lambda\to\infty} \lim_{T\to\infty} \mathbb{E}\left\|\mathbf{w}_T - \mathbf{w}^\star\right\|^2 = 0.$$

Proof. We split the proof into two parts.

E.3.1. PROVING CONVERGENCE

Note that for any fixed i,

$$\begin{aligned}\lim_{\lambda\to\infty} \lim_{T\to\infty} \mathbb{E}\left\|\mathbf{w}_T - \mathbf{w}^\star\right\|^2 &= \lim_{\lambda\to\infty} \lim_{T\to\infty} \mathbb{E}\left[\left\|\mathbf{w}_T - (\mathbf{w}_i^\star + \boldsymbol{\xi}_i)\right\|^2\right] \\ &= \lim_{\lambda\to\infty} \lim_{T\to\infty} \left(\mathbb{E}\left\|\mathbf{w}_T - \mathbf{w}_i^\star\right\|^2 + 2\mathbb{E}\left[\left(\mathbf{w}_T - \mathbf{w}_i^\star\right)^\top \boldsymbol{\xi}_i\right] + \underbrace{\mathbb{E}\left\|\boldsymbol{\xi}_i\right\|^2}_{\mathrm{Tr}(\boldsymbol{\Sigma})}\right).\end{aligned}$$

We proceed by calculating $\lim_{\lambda\to\infty} \lim_{T\to\infty} \mathbb{E}\left[\left\|\mathbf{w}_T - \mathbf{w}_i^\star\right\|^2\right]$. Recall that by Eq. (27), for any fixed $i \geq 2$,

$$\begin{aligned}&\lim_{T\to\infty} \mathbb{E}\left[\left\|\mathbf{w}_T - \mathbf{w}_i^\star\right\|^2\right] \\ &= \underbrace{\frac{v_z}{2v_x}\left(\frac{v_x(1+\alpha)+\lambda}{\sqrt{\lambda^2 + 2\lambda v_x(1+\alpha) + v_x^2(1-\alpha)^2}} - 1\right)\frac{1}{1-a}}_{\text{term 1}} + \underbrace{2\,\mathrm{Tr}(\boldsymbol{\Sigma})\left(\frac{1-b}{1-a}\right)}_{\text{term 2}}.\end{aligned}$$

Then, by Lemma D.19, as $\lambda \to \infty$, we have $\frac{1-b}{1-a} = \frac{1}{2} + O\left(\lambda^{-1}\right)$, implying that, $\lim_{\lambda\to\infty} \frac{1-b}{1-a} = \frac{1}{2}$. Furthermore, Lemmas D.7 and D.20 suggest that as $\lambda \to \infty$, it holds that,

$$\frac{v_z}{2v_x}\left(\frac{v_x\left(1+\alpha\right)+\lambda}{\sqrt{\lambda^2 + 2\lambda v_x\left(1+\alpha\right) + v_x^2\left(1-\alpha\right)^2}} - 1\right)\frac{1}{1-a} = O\left(\lambda^{-1}\right),$$

implying $\lim_{\lambda\to\infty} \frac{v_z}{2v_x}\left(\frac{v_x(1+\alpha)+\lambda}{\sqrt{\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2}} - 1\right)\frac{1}{1-a} = 0$. So overall,

$$\lim_{\lambda\to\infty} \lim_{T\to\infty} \mathbb{E}\left[\left\|\mathbf{w}_T - \mathbf{w}_i^\star\right\|^2\right] = \mathrm{Tr}\left(\boldsymbol{\Sigma}\right).$$

Combining all of the above, for some fixed $i \geq 3$:

$$\begin{aligned}\lim_{\lambda\to\infty} \lim_{T\to\infty} \mathbb{E}\left\|\mathbf{w}_T - \mathbf{w}^\star\right\|^2 &= \lim_{\lambda\to\infty} \lim_{T\to\infty} \left(2\mathrm{Tr}\left(\boldsymbol{\Sigma}\right) + 2\mathbb{E}\left[\left(\mathbf{w}_T - \mathbf{w}_i^\star\right)^\top \boldsymbol{\xi}_i\right]\right) \\ [1] &= 2\mathrm{Tr}\left(\boldsymbol{\Sigma}\right) - 2\mathbb{E}\left[\left(\mathbf{w}_i^\star\right)^\top \boldsymbol{\xi}_i\right] = 2\mathrm{Tr}\left(\boldsymbol{\Sigma}\right) - 2\mathbb{E}\left[\left(\mathbf{w}^\star + \boldsymbol{\xi}_i\right)^\top \boldsymbol{\xi}_i\right] \\ &= 2\mathrm{Tr}\left(\boldsymbol{\Sigma}\right) - 2\underbrace{\mathbb{E}\left[\left\|\boldsymbol{\xi}_i\right\|^2\right]}_{\mathrm{Tr}(\boldsymbol{\Sigma})} = 0,\end{aligned}$$

where [1] follows by Eq. (10) where,

$$\mathbf{w}_T = \left[\prod_{m=T}^{1} \mathbf{P}_m\right](\mathbf{w}_0 - \mathbf{w}_1^\star) + \frac{1}{\lambda d}\sum_{k=1}^{T}\left[\prod_{m=T}^{k} \mathbf{P}_m\right]\mathbf{X}_k^\top \mathbf{z}_k + \sum_{k=2}^{T}\left[\prod_{m=T}^{k} \mathbf{P}_m\right]\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \mathbf{w}_T^\star.$$

Recall the identity and Eq. (5),

$$\begin{aligned}\mathbb{E}\left[\prod\nolimits_{m=T}^{i}\mathbf{P}_m\right] &= \mathbb{E}\left[\mathbf{P}_T\cdots\mathbf{P}_i\right] = \mathbb{E}_{\mathbf{P}_i,\ldots,\mathbf{P}_{T-1}}\left[\mathbb{E}_{\mathbf{P}_T}\left[\mathbf{P}_T\cdots\mathbf{P}_i\middle|\mathbf{P}_1,\ldots,\mathbf{P}_{T-1}\right]\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{P}_T\right]\mathbf{P}_{T-1}\cdots\mathbf{P}_i\right] = \mathbb{E}\left[\mathbf{P}_T\right]\mathbb{E}\left[\mathbf{P}_{T-1}\ldots\mathbf{P}_i\right]\\ &= \ldots = b^{T-i+1}\mathbf{I}\,.\end{aligned}$$

For any fixed $\lambda > 0$, we have $0 < b < 1$ (by Lemma D.2).

Hence, for any fixed i, by independence of $\boldsymbol{\xi}_i$ of all other random variables we have,

$$\lim_{T\to\infty}\mathbb{E}\left[\boldsymbol{\xi}_i^\top \mathbf{w}_T\right] = \lim_{T\to\infty}\mathrm{Tr}\left(\mathbb{E}\left[\left(\Pi_{m=T}^{i+1}\mathbf{P}_m - \Pi_{m=T}^{i}\mathbf{P}_m\right)\right]\boldsymbol{\Sigma}\right) = \lim_{T\to\infty}\mathrm{Tr}\left(b^{T-i}(1-b)\boldsymbol{\Sigma}\right) \to 0,$$

here, we used the fact that $\|\boldsymbol{\Sigma}\|_* \le C < \infty$.

E.3.2. PROVING MONOTONICITY

We show that each one of the terms in the following expression is monotonic decreasing with λ.

Recall Corollary 1, according to which,

$$\lim_{T\to\infty}\mathbb{E}\left[G_T\right] = \underbrace{2\mathrm{Tr}\left(\boldsymbol{\Sigma}\right)\left(\tfrac{1-b}{1-a}\right)}_{\text{term 1}} + \underbrace{\frac{v_z}{2v_x}\left(\frac{v_x\left(1+\alpha\right)+\lambda}{\sqrt{\lambda^2+2\lambda v_x\left(1+\alpha\right)+v_x^2\left(1-\alpha\right)^2}} - 1\right)\tfrac{1}{1-a}}_{\text{term 2}}. \quad (28)$$

First, we analyze the **first term** in the summation:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}2\mathrm{Tr}\left(\boldsymbol{\Sigma}\right)\left(\frac{1-b}{1-a}\right) = 2\mathrm{Tr}\left(\boldsymbol{\Sigma}\right)\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{1-b}{1-a}\,.$$

Next, show that $\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{1-b}{1-a} < 0$. Recall the notations from Table 2,

$$\begin{aligned}1-b &= 1-\frac{1}{2}\left(1-\alpha-\tilde{\lambda}+D\right) = \frac{1}{2}\left(1+\alpha+\tilde{\lambda}-D\right)\\ 1-a &= 1-\frac{1}{2}\left(1-\alpha+\frac{N}{D}\right) = \frac{1}{2}\left(1+\alpha-\frac{N}{D}\right)\,.\end{aligned}$$

Using the above closed-form expression and by Lemmas D.3 and D.4,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{1-b}{1-a} = \frac{1-b}{(1-a)^2}\cdot\frac{\mathrm{d}a}{\mathrm{d}\lambda} - \frac{1}{1-a}\cdot\frac{\mathrm{d}b}{\mathrm{d}\lambda} \propto (1-b)\frac{\mathrm{d}a}{\mathrm{d}\lambda} - (1-a)\frac{\mathrm{d}b}{\mathrm{d}\lambda} \\
&= \frac{1}{2}\left(1+\alpha+\tilde{\lambda}-D\right)\frac{1}{2v_x}\frac{4\tilde{\lambda}\alpha}{D^3} - \frac{1}{2}\left(1+\alpha-\frac{N}{D}\right)\left(\frac{1}{2v_x}\left(\frac{-D+\tilde{\lambda}+1+\alpha}{D}\right)\right) \\
&= \frac{1}{4v_x}\left(\frac{1+\alpha+\tilde{\lambda}-D}{D}\right)\frac{4\tilde{\lambda}\alpha}{D^2} - \left(\frac{1}{4v_x}\left(\frac{1+\alpha+\tilde{\lambda}-D}{D}\right)\right)\left(1+\alpha-\frac{N}{D}\right) \\
&= \frac{1}{4v_x}\left(\frac{1+\alpha+\tilde{\lambda}-D}{D}\right)\left(\frac{4\tilde{\lambda}\alpha}{D^2} - \left(\frac{(1+\alpha)D^2-DN}{D^2}\right)\right) \\
&\overset{(1)}{=} \frac{1}{4v_x}\left(\frac{1+\alpha+\tilde{\lambda}-D}{D}\right)\left(\frac{(1+\alpha)D^2 - N\left(\tilde{\lambda}+1+\alpha\right) - (1+\alpha)D^2 + DN}{D^2}\right) \\
&= \left(\frac{1}{4v_x}\left(\frac{1+\alpha+\tilde{\lambda}-D}{D}\right)\right)\left(\frac{-N\left(1+\alpha+\tilde{\lambda}-D\right)}{D^2}\right) \\
&= \frac{-N}{4v_x}\frac{\left(1+\alpha+\tilde{\lambda}-D\right)^2}{D^3} < 0,
\end{aligned} \tag{29}$$

where [1] follows from Eq. (14).

Overall, since $\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{1-b}{1-a} < 0$ and $\mathrm{Tr}(\mathbf{\Sigma}) \geq 0$, the derivative of the first term with respect to λ is non-positive.

For the **second term** in the summation, we define

$$k \triangleq v_z \frac{1}{2v_x}\left(\frac{v_x(1+\alpha)+\lambda}{\sqrt{\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2}} - 1\right)\frac{1}{1-a}.$$

We wish to show that $\frac{\mathrm{d}}{\mathrm{d}\lambda}k < 0$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda}k =& v_z\frac{1}{2v_x}\underbrace{\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{v_x(1+\alpha)+\lambda}{\sqrt{\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2}} - 1\right)}_{\text{term 1}}\frac{1}{1-a} \\
&+ v_z\frac{1}{2v_x}\left(\frac{v_x(1+\alpha)+\lambda}{\sqrt{\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2}} - 1\right)\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{1}{1-a}.
\end{aligned}$$

Using Lemma D.3, we get,

$$\frac{\mathrm{d}\frac{1}{1-a}}{\mathrm{d}\lambda} = \frac{1}{(1-a)^2}\frac{2\tilde{\lambda}\alpha}{v_x D^3} > 0.$$

Evaluating the derivative with respect to λ of the first term,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{v_x(1+\alpha)+\lambda}{\sqrt{\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2}}-1\right]\\
&=\frac{1}{\sqrt{\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2}}-\frac{(\lambda+v_x(1+\alpha))(v_x(1+\alpha)+\lambda)}{\left(\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2\right)^{\frac{3}{2}}}\\
&=\frac{\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2-\left(\lambda^2+(v_x(1+\alpha))^2+2\lambda v_x(1+\alpha)\right)}{\left(\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2\right)^{\frac{3}{2}}}\\
&=\frac{v_x^2(1-\alpha)^2-v_x^2(1+\alpha)^2}{\left(\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2\right)^{\frac{3}{2}}}=\frac{-4v_x^2\alpha}{\left(\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2\right)^{\frac{3}{2}}}<0.
\end{aligned}$$

Combining the derivations above, we obtain

$$\begin{aligned}
\frac{\mathrm{d}k}{\mathrm{d}\lambda}&=\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[v_z\frac{1}{2v_x}\left(\frac{v_x(1+\alpha)+\lambda}{\sqrt{\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2}}-1\right)\right]\cdot\frac{1}{1-a}+\\
&\qquad\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{1-a}\right)\cdot v_z\frac{1}{2v_x}\left(\frac{v_x(1+\alpha)+\lambda}{\sqrt{\lambda^2+2\lambda v_x(1+\alpha)+v_x^2(1-\alpha)^2}}-1\right)\\
&=\frac{v_z}{2v_x}\left(\frac{-4v_x^2\alpha}{v_x^3D^3}\right)\cdot\frac{1}{1-a}+\frac{1}{(1-a)^2}\frac{4\tilde{\lambda}\alpha}{2v_xD^3}\frac{v_z}{2v_x}\left(\frac{v_x(1+\alpha)+\lambda}{v_xD}-1\right)\\
&=\frac{2v_z}{v_x^2}\frac{\alpha}{(1-a)D^3}\left(\frac{1}{(1-a)}\frac{\tilde{\lambda}}{2v_xD}(v_x(1+\alpha-D)+\lambda)-1\right)\\
&=\frac{2v_z}{v_x^2}\frac{\alpha}{(1-a)^2D^3}\left(\frac{\tilde{\lambda}}{2v_xD}(v_x(1+\alpha-D)+\lambda)-\frac{1+\alpha}{2}-\frac{\tilde{\lambda}(1+\alpha)+(1-\alpha)^2}{2D}\right)\\
&=\frac{v_z}{v_x^2}\frac{\alpha}{(1-a)^2D^4}\left(\tilde{\lambda}\left((1+\alpha-D)+\tilde{\lambda}\right)-\left(D(1+\alpha)-\tilde{\lambda}(1+\alpha)-(1-\alpha)^2\right)\right)\\
&=\frac{v_z}{v_x^2}\frac{\alpha}{(1-a)D^4(1-a)}\left(\tilde{\lambda}^2+2(1+\alpha)\tilde{\lambda}+(1-\alpha)^2-\left(1+\alpha+\tilde{\lambda}\right)D\right)\\
&=\frac{v_z}{v_x^2}\frac{\alpha}{(1-a)^2D^3}\left(D-\left(1+\alpha+\tilde{\lambda}\right)\right) \qquad (30)\\
&\overset{[\alpha>0]}{\leq}\frac{v_z}{v_x^2}\frac{\alpha}{(1-a)^2D^3}\left(\sqrt{\left(\tilde{\lambda}+(1+\alpha)\right)^2}-\left(1+\alpha+\tilde{\lambda}\right)\right)=0\,, \qquad (31)
\end{aligned}$$

implying that $\lim_{T\to\infty}\mathbb{E}[G_T]$ is monotonic decreasing as a function of λ for all $\lambda>0$. ■

Remark E.2 *If either* $\mathrm{Tr}(\mathbf{\Sigma})>0$ *or* $v_z>0$*, then* $\lim_{T\to\infty}\mathbb{E}[G_T]$ *is strictly decreasing in* λ *for all* $\lambda>0$*: when* $v_z>0$*, the inequality in Eq.* (31) *becomes strict; and when* $\mathrm{Tr}(\mathbf{\Sigma})>0$*, the derivative of the first term in Eq.* (28) *is strictly negative.*

E.4. Proof of Lemma 5: More Tasks Can Be Useful or Harmful

Recall Lemma 5. Consider the i.i.d. teacher setting (Assumption 1) with a label noise $v_z > 0$.

- **Single teacher:** When $\mathbf{w}_1^\star = \cdots = \mathbf{w}_T^\star$ (i.e., $\mathrm{Tr}\left(\mathbf{\Sigma}\right) = 0$), the expected generalization loss with an *optimal* horizon-dependent strength $\lambda^\star(T)$ is monotonically *decreasing* in the task horizon T.
- **Multiple i.i.d. teachers:** When teachers vary across tasks (i.e., $\mathrm{Tr}(\mathbf{\Sigma}) > 0$), then under any *fixed* strength $\lambda > 0$, the expected generalization loss is monotonically *increasing* in the task horizon T for all $T \geq T'$ for a sufficiently large T'.

Next, we prove each case separately.

E.4.1. PROVING LEMMA 5 IN THE SINGLE TEACHER CASE

Proof. Recall Theorem 3 when $\mathbf{w}_t^\star = \mathbf{w}^\star$ for all $t \in [T]$, and $v_z > 0$ the expression corresponding to $\mathbb{E}[G_T]$ can be simplified to a more compact form,

$$\begin{aligned}\mathbb{E}[G_T] &= \mathbb{E}\left\|\mathbf{w}_T - \mathbf{w}^\star\right\|^2 \\ &= a^T \left\|\mathbf{w}_0 - \mathbf{w}^\star\right\|^2 + \frac{v_z}{v_x}\frac{1-a^T}{\frac{\lambda}{v_x} + \sqrt{\left(\frac{\lambda}{v_x}\right)^2 + 2\frac{\lambda}{v_x}\left(1+\alpha\right) + \left(1-\alpha\right)^2}},\end{aligned} \tag{32}$$

Define $s = \frac{\lambda}{v_x} + \sqrt{\left(\frac{\lambda}{v_x}\right)^2 + 2\frac{\lambda}{v_x}\left(1+\alpha\right) + \left(1-\alpha\right)^2}$, and

$$\begin{aligned}J\left(t,\lambda\right) &= a^t \left\|\mathbf{w}_0 - \mathbf{w}^\star\right\|^2 + \frac{v_z}{v_x}\frac{1-a^t}{s} \\ \frac{\mathrm{d}}{\mathrm{d}\lambda}J\left(t,\lambda\right) &= \underbrace{\left[\left\|\mathbf{w}_0 - \mathbf{w}^\star\right\|^2 - \frac{v_z}{v_x}\frac{1}{s}\right] t a^{t-1}\frac{\mathrm{d}a}{\mathrm{d}\lambda}}_{\text{term 1}} - \underbrace{\frac{v_z}{v_x}\frac{1-a^t}{s^2}\frac{\mathrm{d}s}{\mathrm{d}\lambda}}_{\text{term 2}}.\end{aligned} \tag{33}$$

We will show that $\frac{\mathrm{d}}{\mathrm{d}\lambda}J\left(t,\lambda\right)\big|_{\lambda=0} < 0$ when $\alpha < 1$, otherwise if $\alpha = 1$, the existence of $\bar{\lambda} > 0$ such that $\frac{\mathrm{d}}{\mathrm{d}\lambda}J\left(t,\lambda\right)\big|_{\lambda=\bar{\lambda}} < 0$ and $\lim_{\tilde{\lambda}\to\infty}\frac{\mathrm{d}}{\mathrm{d}\lambda}J\left(t,\lambda\right)\big|_{\lambda=\tilde{\lambda}} > 0$ for the defined scaled parameter $\tilde{\lambda} \triangleq \frac{\lambda}{v_x}$, note that if $\tilde{\lambda} \to \infty$ then $\lambda \to \infty$, continuity implies at least one stationary point $\lambda^\star \in (0,\infty)$ and it is a minimum.

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}J\left(t,\lambda\right)\bigg|_{\lambda=0} = \frac{\mathrm{d}a}{\mathrm{d}\lambda}\big|_{\lambda=0}=0 \; -\frac{v_z}{v_x}\frac{1-\left(1-\alpha\right)^t}{\left(1-\alpha\right)^2}\frac{\mathrm{d}s}{\mathrm{d}\lambda} \overset{[0<\alpha<1]}{<} 0 \tag{34}$$

$$\begin{aligned}\lim_{\tilde{\lambda}\to\infty}\frac{\mathrm{d}}{\mathrm{d}\lambda}J\left(t,\lambda\right)\bigg|_{\lambda=\tilde{\lambda}} &= \lim_{\tilde{\lambda}\to\infty}\left[\left\|\mathbf{w}_0 - \mathbf{w}^\star\right\|^2 - \underbrace{\frac{v_z}{v_x}\frac{1}{s}}_{\to 0}\right] t a^{t-1}\frac{\mathrm{d}a}{\mathrm{d}\tilde{\lambda}} - \frac{v_z}{v_x}\frac{1-a^t}{s^2}\frac{\mathrm{d}s}{\mathrm{d}\tilde{\lambda}} \\ &\overset{[1]}{\geq} \lim_{\tilde{\lambda}\to\infty}\tfrac{1}{2}\left\|\mathbf{w}_0 - \mathbf{w}^\star\right\|^2 t\left(1+O\left(\tilde{\lambda}^{-1}\right)\right)\left(\frac{2\alpha}{\tilde{\lambda}^2} + O\left(\tilde{\lambda}^{-3}\right)\right) - \frac{v_z}{v_x}\frac{\frac{2\alpha t}{\tilde{\lambda}} + O\left(\tilde{\lambda}^{-2}\right)}{\left(2\tilde{\lambda} + O(1)\right)^2}\left(1+O\left(\tilde{\lambda}^{-2}\right)\right) \\ &= \lim_{\tilde{\lambda}\to\infty}\frac{1}{2}\left\|\mathbf{w}_0 - \mathbf{w}^\star\right\|^2 t\left(\frac{2\alpha}{\tilde{\lambda}^2} + O\left(\tilde{\lambda}^{-3}\right)\right) - \frac{v_z}{v_x}O\left(\tilde{\lambda}^{-3}\right)\left(1 + O\left(\tilde{\lambda}^{-2}\right)\right) > 0\,,\end{aligned} \tag{35}$$

where [1] follows by Lemmas D.16 and D.18 (as all terms are analytic at ∞ being a concatenation, sum, and division of analytic functions) and $s \to \infty$ as $\tilde{\lambda} \to \infty$.

For $\alpha = 1$, we note that $a \to 0$ and $\frac{da}{d\lambda} \to 0$ as $\lambda \to 0^+$ by Lemmas D.3 and D.10, since the exponential decay in a^t dominates any polynomial growth in t. Moreover, $\frac{\mathrm{d}s}{\mathrm{d}\lambda} > \frac{1}{v_x} > 0$ for all $\lambda > 0$. Thus, the second term in $\frac{d}{d\lambda} J(t, \lambda)$ (Eq. (33)) dominates as $\lambda \to 0^+$, while remaining strictly negative. By continuity, there exists $\bar{\lambda} > 0$ such that for all $\lambda \in (0, \bar{\lambda}]$ we have $\frac{\mathrm{d}}{\mathrm{d}\lambda} J(t, \lambda) < 0$. Hence, near $\lambda = 0$ the derivative is negative and the second term dominates.

At a stationary point we have

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} J\left(t, \lambda^\star\right) = 0 \iff \left[\|\mathbf{w}_0 - \mathbf{w}^\star\|^2 - \frac{v_z}{v_x}\frac{1}{s}\right] t a^{t-1} \left.\frac{\mathrm{d}a}{\mathrm{d}\lambda}\right|_{\lambda=\lambda^\star} = \frac{v_z}{v_x}\frac{1-a^t}{s^2} \left.\frac{\mathrm{d}s}{\mathrm{d}\lambda}\right|_{\lambda=\lambda^\star},$$

When differentiating with respect to t,

$$\frac{\mathrm{d}}{\mathrm{d}t} J\left(t, \lambda^\star\right) = a^t \ln a \left[\|\mathbf{w}_0 - \mathbf{w}^\star\|^2 - \frac{v_z}{v_x}\frac{1}{s}\right] \overset{[2]}{=} \underbrace{\frac{a}{t \left.\frac{\mathrm{d}a}{\mathrm{d}\lambda}\right|_{\lambda=\lambda^\star}}}_{>0} \underbrace{\ln a}_{<0} \underbrace{\frac{v_z}{v_x}\frac{1-a^t}{s^2} \left.\frac{\mathrm{d}s}{\mathrm{d}\lambda}\right|_{\lambda=\lambda^\star}}_{>0} < 0. \tag{36}$$

Where [2] follows by Lemmas D.2 and D.3 as $\lambda^\star > 0$ and since $v_z > 0$.

From Eq. (35) it follows that there exists Λ such that for all $\lambda \geq \Lambda$ we have $\frac{d}{d\lambda} J(t, \lambda) > 0$. Moreover, if $\alpha < 1$ then $\left.\frac{d}{d\lambda} J(t, \lambda)\right|_{\lambda=0} < 0$, while if $\alpha = 1$ there exists $\lambda' > 0$ such that $\left.\frac{d}{d\lambda} J(t, \lambda)\right|_{\lambda=\lambda'} < 0$. from Eq. (34), the Intermediate Value Theorem guarantees a point $\lambda^\star \in (0, \Lambda]$ such that $\frac{\mathrm{d}}{\mathrm{d}\lambda^\star} J\left(t, \lambda^\star\right) = 0$. On the compact interval $I = [0, \Lambda]$ if $\alpha < 1$, and $I = [\bar{\lambda}, \Lambda]$ if $\alpha = 1$, the function $J(t, \cdot)$ attains a global minimum by Weierstrass, since the derivative is negative at the left end-point and is positive at Λ, the global minimizer is a stationary point, by Eq. (36) the global minimum is decreasing with respect to t. ■

E.4.2. PROVING LEMMA 5 IN THE CASE OF MULTIPLE I.I.D. TEACHERS

Proof. Recall Eq. (26), for all $T > 1$ we have,

$$\begin{aligned}\mathbb{E}[G_T] &= \left(a^T - 2a^{T-1}b\right)\mathrm{Tr}\left(\mathbf{\Sigma}\right) + a^T\left\|\mathbf{w}^\star\right\|^2 + 2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{a\left(1-a^{T-1}\right)}{1-a} - b\frac{a\left(1-a^{T-2}\right)}{1-a}\right)\\ &+ v_z\frac{1}{2v_x}\left(\frac{v_x\left(1+\alpha\right)+\lambda}{\sqrt{\lambda^2+2\lambda v_x\left(1+\alpha\right)+v_x^2\left(1-\alpha\right)^2}} - 1\right)\frac{1-a^T}{1-a}\\ &+ \frac{2}{T}\left(2b^T + T - 1 - Tb\right)\mathrm{Tr}\left(\mathbf{\Sigma}\right).\end{aligned}$$

Define $g(t) = \mathbb{E}\left[G_t\right]$, by the assumptions $\mathrm{Tr}\left(\mathbf{\Sigma}\right) > 0$. Then,

$$\begin{aligned}g'\left(t\right) &= \mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(a^t\ln a - 2ba^{t-1}\ln a\right) + \left\|\mathbf{w}^\star\right\|^2 a^t\ln a +\\ &+ \frac{2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(-a^t\ln a + ba^{t-1}\ln a\right)}{1-a} + \frac{4\mathrm{Tr}\left(\mathbf{\Sigma}\right)}{t}b^t\left(\ln\left(b\right) - \frac{1}{t}\right) + \frac{2}{t^2}\mathrm{Tr}\left(\mathbf{\Sigma}\right) - \frac{cv_z a^t\ln a}{1-a}\\ &= a^t\Bigg[\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(\ln a - \frac{2b}{a}\ln a\right) + \left\|\mathbf{w}^\star\right\|^2\ln a + \frac{2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(-\ln a + \frac{b}{a}\ln a\right)}{1-a}\\ &\quad + \frac{4\mathrm{Tr}\left(\mathbf{\Sigma}\right)}{a^t t}b^t\left(\ln\left(b\right) - \frac{1}{t}\right) + \frac{2}{a^t t^2}\mathrm{Tr}\left(\mathbf{\Sigma}\right) - \frac{cv_z\ln a}{1-a}\Bigg].\end{aligned}$$

Defining $A = \mathrm{Tr}\left(\mathbf{\Sigma}\right)\ln a\left[\left(1-\frac{2b}{a}\right) + \frac{\left\|\mathbf{w}^\star\right\|^2}{\mathrm{Tr}(\mathbf{\Sigma})} + \frac{2\left(\frac{b}{a}-1\right)}{1-a}\right] - \frac{cv_z\ln a}{1-a}$, we have,

$$\frac{g'\left(t\right)}{a^t} = A + \frac{4\mathrm{Tr}\left(\mathbf{\Sigma}\right)}{t}\left(\frac{b}{a}\right)^t\left(\ln\left(b\right) - \frac{1}{t} + \frac{1}{2tb^t}\right).$$

There exists t' such that for all $t \geq t'$ we have $\ln\left(b\right) - \frac{1}{t} + \frac{1}{2tb^t} > 0$ since $0 < b < 1$ by Lemma D.2. Since $b > a$ by Lemma D.1 for all $\lambda > 0$, then there exists $t^\star > t'$ which for all $t \geq t^\star$ we have $\underbrace{A}_{\text{constant in t}} + \frac{4\mathrm{Tr}(\mathbf{\Sigma})}{t}\left(\frac{b}{a}\right)^t\underbrace{\left(\ln\left(b\right) - \frac{1}{t} + \frac{1}{2tb^t}\right)}_{>0} > 0$ then $g'\left(t\right) > 0$ implying that $g\left(t\right)$ is increasing for all $t \geq t^\star$, thus $\mathbb{E}[G_t]$ is increasing for all $t \geq \max(2, t^\star)$. ■

Before proving Theorem 6, we introduce several auxiliary lemmas.

Lemma E.3 $\frac{\mathrm{d}}{\mathrm{d}\lambda}\lim_{T\to\infty}\mathbb{E}\left[G_T\right] = -\frac{1}{2v_x}\left(\frac{v_z}{v_x}+\mathrm{Tr}(\mathbf{\Sigma})\left(\alpha+1\right)\right)\tilde{\lambda}^{-2}+O\left(\tilde{\lambda}^{-3}\right)$.

Proof. In the proof of Theorem 7, we have shown that $\frac{\mathrm{d}}{\mathrm{d}\lambda}\lim_{T\to\infty}\mathbb{E}[G_t]$ decomposes into two terms:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\lim_{T\to\infty}\mathbb{E}[G_t] = \frac{v_z}{v_x^2}\frac{\alpha}{(1-a)^2D^3}\Big(D-(1+\alpha+\tilde{\lambda})\Big)+\mathrm{Tr}(\mathbf{\Sigma})\frac{-N}{2v_x}\frac{\Big(1+\alpha+\tilde{\lambda}-D\Big)^2}{(1-a)^2D^3}. \quad (37)$$

From Lemma D.21, we know that $D=\tilde{\lambda}+\alpha+1-\frac{2\alpha}{\tilde{\lambda}+\alpha+1}+O\left(\tilde{\lambda}^{-3}\right)$. Thus,

$$1+\alpha+\tilde{\lambda}-D=\frac{1}{\tilde{\lambda}}\frac{2\alpha}{1+\frac{\alpha+1}{\tilde{\lambda}}}+O\left(\tilde{\lambda}^{-3}\right)=\frac{2\alpha}{\tilde{\lambda}}+O\left(\tilde{\lambda}^{-2}\right).$$

Expanding them term by term, using Lemma D.18, we begin with the first term in Eq. (37),

$$\begin{aligned}\frac{v_z}{v_x^2}\frac{\alpha}{(1-a)^2D^3}\Big(D-(1+\alpha+\tilde{\lambda})\Big) &= -\frac{v_z}{v_x^2}\left(\frac{\alpha\,\tilde{\lambda}^2}{4\alpha^2}+O(\tilde{\lambda})\right)\left(\tilde{\lambda}^{-3}+O(\tilde{\lambda}^{-4})\right)\left(\frac{2\alpha}{\tilde{\lambda}}+O(\tilde{\lambda}^{-2})\right)\\ &= -\frac{v_z}{2v_x^2}\,\tilde{\lambda}^{-2}+O(\tilde{\lambda}^{-3}),\end{aligned}$$

and the second term in Eq. (37) becomes,

$$\begin{aligned}&\mathrm{Tr}(\mathbf{\Sigma})\,\frac{-N}{2v_x}\left(\frac{4\alpha^2}{\tilde{\lambda}^2}+O(\tilde{\lambda}^{-3})\right)\left(\frac{\tilde{\lambda}^2}{4\alpha^2}+O(\tilde{\lambda})\right)\left(\tilde{\lambda}^{-3}+O(\tilde{\lambda}^{-4})\right)\\ &=\mathrm{Tr}(\mathbf{\Sigma})\,\frac{-N}{2v_x}\left(\tilde{\lambda}^{-3}+O(\tilde{\lambda}^{-4})\right).\end{aligned}$$

Since $N\triangleq(1+\alpha)\tilde{\lambda}+(1-\alpha)^2$,

$$\mathrm{Tr}(\mathbf{\Sigma})\,\frac{-1}{2v_x}\Big((1+\alpha)\tilde{\lambda}+(1-\alpha)^2\Big)\left(\tilde{\lambda}^{-3}+O(\tilde{\lambda}^{-4})\right)=-\mathrm{Tr}(\mathbf{\Sigma})\,\frac{1+\alpha}{2v_x}\,\tilde{\lambda}^{-2}+O(\tilde{\lambda}^{-3}).$$

Overall, we showed that the second term in Eq. (37) is,

$$\mathrm{Tr}(\mathbf{\Sigma})\,\frac{-N}{2v_x}\frac{\Big(1+\alpha+\tilde{\lambda}-D\Big)^2}{(1-a)^2D^3}=-\mathrm{Tr}(\mathbf{\Sigma})\,\frac{1+\alpha}{2v_x}\,\tilde{\lambda}^{-2}+O\left(\tilde{\lambda}^{-3}\right).$$

■

Additionally, we prove a slightly weaker Lemma than Theorem 6. This establishes a nearly linear scaling of $\lambda^\star(T)$, which will help us derive the main result in Theorem 6.

E.5. Auxiliary Lemma: Optimal Regularization is Near Linear in T

Lemma E.4 (Optimal regularization strength is near linear in T) *Under the i.i.d.-teacher setting (Assumption 1) with non-zero mean teacher, if either $\mathrm{Tr}(\mathbf{\Sigma}) > 0$ or $v_z > 0$ the optimal fixed regularization strength $\lambda^\star$ that minimizes the expected generalization loss after T iterations satisfies $C_1 T^\rho < \lambda^\star < C_0 T$, for all sufficiently large T and every $\rho < 1$, where $C_0, C_1 > 0$ are constants depending only on the problem parameters.*

We first provide a brief proof outline.

Proof outline. Set $\tilde{\lambda} \triangleq \lambda / v_x$ and fix a cutoff $\Lambda(T) > 0$; we analyze $\tilde{\lambda} \in [0, \Lambda(T)]$. Let $D = \sqrt{\tilde{\lambda}^2 + 2(1+\alpha)\tilde{\lambda} + (1-\alpha)^2}$. We show that $a, b \le \zeta \triangleq 1 - \frac{\alpha}{1+\alpha+\Lambda(T)}$.

1. **Limit objective is strictly decreasing.** By the closed form of the infinite-horizon loss,

$$\lim_{T'\to\infty} \mathbb{E}[G_{T'}] = 2\mathrm{Tr}(\mathbf{\Sigma})\Big(\tfrac{1-b}{1-a}\Big) + v_z \frac{1}{2v_x}\Big(\frac{1+\alpha+\tilde{\lambda}}{D} - 1\Big)\frac{1}{1-a}.$$

   From Remark E.2 we have

$$\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}} \lim_{T'\to\infty} \mathbb{E}[G_{T'}] \;<\; 0 \qquad \forall\, \tilde{\lambda} > 0.$$

2. **In the sublinear regime, the derivative is also negative.** Using the decomposition (algebra is provided in the full proof),

$$\mathbb{E}[G_T] - \lim_{T'\to\infty} \mathbb{E}[G_{T'}] = a^T\Big(\|\mathbf{w}^\star\|^2 + \mathrm{Tr}(\mathbf{\Sigma}) - 2\mathrm{Tr}(\mathbf{\Sigma})\tfrac{1-b}{1-a} - \tfrac{f}{1-a}\Big) + \tfrac{2\mathrm{Tr}(\mathbf{\Sigma})}{T}(2b^T - 1), \quad (38)$$

   Differentiating and using $a, b \le \zeta$, boundedness of $\frac{1-b}{1-a}$ and its derivative, and boundedness of f and its derivative (see details in the proof), gives the uniform bound

$$\sup_{\tilde{\lambda}\in[0,\Lambda(T)]} \Big|\tfrac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\big(\mathbb{E}[G_T] - \lim_{T'\to\infty}\mathbb{E}[G_{T'}]\big)\Big| \;\le\; C\,T\,\zeta^{T-5}\,(1-a)^{-2}\,.$$

   Since $(1-a)^{-2} = O(\Lambda(T)^2)$ and $T\zeta^{T-5} = T\exp(-\Theta(T/\Lambda(T)))$, if $\Lambda(T) \in o(T^\rho)$ for any fixed $0 < \rho < 1$, the RHS decay exponentially. Meanwhile,

$$\inf_{\tilde{\lambda}\in[0,\Lambda(T)]} \Big(-\tfrac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}} \lim_{T'\to\infty} \mathbb{E}[G_{T'}]\Big) = \Omega(\Lambda(T)^{-2}) = \omega(T^{-2\rho})\,.$$

   Hence, for large T, the derivative stays negative in $[0, \Lambda(T)]$. Therefore, $\tilde{\lambda}^\star \notin o(T^\rho)$ for every $0 < \rho < 1$.

3. **In superlinear regimes, the derivative turns positive.** For $\tilde{\lambda} \in \omega(T)$, Laurent expansions give
$$a^T = 1 - \frac{2\alpha T}{\tilde{\lambda}} + O(T^2\tilde{\lambda}^{-2}), \quad b^T = 1 - \frac{\alpha T}{\tilde{\lambda}} + O(T^2\tilde{\lambda}^{-2}), \quad \frac{1-b}{1-a} = \frac{1}{2} + O(\tilde{\lambda}^{-1}),$$
$$c = O(\tilde{\lambda}^{-2}), \quad (1-a)^{-1} = O(\tilde{\lambda}).$$
Plugging into the expression above (Eq. (38)),
$$\mathbb{E}[G_T] - \lim_{T'\to\infty} \mathbb{E}[G_{T'}] = \|\mathbf{w}^\star\|^2 - \frac{2\alpha T}{\tilde{\lambda}}\|\mathbf{w}^\star\|^2 + \frac{2\mathrm{Tr}(\mathbf{\Sigma})}{T} + O(T^2\tilde{\lambda}^{-2}) + O(\tilde{\lambda}^{-1}),$$
$$\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\mathbb{E}[G_T] = \frac{2\alpha T}{\tilde{\lambda}^2} + O(T^2\tilde{\lambda}^{-3}) + O(\tilde{\lambda}^{-2}) + \frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}} \lim_{T'\to\infty} \mathbb{E}[G_{T'}] \ > \ 0,$$
and $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}} \lim_{T'\to\infty} \mathbb{E}[G_{T'}] = O(\tilde{\lambda}^{-2})$. Hence, no optimizer lies in $\omega(T)$.

4. **Conclusion.** Combining (2) + (3), any minimizer satisfies $\tilde{\lambda}^\star \in O(T)$ and $\tilde{\lambda}^\star \in \Omega(T^\rho)$ for all $0 < \rho < 1$, since for $\tilde{\lambda} = \frac{\lambda}{v_x}$ the same asymptotic bounds hold for $\lambda^\star$.

■

Proof for Lemma E.4. For the actual proof, we first analyze the asymptotic regime $T \to \infty$.

In the sublinear regime, the derivative is negative. Let $\Lambda(T)$ be a positive function of T, whose form will be specified later. We focus on regularization strengths satisfying $\lambda \in [0, \Lambda(T)]$, and define the corresponding scaled parameter $\tilde{\lambda} \triangleq \lambda/v_x$.

Let $\bar{D} = \sup_{[0,\Lambda(T)]} D$; $\underline{D} = \inf_{[0,\Lambda(T)]} D$; and $\zeta = 1 - \frac{\alpha}{1+\alpha+\Lambda(T)}$. Note that $a, b \le \zeta$ by Lemma D.7. Recall Remark E.2 which explained that $\frac{\mathrm{d}}{\mathrm{d}\lambda} \lim_{T\to\infty} \mathbb{E}[G_T] < 0, \ \forall \lambda > 0$. Clearly, the same holds for the scaled parameter $\tilde{\lambda} = \frac{\lambda}{v_x}$, i.e., $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}} \lim_{T\to\infty} \mathbb{E}[G_T] < 0, \ \forall \tilde{\lambda} > 0$.

We now wish to use this fact. If the difference between $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}} \lim_{T'\to\infty} \mathbb{E}[G_{T'}]$ and $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\mathbb{E}[G_T]$ is sufficiently small, then we can ensure that $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\mathbb{E}[G_T] < 0$.

Recall $f \triangleq v_z \frac{1}{2v_x}\left(\frac{(1+\alpha)+\tilde{\lambda}}{\sqrt{\tilde{\lambda}^2+2\tilde{\lambda}(1+\alpha)+(1-\alpha)^2}} - 1\right)$ then by Corollary 1 and Eq. (26), we have that

$$\begin{aligned}
&\mathbb{E}[G_T] - \lim_{T'\to\infty} \mathbb{E}[G_{T'}] = \\
&= \underbrace{\left(a^T - 2a^{T-1}b\right)\mathrm{Tr}(\mathbf{\Sigma}) + a^T\|\mathbf{w}^\star\|^2 + 2\mathrm{Tr}(\mathbf{\Sigma})\left(\frac{a\left(1-a^{T-1}\right)}{1-a} - b\cdot\frac{a\left(1-a^{T-2}\right)}{1-a}\right)}_{\text{first part of } \mathbb{E}[G_T]} \\
&\quad \underbrace{+f\frac{1-a^T}{1-a} + \frac{2}{T}\left(2b^T + T - 1 - Tb\right)\mathrm{Tr}(\mathbf{\Sigma})}_{\text{other part of } \mathbb{E}[G_T]} - \underbrace{\left(2\mathrm{Tr}(\mathbf{\Sigma})\left(\frac{a-ba}{1-a} - b + 1\right) + f\frac{1}{1-a}\right)}_{\lim_{T'\to\infty}\mathbb{E}[G_{T'}]} \\
&= \left(a^T - 2a^{T-1}b\right)\mathrm{Tr}(\mathbf{\Sigma}) + a^T\|\mathbf{w}^\star\|^2 + 2\mathrm{Tr}(\mathbf{\Sigma})\left(\frac{-a^T}{1-a} + b\cdot\frac{a^{T-1}}{1-a}\right) \\
&\qquad + f\frac{-a^T}{1-a} + \frac{2}{T}\left(2b^T - 1\right)\mathrm{Tr}(\mathbf{\Sigma}) \\
&= a^T\left(\|\mathbf{w}^\star\|^2 + \mathrm{Tr}(\mathbf{\Sigma})\right) + f\frac{-a^T}{1-a} - 2\mathrm{Tr}(\mathbf{\Sigma})\,a^T\left(\frac{1-b}{1-a}\right) + \frac{2\mathrm{Tr}(\mathbf{\Sigma})}{T}\left(2b^T - 1\right). \qquad (39)
\end{aligned}$$

Using Eq. (39) taking the derivative according to $\tilde{\lambda}$

$$\left|\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\left[\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right]-\mathbb{E}\left[G_T\right]\right]\right| \tag{40}$$

$$=\Bigg|Ta^{T-1}\frac{\mathrm{d}a}{\mathrm{d}\tilde{\lambda}}+f\frac{-\left(Ta^{T-1}-Ta^T+a^T\right)}{\left(1-a\right)^2}\frac{\mathrm{d}a}{\mathrm{d}\tilde{\lambda}}+\frac{\mathrm{d}f}{\mathrm{d}\tilde{\lambda}}\frac{-a^T}{1-a}$$
$$-2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(Ta^{T-1}\frac{\mathrm{d}a}{\mathrm{d}\tilde{\lambda}}\left(\frac{1-b}{1-a}\right)+a^T\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\left(\frac{1-b}{1-a}\right)\right)+\frac{4\mathrm{Tr}\left(\mathbf{\Sigma}\right)}{T}Tb^{T-1}\frac{\mathrm{d}b}{\mathrm{d}\tilde{\lambda}}\Bigg|. \tag{41}$$

If $\alpha<1$, we first show that $\frac{1-b}{1-a}$ and $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\frac{1-b}{1-a}$ are bounded. Consider the Laurent expansion of $\frac{1-b}{1-a}=\frac{1}{2}+O\left(\tilde{\lambda}^{-1}\right)$ by Lemma D.19, which is analytic at ∞ (being a concatenation, sum, and division of analytic functions). By Lemma D.16, we then obtain $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\frac{1-b}{1-a}=O\left(\tilde{\lambda}^{-2}\right)$. Since both terms are continuous on the region $\tilde{\lambda}\geq 0$, Lemma D.17 implies that both terms are bounded.

Consider the Laurent expansion of $f=O\left(\tilde{\lambda}^{-2}\right)$ from Lemma D.20. As above, it is analytic at ∞. By Lemma D.16, $\frac{\mathrm{d}f}{\mathrm{d}\tilde{\lambda}}=O\left(\tilde{\lambda}^{-3}\right)$. Since both $f,\frac{\mathrm{d}f}{\mathrm{d}\tilde{\lambda}}$ are continuous at the region where $\tilde{\lambda}\geq 0$ by Lemma D.17 both terms are bounded.

Otherwise, if $\alpha=1$, it is not clear that $f,\frac{1-b}{1-a},a,b$ and their corresponding derivatives are continuous at the region $\tilde{\lambda}\geq 0$, as they may diverge at $\tilde{\lambda}=0$. From Lemmas D.11, D.13 and D.15 we know that $a^4\frac{\mathrm{d}f}{\mathrm{d}\tilde{\lambda}},a^2f,a^2\frac{\mathrm{d}a}{\mathrm{d}\tilde{\lambda}},b^3\frac{\mathrm{d}b}{\mathrm{d}\tilde{\lambda}},\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\frac{1-b}{1-a},\frac{1-b}{1-a}$ are bounded at the region $\tilde{\lambda}>0$ in that case.

Combining with Lemma D.9 if $\alpha<1$, otherwise if $\alpha=1$ we obtain that each prefactor multiplying the terms involving a^{T-5} or b^{T-5} in Eq. (41) is bounded, so there exist constants $C_1,C_2>0$ such that,

$$\left|\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\left[\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right]-\mathbb{E}\left[G_T\right]\right]\right|\leq\frac{C_1Ta^{T-5}}{\left(1-a\right)^2}+C_2Tb^{T-5}\leq\left(C_1+C_2\right)\frac{T\zeta^{T-5}}{\left(1-a\right)^2},$$

where we used Lemma D.2.

Furthermore,

$$\sup_{\lambda\in[0,\Lambda(T)]}\left|\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\left[\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right]-\mathbb{E}\left[G_T\right]\right]\right|\leq\left(C_1+C_2\right)T\zeta^{T-5}O\left(\Lambda(T)^2\right), \tag{42}$$

where the last inequality follows from Lemma D.7.

Proving $\lambda^\star=\omega\left(T^\rho\right),\forall\rho\in(0,1)$. First, we show that the optimal regularizer $\lambda^\star$ is not in $o\left(T^\rho\right)$ for all $0<\rho<1$ as stated in the theorem. Next, we ensure that $\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathbb{E}\left[G_T\right]$ remains negative when $\Lambda\left(T\right)\in o\left(T^\rho\right)$ for arbitrary $0<\rho<1$. It suffices to show that the gap between the derivatives (Eq. (42)) converges to zero uniformly in $\tilde{\lambda}$ faster than $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right]$ does.

Pick $\Lambda\left(T\right)\in o\left(T^\rho\right)$ for some $\alpha<1$. From Corollary 1, we have

$$\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right]=\underbrace{2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{1-b}{1-a}\right)}_{\text{term 1}}+\underbrace{\frac{v_z}{2v_x}\left(\frac{v_x\left(1+\alpha\right)+\lambda}{\sqrt{\lambda^2+2\lambda v_x\left(1+\alpha\right)+v_x^2\left(1-\alpha\right)^2}}-1\right)\frac{1}{1-a}}_{\text{term 2}}.$$

Recall Theorem 7 where we proved that the derivative of both terms are negative for all $\lambda > 0$.

We aim to establish a lower bound for $\inf_{\tilde{\lambda}\in[0,\Lambda(T)]}\left(-\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right]\right)$, providing a uniform rate of decrease for $\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right]$. If either $\mathrm{Tr}(\mathbf{\Sigma}) > 0$ or $v_z > 0$ then, by Lemma E.3, since $\tilde{\lambda} = \frac{\lambda}{v_x} \in [0, \frac{1}{v_x}\Lambda(T)]$, we have,

$$\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right] \in \Omega\left(\left(\Lambda\left(T\right)\right)^{-2}\right) = \omega(T^{-2\rho})\,.$$

Let $\eta = \inf_{\tilde{\lambda}\in[0,\Lambda(T)]}\left(-\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right]\right) = \omega\left(T^{-2\rho}\right)$ and $\bar{\lambda} = \inf_{\tilde{\lambda}\in[0,\Lambda(T)]}\left(\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\mathbb{E}\left[G_T\right]\right)$. Then,

$$\frac{\mathrm{d}}{\mathrm{d}\bar{\lambda}}\mathbb{E}\left[G_T\right] \le \frac{\mathrm{d}}{\mathrm{d}\bar{\lambda}}\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right] + \left|\frac{\mathrm{d}}{\mathrm{d}\bar{\lambda}}\left[\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right] - \mathbb{E}\left[G_T\right]\right]\right| \le -\eta + \frac{1}{2}\eta \le -\frac{1}{2}\eta < 0,$$

where the second inequality follows from the uniform bound,

$$\sup_{\tilde{\lambda}\in[0,\Lambda(T)]}\left|\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right] - \mathbb{E}\left[G_T\right]\right]\right| \le \frac{1}{2}\eta\,,$$

since the exponential factor

$$\zeta^{T-5} = \left(1 - \frac{\alpha}{\Lambda\left(T\right) + 1 + \alpha}\right)^{T-5} \le \exp\left(-\frac{\alpha\left(T-5\right)}{\Lambda\left(T\right) + 1 + \alpha}\right)$$

decays faster than any polynomial in T. In particular, the factor $\frac{\alpha(T-5)}{\Lambda(T)+1+\alpha} \in \omega\left(T^{1-\rho}\right)$.

In superlinear regimes, the derivative turns positive. Next, we analyze the case of $\tilde{\lambda} \in \omega\left(T\right)$, and later, we show that $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\mathbb{E}\left[G_T\right] > 0$ which completes our proof.

Recall Eq. (39),

$$\begin{aligned}
&\mathbb{E}\left[G_T\right] - \lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right] \\
&= \underbrace{a^T\left(\|\mathbf{w}^\star\|^2 + \mathrm{Tr}\left(\mathbf{\Sigma}\right)\right) + \frac{2\mathrm{Tr}\left(\mathbf{\Sigma}\right)}{T}\left(2b^T - 1\right)}_{\text{term 1}} - \underbrace{2\mathrm{Tr}\left(\mathbf{\Sigma}\right)a^T\left(\frac{1-b}{1-a}\right)}_{\text{term 2}} + \underbrace{f\frac{-a^T}{1-a}}_{\text{term 3}}\,.
\end{aligned}$$

Using Lemma D.18 we adopt a Laurent expansion for the expression $\mathbb{E}\left[G_T\right] - \lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right]$, term by term, with this expression we intend to show that $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\mathbb{E}\left[G_T\right] > 0$ when $\tilde{\lambda} = \omega(T)$. Here, we analyze the asymptotic case where $T \to \infty$ and $\tilde{\lambda} = \omega(T)$, and thus, the $O(\cdot)$ notations below hide all bounded variables, but not $T, \tilde{\lambda}$. Term 1 becomes,

$$\begin{aligned}
&a^T\left(\|\mathbf{w}^\star\|^2 + \mathrm{Tr}\left(\mathbf{\Sigma}\right)\right) + \frac{2\mathrm{Tr}\left(\mathbf{\Sigma}\right)}{T}\left(2b^T - 1\right) \\
&= \left(1 - \frac{2\alpha T}{\tilde{\lambda}} + O\left(T^2\tilde{\lambda}^{-2}\right)\right)\left(\|\mathbf{w}^\star\|^2 + \mathrm{Tr}\left(\mathbf{\Sigma}\right)\right) + \frac{2\mathrm{Tr}\left(\mathbf{\Sigma}\right)}{T}\left(1 - \frac{2\alpha T}{\tilde{\lambda}} + O\left(T^2\tilde{\lambda}^{-2}\right)\right) \\
&= \|\mathbf{w}^\star\|^2 + \mathrm{Tr}\left(\mathbf{\Sigma}\right) - \frac{2\alpha T}{\tilde{\lambda}}\left(\|\mathbf{w}^\star\|^2 + \mathrm{Tr}\left(\mathbf{\Sigma}\right)\right) + \frac{2\mathrm{Tr}\left(\mathbf{\Sigma}\right)}{T} + O\left(T^2\tilde{\lambda}^{-2}\right) + O\left(\tilde{\lambda}^{-1}\right)\,.
\end{aligned}$$

Term 2 is,

$$
\begin{aligned}
2\mathrm{Tr}\left(\mathbf{\Sigma}\right) a^T \left(\frac{1-b}{1-a}\right) &= 2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(1-\frac{2\alpha T}{\tilde{\lambda}}+O\left(T^2\tilde{\lambda}^{-2}\right)\right)\frac{\frac{\alpha}{\tilde{\lambda}}+O\left(\tilde{\lambda}^{-2}\right)}{\frac{2\alpha}{\tilde{\lambda}}+O\left(\tilde{\lambda}^{-2}\right)} \\
\left[\text{Lemma D.19}\right] &= 2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(1-\frac{2\alpha T}{\tilde{\lambda}}+O\left(T^2\tilde{\lambda}^{-2}\right)\right)\left(\frac{1}{2}+O\left(\tilde{\lambda}^{-1}\right)\right) \\
&= 2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{1}{2}-\frac{\alpha T}{\tilde{\lambda}}+O\left(T^2\tilde{\lambda}^{-2}\right)+O\left(\tilde{\lambda}^{-1}\right)\right) \\
&= \mathrm{Tr}\left(\mathbf{\Sigma}\right)-2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\frac{\alpha T}{\tilde{\lambda}}+O\left(T^2\tilde{\lambda}^{-2}\right)+O\left(\tilde{\lambda}^{-1}\right).
\end{aligned}
$$

Finally, Term 3 is,

$$
\begin{aligned}
& \frac{v_z}{2v_x}\left(\frac{\tilde{\lambda}+\alpha+1}{\sqrt{\left(\tilde{\lambda}+\alpha+1\right)^2-4\alpha}}-1\right)\frac{-a^T}{1-a} \\
&= \frac{v_z}{2v_x}\left(\frac{1}{\sqrt{1-\frac{4\alpha}{\left(\tilde{\lambda}+\alpha+1\right)^2}}}-1\right)\frac{-a^T}{1-a} \\
\left[(1-u)^{-\frac{1}{2}}=1+\tfrac{1}{2}u+O\left(u^2\right)\right] &= v_z\frac{1}{2v_x}\left(1+\frac{2\alpha}{\left(\tilde{\lambda}+\alpha+1\right)^2}+O\left(\tilde{\lambda}^{-4}\right)-1\right)\frac{-a^T}{1-a} \\
&= O\left(\tilde{\lambda}^{-2}\right)\left(\frac{-1+\frac{2\alpha T}{\tilde{\lambda}}+O\left(T^2\tilde{\lambda}^{-2}\right)}{\frac{2\alpha}{\tilde{\lambda}}+O\left(\tilde{\lambda}^{-2}\right)}\right) \\
&= O\left(\tilde{\lambda}^{-2}\right)\left(-\frac{\tilde{\lambda}}{2\alpha}+T+O\left(T^2\tilde{\lambda}^{-1}\right)\right)=O\left(\tilde{\lambda}^{-1}\right).
\end{aligned}
$$

Overall, when summing,

$$
\mathbb{E}\left[G_T\right]-\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right]=\left\|\mathbf{w}^\star\right\|^2-\frac{2\alpha T}{\tilde{\lambda}}\left\|\mathbf{w}^\star\right\|^2+\frac{2\mathrm{Tr}\left(\mathbf{\Sigma}\right)}{T}+O\left(T^2\tilde{\lambda}^{-2}\right)+O\left(\tilde{\lambda}^{-1}\right).
$$

By Lemma D.16, as of the terms above is analytic,[11] and analyticity is preserved under concatenation, summing and multiplication and division,

$$\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\left[\mathbb{E}\left[G_T\right]-\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right]\right]=\frac{2\alpha T}{\tilde{\lambda}^2}\left\|\mathbf{w}^\star\right\|^2+O\left(T^2\tilde{\lambda}^{-3}\right)+O\left(\tilde{\lambda}^{-2}\right)\overset{[1]}{>}0$$
$$\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\mathbb{E}\left[G_T\right]=\frac{2\alpha T}{\tilde{\lambda}^2}\left\|\mathbf{w}^\star\right\|^2+O\left(T^2\tilde{\lambda}^{-3}\right)+O\left(\tilde{\lambda}^{-2}\right)+\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\lim_{T\to\infty}\mathbb{E}\left[G_T\right]\overset{[1]+[2]}{>}0\,,$$

where [1] follows from $\lim_{T\to\infty}\frac{2\alpha T\tilde{\lambda}^{-2}}{T^2\tilde{\lambda}^{-3}}=\lim_{T\to\infty}\frac{2\alpha\tilde{\lambda}}{T}\overset{\tilde{\lambda}=\omega(T)}{=}\infty$ and $\lim_{T\to\infty}\frac{2\alpha T\tilde{\lambda}^{-2}}{\tilde{\lambda}^{-2}}=\infty$ implying the sign is determined by $\frac{2\alpha T}{\tilde{\lambda}^2}>0$ as $\mathbf{w}^\star\neq\mathbf{0}$; and [2] follows since $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\lim_{T'\to\infty}\mathbb{E}\left[G_{T'}\right]=O\left(\tilde{\lambda}^{-2}\right)$ by Lemma E.3.

Conclusion. Since for all $\tilde{\lambda}\in o(T^\rho)$ (equivalently, $\lambda\in o(T^\rho)$) with $0<\rho<1$, we have $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\mathbb{E}[G_T]<0$, and for all $\tilde{\lambda}\in\omega(T)$ (equivalently, $\lambda\in\omega(T)$), $\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\mathbb{E}[G_T]>0$, it follows that any optimal regularizer $\lambda^\star$ must satisfy $\lambda^\star\in O(T)$ and $\lambda^\star\in\Omega(T^\rho)$ for all $0<\rho<1$. ■

11. An analytic function is a function locally given by a convergent power series.

E.6. Proof of Theorem 6: Key Result on Optimal Regularization Scaling

Recall Theorem 6. Under i.i.d. teachers (Assumption 1) with non-zero mean teacher $\mathbf{w}^\star$, the optimal fixed regularization strength that minimizes the expected generalization loss after T iterations satisfies $\lambda^\star = \lambda^\star(T) \asymp \frac{T}{\ln T}$. More precisely, for any $\epsilon > 0$, there exists $T_0 = O\big(1/\epsilon^2\big)$ such that for all $T \geq T_0$,

$$(1-\epsilon)\frac{2v_x\alpha T}{\ln\left(\frac{4\alpha T\|\mathbf{w}^\star\|^2 v_x}{v_z+v_x\operatorname{Tr}(\mathbf{\Sigma})(1+\alpha)}\right)} < \lambda^\star < (1+\epsilon)\frac{2v_x\alpha T}{\ln\left(\frac{4\alpha T\|\mathbf{w}^\star\|^2 v_x}{v_z+v_x\operatorname{Tr}(\mathbf{\Sigma})(1+\alpha)}\right)}.$$

Furthermore, in the degenerate noiseless case of $\operatorname{Tr}(\mathbf{\Sigma}) = v_z = 0$, it holds that $\lambda^\star \to 0$.

Proof. If both $\operatorname{Tr}(\mathbf{\Sigma}) = 0$ and $v_z = 0$, then by Eq. (32),

$$\mathbb{E}[G_T] = \mathbb{E}\left\|\mathbf{w}_T - \mathbf{w}^\star\right\|^2 = a^T\left\|\mathbf{w}_0 - \mathbf{w}^\star\right\|^2.$$

As a function of λ, the derivative of a is positive for all $\lambda > 0$ (Lemma D.3). Therefore, the optimal regularizer is $\lambda^\star \to 0$.

Otherwise, let $\lambda = \psi(T)\,T$, where $\psi(T) \in \omega\left(T^{-1/4}\right)$ and $\psi(T) \in O(1)$. By Lemma E.4, this entire region contains the optimal regularizer $\lambda^\star$ for all sufficiently large T. Recall Eq. (39),

$$\mathbb{E}\left[G_T\right] - \lim_{T\to\infty}\mathbb{E}\left[G_T\right] = \underbrace{\frac{2\operatorname{Tr}\left(\mathbf{\Sigma}\right)}{T}\left(2b^T-1\right)}_{\text{term 1}} + \underbrace{\operatorname{Tr}\left(\mathbf{\Sigma}\right)a^T\left(1-2\frac{1-b}{1-a}\right)}_{\text{term 2}} + \underbrace{a^T\left\|\mathbf{w}^\star\right\|^2 + f\frac{-a^T}{1-a}}_{\text{term 3}},$$

Using Lemma D.18, we begin with developing an equivalent Laurent expansion of each one of the terms. Term 1 becomes,

$$\begin{aligned}\frac{2\operatorname{Tr}(\mathbf{\Sigma})}{T}\left(2b^T-1\right) &= \frac{2\operatorname{Tr}(\mathbf{\Sigma})}{T}\left(2\exp\left(-\frac{\alpha T}{\tilde{\lambda}}\right)\left(1+O(T\tilde{\lambda}^{-2})\right)-1\right)\\ &= \frac{2\operatorname{Tr}(\mathbf{\Sigma})}{T}\left(2\exp\left(-\frac{\alpha v_x}{\psi(T)}\right)\left(1+O\left(T^{-1}\psi(T)^{-2}\right)\right)-1\right).\end{aligned} \tag{43}$$

Term 2 is,

$$\begin{aligned}&\operatorname{Tr}\left(\mathbf{\Sigma}\right)a^T\left[1-2\left(\frac{1-b}{1-a}\right)\right]\\ &= \operatorname{Tr}\left(\mathbf{\Sigma}\right)\left(\exp\left(-\frac{2\alpha v_x}{\psi(T)}\right)\left(1+O\left(\psi(T)^{-2}T^{-1}\right)\right)\right)O\left(\psi(T)^{-1}T^{-1}\right).\end{aligned} \tag{44}$$

Term 3 is,

$$
\begin{aligned}
&a^T\|\mathbf{w}^\star\|^2 + f\frac{-a^T}{1-a} \\
&= \exp\left(-\frac{2\alpha v_x}{\psi(T)} + O\left(\psi(T)^{-2}T^{-1}\right)\right)\|\mathbf{w}^\star\|^2 \\
&\quad - \tfrac{\psi(T)Tv_z}{4\alpha v_x^2}\left(\frac{2\alpha}{\left(\frac{\psi(T)T}{v_x}+\alpha+1\right)^2} + O\left((\psi(T)T)^{-4}\right)\right)\exp\left(-\tfrac{2\alpha v_x}{\psi(T)} + O\left(\psi(T)^{-2}T^{-1}\right)\right) \\
&= \exp\left(-\tfrac{2\alpha v_x}{\psi(T)}\right)\left(1 + O\left(\psi(T)^{-2}T^{-1}\right)\right)\|\mathbf{w}^\star\|^2 \\
&\quad - \tfrac{v_z}{2}\left(\frac{1}{\psi(T)T} + O\left((\psi(T)T)^{-2}\right)\right)\left(\exp\left(-\tfrac{2\alpha v_x}{\psi(T)}\right)\left(1 + O\left(\psi(T)^{-2}T^{-1}\right)\right)\right). \qquad (45)
\end{aligned}
$$

Using Lemma D.16, we differentiate $\frac{\mathrm{d}}{\mathrm{d}\psi(T)}\left[\mathbb{E}[G_T] - \lim_{T\to\infty}\mathbb{E}[G_T]\right]$, term by term.

We begin with Eq. (43),

$$
\frac{\mathrm{d}}{\mathrm{d}\psi(T)}\frac{2\,\mathrm{Tr}(\mathbf{\Sigma})}{T}\left(2b^T - 1\right) = \frac{2\,\mathrm{Tr}(\mathbf{\Sigma})}{T}\left(\frac{2\alpha v_x}{\psi(T)^2}\exp\left(-\frac{\alpha v_x}{\psi(T)}\right)\left(1 + O\left(\psi(T)^{-2}T^{-1}\right)\right)\right) + \frac{2\,\mathrm{Tr}(\mathbf{\Sigma})}{T}\left(2\exp\left(-\frac{\alpha v_x}{\psi(T)}\right)O\left(\psi(T)^{-3}T^{-1}\right)\right). \qquad (46)
$$

Differentiating the expression in Eq. (44), we get,

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\psi(T)}\mathrm{Tr}\left(\mathbf{\Sigma}\right)a^T\left[1 - 2\left(\frac{1-b}{1-a}\right)\right] \\
&= \mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{2\alpha v_x}{\psi(T)^2}\exp\left(-\frac{2\alpha v_x}{\psi(T)}\right)\left(1 + O\left(\psi(T)^{-2}T^{-1}\right)\right)\right. \\
&\qquad\qquad \left. + \exp\left(-\frac{2\alpha v_x}{\psi(T)}\right)O\left(\psi(T)^{-3}T^{-1}\right)\right)O\left(\psi(T)^{-1}T^{-1}\right) \\
&\quad + \mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(\exp\left(-\frac{2\alpha v_x}{\psi(T)}\right)\left(1 + O\left(\psi(T)^{-2}T^{-1}\right)\right)\right)O\left(\psi(T)^{-2}T^{-1}\right) \qquad (47)
\end{aligned}
$$

And Eq. (45) gives

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\psi\left(T\right)}\left[a^T\|\mathbf{w}^\star\|^2+f\frac{-a^T}{1-a}\right]\\
&=\tfrac{v_z}{2}\left(\frac{1}{\psi(T)^2T}+O\left(\psi(T)^{-3}T^{-2}\right)\right)\left(\exp\left(-\tfrac{2\alpha v_x}{\psi(T)}\right)\left(1+O\left(\psi(T)^{-2}T^{-1}\right)\right)\right)\\
&\qquad-\tfrac{v_z}{2}\left(\frac{1}{\psi(T)T}+O\left((\psi(T)T)^{-2}\right)\right)\left(\tfrac{2\alpha v_x}{\psi(T)^2}\exp\left(-\tfrac{2\alpha v_x}{\psi(T)}\right)\right.\\
&\qquad\qquad\left.\left(1+O\left(\psi(T)^{-2}T^{-1}\right)\right)+\exp\left(-\tfrac{2\alpha v_x}{\psi(T)}\right)O\left(\psi(T)^{-3}T^{-1}\right)\right)\\
&\quad+\frac{2\alpha v_x}{\psi(T)^2}\exp\left(-\frac{2\alpha v_x}{\psi(T)}\right)\left(1+O\left(\psi(T)^{-2}T^{-1}\right)\right)\|\mathbf{w}^\star\|^2\\
&\quad+\exp\left(-\frac{2\alpha v_x}{\psi(T)}\right)O\left(\psi(T)^{-3}T^{-1}\right)\|\mathbf{w}^\star\|^2.
\end{aligned}\tag{48}$$

Retaining only the leading–order terms, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\psi(T)}[\mathbb{E}\left[G_T\right]-\lim_{T\to\infty}\mathbb{E}\left[G_T\right]]\\
&=\frac{2\alpha v_x}{\psi(T)^2}\exp\left(-q\right)\|\mathbf{w}^\star\|^2+\frac{4\,\mathrm{Tr}(\mathbf{\Sigma})}{T}\frac{\alpha v_x}{\psi(T)^2}\exp\left(-\frac{q}{2}\right),
\end{aligned}$$

where $q=\frac{2\alpha v_x}{\psi(T)}$.

Let

$$g\left(\psi(T)\right)=\frac{v_zv_x}{2T\psi(T)^2}\left[\underbrace{\exp\left(-q\right)\frac{4\alpha T\left\|\mathbf{w}^\star\right\|^2}{v_z}+\frac{8\alpha\mathrm{Tr}(\mathbf{\Sigma})}{v_z}\sqrt{\exp\left(-q\right)}}_{H(q)}\right],\tag{49}$$

Recall Lemma E.3 where we have shown that

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\lim_{T\to\infty}\mathbb{E}\left[G_T\right]=-\frac{1}{2v_x}\left(\frac{v_z}{v_x}+\mathrm{Tr}(\mathbf{\Sigma})\left(\alpha+1\right)\right)\tilde{\lambda}^{-2}+O\left(\tilde{\lambda}^{-3}\right).$$

We would like to show that as a consequence $g\left(\psi(T)\right)-\frac{\mathrm{d}}{\mathrm{d}\psi(T)}\lim_{T\to\infty}\mathbb{E}\left[G_T\right]$ has a single root. The term $\frac{\mathrm{d}}{\mathrm{d}\psi(T)}\lim_{T\to\infty}\mathbb{E}[G_T]$ has a contribution of $g=-\left(\frac{1}{v_x}+\mathrm{Tr}(\mathbf{\Sigma})\frac{\alpha+1}{v_z}\right)$ (retaining the leading term) to $H(q)$, therefore we define $\tilde{H}(q)=H(q)+g$.

The derivative of $\tilde{H}$ is given by $\frac{\mathrm{d}}{\mathrm{d}q}\tilde{H}=-\exp(-q)\frac{4\alpha T\|\mathbf{w}^\star\|^2}{v_z}-\frac{4\alpha\mathrm{Tr}(\mathbf{\Sigma})}{v_z}\exp(-\frac{q}{2})$, which is strictly negative. Consequently, $\tilde{H}(q)$ is strictly decreasing in this regime. Moreover, if $q\in o\left(\ln\left(T\right)\right)$ then $\tilde{H}(q)=\exp(-q)\frac{4\alpha T\|\mathbf{w}^\star\|^2}{v_z}+g+o(1)$ which is positive for all sufficiently large T. Furthermore, if $q\in\omega(\ln\left(T\right))$ then $\tilde{H}(q)=g+o(1)$ hence for all sufficiently large T, we have that $\tilde{H}(q)$ is negative. Therefore, it suffices to identify a value $q^\star$ such that $\tilde{H}\left(q^\star\right)\approx 0$. To this

end, we solve $H(q^\star) \approx -g$ by retaining the leading-order term for $q^\star \in \Theta(\ln(T))$ we evaluate $\frac{4\alpha T\|\mathbf{w}^\star\|^2 \exp(-q^\star)}{v_z} \approx -g$ which leaves us with $q^\star \approx \ln \frac{4\alpha T\|\mathbf{w}^\star\|^2}{-v_z g}$. We proceed to prove the Theorem using the approximation discussed above.

Let $\epsilon > 0$,

$$\tilde{H}\left(\frac{1}{1+\epsilon}q^\star\right) = -g \cdot \left(\frac{4\alpha T\|\mathbf{w}^\star\|^2}{-v_z g}\right)^{\frac{\epsilon}{1+\epsilon}} + g + o(1) \xrightarrow[T\to\infty]{} +\infty, \tag{50}$$

and hence $\tilde{H}\left(\frac{1}{1+\epsilon}q^\star\right) > 0$ for all sufficiently large T. Similarly,

$$\tilde{H}\left(\frac{1}{1-\epsilon}q^\star\right) = -g \cdot \left(\frac{4\alpha T\|\mathbf{w}^\star\|^2}{-v_z g}\right)^{-\frac{\epsilon}{1-\epsilon}} + g + o(1) \xrightarrow[T\to\infty]{} g < 0, \tag{51}$$

so $\tilde{H}\left(\frac{1}{1-\epsilon}q^\star\right) < 0$ for all sufficiently large T.

By continuity (and the strict monotonicity of $\tilde{H}$), there exists a unique solution satisfying $\tilde{H}(q^\star) = 0$. Converting back via $q = \frac{2\alpha v_x}{\psi(T)}$ yields

$$(1-\epsilon)\frac{2\alpha v_x T}{\ln \frac{4\alpha T\|\mathbf{w}^\star\|^2}{\frac{v_z}{v_x}+\mathrm{Tr}(\mathbf{\Sigma})(\alpha+1)}} < \lambda^\star < (1+\epsilon)\frac{2\alpha v_x T}{\ln \frac{4\alpha T\|\mathbf{w}^\star\|^2}{\frac{v_z}{v_x}+\mathrm{Tr}(\mathbf{\Sigma})(\alpha+1)}}.$$

Recall Eqs. (46) to (48), where the last term stems from the non-leading factor we discarded in the term $\frac{\mathrm{d}}{\mathrm{d}\psi(T)} \lim_{T\to\infty} \mathbb{E}[G_T]$. Since $\psi(T) = \Theta\left(\frac{1}{\ln T}\right)$ the non-leading terms contribute $o(1)$ perturbation to $H(q)$, which is absorbed into the $o(1)$ in Eqs. (50) and (51). We now investigate the asymptotic behavior under which the theorem holds for all $T > T_0$.

To satisfy $\tilde{H}\left(\frac{1}{1+\epsilon}q^\star\right) > 0$, we begin with Eq. (50). In this case, the non-leading terms contribute $O\left(T^{-\frac{1}{2(1+\epsilon)}}\right)$ (as $\exp(-q/2)$, the dominating term, in Eq. (49) becomes $O\left(T^{-\frac{1}{2(1+\epsilon)}}\right)$).

$$\begin{aligned}
&-g \cdot \left(\left(\frac{4\alpha T\|\mathbf{w}^\star\|^2}{-v_z g}\right)^{\frac{\epsilon}{1+\epsilon}} - 1\right) + O\left(T^{-\frac{1}{2(1+\epsilon)}}\right) \\
&= -g \cdot \left(\exp\left(\frac{\epsilon}{1+\epsilon}\ln\left(\frac{4\alpha T\|\mathbf{w}^\star\|^2}{-v_z g}\right)\right) - 1\right) + O\left(T^{-\frac{1}{2(1+\epsilon)}}\right) \\
[e^x-1\geq x] \quad &\geq -g \cdot \frac{\epsilon}{1+\epsilon}\ln\left(\frac{4\alpha T\|\mathbf{w}^\star\|^2}{-v_z g}\right) + O\left(T^{-\frac{1}{2(1+\epsilon)}}\right) \\
&\gtrsim -g \cdot \frac{\epsilon}{1+\epsilon}\ln(T) + O\left(T^{-\frac{1}{2(1+\epsilon)}}\right).
\end{aligned}$$

There exists a constant K_1 such that for all sufficiently large T we have $O\left(T^{-\frac{1}{2(1+\epsilon)}}\right) \leq K_1 T^{-\frac{1}{2(1+\epsilon)}}$, and hence we require, $-g\frac{\epsilon}{1+\epsilon}\ln T \gtrsim K_1 T^{-\frac{1}{2(1+\epsilon)}}$, which is satisfied if $T > T_0 = O\left(\frac{1}{\epsilon^2}\right)$ (as $\lim_{\epsilon\to 0^+} \frac{\epsilon \ln\frac{1}{\epsilon}}{\epsilon^{\frac{1}{1+\epsilon}}} = \infty$). Note that $\frac{4\alpha T\|\mathbf{w}^\star\|^2}{-v_z g} > 0$. We proceed with Eq. (51), in this case, the non-leading terms contribute $O\left(T^{-\frac{1}{2(1-\epsilon)}}\right)$ (as $\exp(-q/2)$, the dominating term, in Eq. (49) becomes $O\left(T^{-\frac{1}{2(1-\epsilon)}}\right)$). Splitting it to cases.

Case (1) When $\frac{\epsilon}{1-\epsilon}\ln\left(\frac{4\alpha T\|\mathbf{w}^\star\|^2}{-v_z g}\right)\in(0,1)$,

$$
\begin{aligned}
&-g\cdot\left(\exp\left(-\frac{\epsilon}{1-\epsilon}\ln\left(\frac{4\alpha T\left\|\mathbf{w}^\star\right\|^2}{-v_z g}\right)\right)-1\right)+O\left(T^{-\frac{1}{2(1-\epsilon)}}\right)\\
\left[e^{-x}-1\le-\tfrac{1}{2}x,\ \forall x\in(0,1)\right]\ &\le g\cdot\frac{\epsilon}{2\left(1-\epsilon\right)}\ln\left(\frac{4\alpha T\left\|\mathbf{w}^\star\right\|^2}{-v_z g}\right)+O\left(T^{-\frac{1}{2(1-\epsilon)}}\right)\\
&\lesssim g\cdot\frac{\epsilon}{1-\epsilon}\ln\left(T\right)+O\left(T^{-\frac{1}{2(1-\epsilon)}}\right).
\end{aligned}
$$

As before, there exists a constant K_2 such that for all sufficiently large T we have $O\left(T^{-\frac{1}{2(1-\epsilon)}}\right)\le K_2\,T^{-\frac{1}{2(1-\epsilon)}}$, and hence we require $-g\frac{\epsilon}{1-\epsilon}\ln T\gtrsim K_2\,T^{-\frac{1}{2(1-\epsilon)}}$. Although this equation does not admit a closed-form solution for T, choosing T_0 such that $T>T_0=O\left(\frac{1}{\epsilon^2}\right)$ (as $\lim\limits_{\epsilon\to0^+}\frac{\epsilon\ln\frac{1}{\epsilon}}{\epsilon^{\frac{1}{1-\epsilon}}}=\infty$), is sufficient.

Case (2) When $\frac{\epsilon}{1-\epsilon}\ln\left(\frac{4\alpha T\|\mathbf{w}^\star\|^2}{-v_z g}\right)\in[1,\infty)$,

$$
\begin{aligned}
&-g\left(\exp\left(-\frac{\epsilon}{1-\epsilon}\ln\left(\frac{4\alpha T\left\|\mathbf{w}^\star\right\|^2}{-v_z g}\right)\right)-1\right)+O\left(T^{-\frac{1}{2(1-\epsilon)}}\right)\\
&\le -g\left(e^{-1}-1\right)+O\left(T^{-\frac{1}{2(1-\epsilon)}}\right).
\end{aligned}
$$

Therefore, we require,

$$
-g\left(e^{-1}-1\right)\gtrsim K_2T^{-\frac{1}{2(1-\epsilon)}},
$$

which holds for all sufficiently large T independent of ϵ. ■