OPTIMAL EXPLICIT AND IMPLICIT REGULARIZATION STRENGTHS IN HIGH-DIMENSIONAL CONTINUAL LINEAR REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study continual linear regression in a high-dimensional regime, where a learner observes a sequence of tasks and updates its parameters sequentially. In this setting, we derive an explicit closed-form expression for the generalization loss of a model trained with gradient descent for a fixed number of steps per task, which holds for arbitrary teachers. We prove that the optimal fixed step size decays nearly inversely with the number of tasks, specifically as $\ln T/T$, for a single teacher and, more generally, i.i.d. teachers. Under this scaling, we show that the resulting generalization loss is asymptotically equivalent to that achieved by optimally tuned L2 (isotropic) regularization. Furthermore, we extend our analysis to generalized ridge regression with diagonal regularization matrices, obtaining a closed-form characterization of the generalization loss and identifying necessary and sufficient conditions under which the optimal regularizer is isotropic. Finally, we identify a regime where the optimal step size remains finite even as the number of tasks tends to infinity, thus revealing a non-trivial tradeoff between task variability and implicit regularization, in contrast to the i.i.d. teacher scaling. Our theoretical predictions are corroborated by experiments on both linear models and neural networks, providing practical guidance for the design of continual learning systems.

1 INTRODUCTION

In continual learning, the learner is expected to learn sequentially from data, without 'forgetting' previous tasks. In recent years, driven by the remarkable success of artificial intelligence (AI) systems (Vaswani et al., 2017; Saharia et al., 2022; Rombach et al., 2022), continual learning has attracted significant attention, as most modern models are static in nature and do not adapt to newly arriving data. Unlike conventional machine learning models, continual learning involves learning from dynamic data distributions and faces a challenge known as catastrophic forgetting (McCloskey & Cohen, 1989; French, 1999), where adaptation to new distributions often leads to a substantial loss of performance on previously learned ones.

Many continual learning works aimed to reduce forgetting by *explicitly* penalizing changes in parameter space using per-parameter regularization terms (e.g., Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018; Chaudhry et al., 2018). Others (Lubana et al., 2022; Smith et al., 2023) showed empirically that isotropic L2 regularization results in similar performance. Practically, it is also common to minimize the current task's *unregularized* loss with a gradient algorithm for a *finite* number of steps (in contrast to theoretically learning to convergence). This cheaper alternative *implicitly* regularizes the model, even in stationary settings (Ali et al., 2019). To better understand catastrophic forgetting and how to mitigate it, most theoretical papers focus on a simplified setting, such as continual linear regression. This simplified yet analytically tractable setting captures essential features of sequential learning and has proven useful for analyzing diverse continual learning systems (Li et al., 2023; Hiratani, 2024; Levinstein et al., 2025; Evron et al., 2026).

Existing optimization-focused work (Evron et al., 2022; 2026; Levinstein et al., 2025; Tsipory et al., 2025) had to make a limiting assumption of joint realizability—i.e., that training data can be perfectly fitted by a single noiseless 'teacher' model. In contrast, we adopt a *statistical* framework, focusing on the generalization loss. Such analysis allows adopting a closed-form characterization of the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

generalization loss that naturally accommodates multiple teachers and label noise, thereby better capturing real-world scenarios. Moreover, our framework yields *exact* generalization expressions, whereas optimization-focused works only allow worst-case upper and lower bounds. Such expressions better align with empirical experiments, highlighting the explanatory role of this framework (Lin et al., 2023; Karpel et al., 2026). Building on this framework, in this paper we investigate several remaining mysteries in continual learning regression. Our contributions are the following:

- We derive an exact closed-form generalization expression for early-stopped gradient descent that accommodates arbitrary teachers and label noise, and show that the optimal step decays as $\ln T/T$, for a single teacher and, more generally, i.i.d. teachers.
- We identify a teacher regime under which the optimal step size need not scale with T.
- We derive the exact generalization loss under task-dependent diagonal regularization in the single-teacher setting, and characterize when isotropic regularization remains optimal.
- We empirically validate our theoretical findings on continual linear models and neural networks, using both synthetic and `MNIST`-based data.

2 Setting: regularized continual learning with implicit or explicit regularization

Learning scheme. We study the continual linear regression setting, which has been widely studied in recent years (Evron et al., 2022; 2026; Zhao et al., 2024; Zhu et al., 2025; Levinstein et al., 2025).

In our setting the learner is exposed to T sequential tasks, that is, $(\mathbf{X}_1, \mathbf{y}_1), (\mathbf{X}_2, \mathbf{y}_2), \ldots, (\mathbf{X}_T, \mathbf{y}_T)$, where each $\mathbf{X}_t \in \mathbb{R}^{n\times d}, \mathbf{y}_t \in \mathbb{R}^n$. During task t, the learner observes only the data $(\mathbf{X}_t, \mathbf{y}_t)$ associated with that task and updates its predictor accordingly, without having access to data from *any* other task. To be capable of learning sequentially, many continual learning algorithms incorporate a regularization towards the last iterate. A prominent line of work proposes using the Fisher information to weight this regularization (Kirkpatrick et al., 2017; Benzing, 2022). However, subsequent studies have observed that a much simpler isotropic regularization performs comparably well in practice (Hsu et al., 2018; Lubana et al., 2022). Following recent theoretical works (Levinstein et al., 2025; Zhao et al., 2025) we focus on L2 (isotropic) regularization.

Algorithm 1 Continual linear regression with L2 regularization

Input: $\{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^T$, regularization strength λ
Initialize $\mathbf{w}_0 \in \mathbb{R}^d$
For each iteration $t = 1, \ldots, T$:
$\quad \mathbf{w}_t \leftarrow \arg\min_{\mathbf{w}} \left\{ \frac{1}{2n}\|\mathbf{X}_t\mathbf{w} - \mathbf{y}_t\|^2 + \frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}_{t-1}\|^2 \right\}$
Output: $\mathbf{w}_T$

Employing explicit regularization may become *impractical* in high-dimensional settings or over long task horizons, as it incurs a high computational cost (as we have to fully minimize the objective in 1). For this reason, we also study a gradient algorithm run for a *finite* number of steps, which naturally induces a form of implicit regularization (Ali et al., 2019; Sonthalia et al., 2024), while maintaining a relative small cost. In line with recent theoretical works (Ding et al., 2024; Levinstein et al., 2025), we will mainly focus on this scheme and its relation to the optimal regularized L2 (isotropic) solution.

Algorithm 2 Continual linear regression with finite step budgets (Early-stopped GD)

Input: $\{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^T$, inner step count N, and step size γ
Initialize $\mathbf{w}_0 \in \mathbb{R}^d$
For each task $t = 1, \ldots, T$:
$\quad$ Initialize $\mathbf{w}^{(0)} \leftarrow \mathbf{w}_{t-1}$
$\quad$ For $s = 1, \ldots, N$:
$\quad\quad \mathbf{w}^{(s)} \leftarrow \mathbf{w}^{(s-1)} - \gamma \nabla \frac{1}{2n} \left\|\mathbf{X}_t\mathbf{w}^{(s-1)} - \mathbf{y}_t\right\|^2$
$\quad \mathbf{w}_t \leftarrow \mathbf{w}^{(N)}$
Output $\mathbf{w}_T$

A widely used alternative that has been popular in recent empirical and theoretical works (Kirkpatrick et al., 2017; Schwarz et al., 2018; Zhao et al., 2024) considers continual learning through a quadratic regularization that penalizes deviations from the previous iterate as $\|\mathbf{w} - \mathbf{w}_{t-1}\|_{\mathbf{M}_t}^2 = (\mathbf{w} - \mathbf{w}_{t-1})^\top \mathbf{M}_t (\mathbf{w} - \mathbf{w}_{t-1})$, where $\mathbf{M}_t \succeq 0$ is a regularization matrix. This formulation generalizes the L2 regularization in Algorithm 1, which is recovered by setting $\mathbf{M}_t = \lambda \mathbf{I}$, it also encompasses regularization schemes such as Fisher-information methods.

Algorithm 3 Generalized L2 continual learning scheme

Input: $\{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^T$, regularization matrices $\{\mathbf{M}_t\}_{t=1}^T$
Initialize $\mathbf{w}_0 \in \mathbb{R}^d$
For each iteration $t = 1, \dots, T$:
$\mathbf{w}_t \leftarrow \arg\min_{\mathbf{w}} \left\{ \frac{1}{2n} \|\mathbf{X}_t \mathbf{w} - \mathbf{y}_t\|^2 + \frac{1}{2} \|\mathbf{w} - \mathbf{w}_{t-1}\|_{\mathbf{M}_t}^2 \right\}$
Output: $\mathbf{w}_T$

In practice, due to memory and computational constraints, continual learning systems typically employ diagonal regularization matrices to selectively penalize deviations from previously learned predictors (Kirkpatrick et al., 2017; Schwarz et al., 2018; Aljundi et al., 2018; Chaudhry et al., 2018), rather than maintaining a full $d \times d$ regularization matrix. Indeed, a full matrix can, in principle, eliminate forgetting entirely; see Evron et al. (2023, Proposition 5.5).

3 Setup: High Dimensional Data

Data model. In this paper, we consider a high-dimensional random design with independent feature matrices $\mathbf{X}_1, \dots, \mathbf{X}_T$, where each $\mathbf{X}_t \in \mathbb{R}^{n \times d}$ has i.i.d. entries with zero mean, and variance $v_x > 0$ with finite $4 + \epsilon$ moment for some $\epsilon > 0$, and an arbitrary initialization $\mathbf{w}_0 \triangleq \mathbf{w}_0^\star$. Moreover, the problem dimensions grow to infinity at a proportional rate $\alpha := \frac{n}{d} \in (0, \infty)$.

For each task $t \in [1, T] \triangleq \{1, \dots, T\}$, labels are generated from a corresponding teacher vector $\mathbf{w}_t^\star \in \mathbb{R}^d$ using 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, with $\mathbb{E}[\mathbf{z}_t] = 0$ and $\mathrm{Cov}(\mathbf{z}_t) = v_z \mathbf{I}$ for some noise variance $v_z \geq 0$, sampled independently of all tasks. To ensure a well-behaved high-dimensional limit, we assume that the teacher correlations converge to finite deterministic limits and are uniformly integrable, i.e.

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

Evaluation metrics. Following recent theoretical work in continual learning (Evron et al., 2022; Levinstein et al., 2025; Evron et al., 2026), we evaluate the model error using the mean-squared-error (MSE) loss, that is, $\frac{1}{n} \left\|\mathbf{X}_i \mathbf{w} - \mathbf{y}_i\right\|^2$, however, under our data model, averaging over independent test data and label noise, gives,

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

Thus, the dependence of the expected loss on the predictor $\mathbf{w}$ is entirely captured by the distance $\|\mathbf{w} - \mathbf{w}_i^\star\|$. Consequently, following recent works (Lin et al., 2023; Zhao et al., 2024; Karpel et al., 2026), we measure performance through the distance between the predictor $\mathbf{w}$ and the corresponding teacher $\mathbf{w}_i^\star$. This reduction allows us to formally define our metrics of interest.

Definition 1 (Task loss)**.** *For any predictor $\mathbf{w} \in \mathbb{R}^d$, the loss of the predictor in task $i \in [1, T]$ is defined as the distance between the predictor and the teacher of the associated task, that is,*

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

Definition 2 (Generalization error)**.** *At iteration t, the generalization error is defined as the average task loss of the predictor $\mathbf{w}_t \in \mathbb{R}^d$ over all tasks observed up to time t, that is,*

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

Notations. Throughout the paper, we use $\xrightarrow{P}$ for convergence in probability. We let G_T^{ridge}, G_T^{esgd}, and G_T^{gen} denote the generalization loss after T iterations by Algorithms 1, 2, and 3, respectively.

4 Early-stopped GD generalization loss and the optimal step-size

In this section, we derive an *exact* closed-form expression for the generalization loss attained by the early-stopped GD algorithm (Algorithm 2). Through this section, we consider only the step sizes that ensure the convergence of the early-stopped GD algorithm, that is, $\gamma < \frac{2}{\lambda_{\max}(\frac{1}{n}\mathbf{X}^\top\mathbf{X})}$, for generic $\mathbf{X}$ following the statistical setting described in Section 3. Our key technical result is the following.

Theorem 1 (Generalization loss)**.** *Recall the high-dimensional setting where the number of samples and the dimension grow at proportional rate $\alpha := \frac{n}{d}$. Moreover, recall that v_x, v_z are the variances of features and labels, respectively. Then, the generalization loss after T tasks of the early-stopped GD algorithm (Algorithm 2) satisfies, as $n, d \to \infty$,*

$$G_T^{\mathrm{esgd}} \xrightarrow{P} v_z c' \frac{1-a'^T}{1-a'} + \frac{1}{T}\sum_{i=1}^{T} \|\mathbf{w}_T^\star - \mathbf{w}_i^\star\|^2 + 2\big(\mathbf{w}_T^\star - \bar{\mathbf{w}}^\star\big)^\top \sum_{i=1}^{T} b'^{T-i+1}\big(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\big) + \sum_{i=1}^{T}\sum_{i'=1}^{T} a'^{T-\max(i,i')+1} b'^{|i-i'|}\big(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\big)^\top\big(\mathbf{w}_{i'-1}^\star - \mathbf{w}_{i'}^\star\big),$$

where we denote $\mathbf{w}_0^\star = \mathbf{w}_0$ and $\bar{\mathbf{w}}^\star = \frac{1}{T}\sum_i \mathbf{w}_i^\star$,

$$a' = 1 - 2Nv_x\gamma + O\big((N\gamma)^2\big), \qquad b' = 1 - Nv_x\gamma + O\big((N\gamma)^2\big), \qquad c' = \frac{N^2}{\alpha}v_x\gamma^2 + O\big((N\gamma)^3\big). \tag{2}$$

The proof of Theorem 1, along with the remaining proofs in this section, is provided in Appendix C, where we also give the full expressions for a', b', and c' (Eqs. (11), (14), (10), respectively). To aid interpretation, we include a numerical validation of the predictions of Theorem 1 in Appendix A.1.

Thus far, the exact generalization prediction in Theorem 1 has accommodated *arbitrary* teacher setup. To obtain more meaningful results, following prior works (Zhu et al., 2025; Karpel et al., 2026), we now consider teachers that are sampled i.i.d. from a common distribution.

Assumption 1 (i.i.d. teachers)**.** *Considering the statistical setting of Section 3, the teachers $\mathbf{w}_1^\star, \ldots, \mathbf{w}_T^\star$ are drawn i.i.d. from a distribution with a mean $\mathbf{w}^\star$ and covariance $\mathbf{\Sigma}$ where $\|\mathbf{\Sigma}\|_* \le C < \infty$, independently of all other random variables (features and labels).*

This assumption yet remains more permissive than theoretical works in continual learning which typically assume a single teacher (Li et al., 2023; Zhao et al., 2025) or joint realizability across *training* sets (Evron et al., 2022; 2026; Levinstein et al., 2025; Tsipory et al., 2025).

A theory-guided selection of the step size γ and the number of gradient steps N is crucial in settings where exhaustive tuning is impractical or infeasible, as in continual learning. This motivates the need for principled tuning rules. Such a rule follows from our next result, which we prove in Appendix C.3, by characterizing the optimal parameters that minimize the generalization loss of Theorem 1.

Theorem 2 (Optimal step size)**.** *Under the i.i.d. teacher setting (Assumption 1) with a nontrivial mean teacher $\mathbf{w}^\star \neq \mathbf{w}_0$[1], as $n, d \to \infty$ the optimal fixed step size $\gamma^\star$ that minimizes the generalization loss of the early-stopped GD algorithm (Algorithm 2) after T tasks satisfies $\gamma^\star \asymp \frac{\ln T}{NT}$. Specifically, for all sufficiently large $T > T_0 = \Theta\big(\frac{1}{\epsilon^2 \ln^2(1/\epsilon)}\big)$[2], the optimal parameters satisfy*

$$\frac{1}{1+\epsilon}\frac{\ln(T) + \ln(\Gamma)}{2Nv_xT} < \gamma^\star < \frac{1}{1-\epsilon}\frac{\ln(T) + \ln(\Gamma)}{2Nv_xT},$$

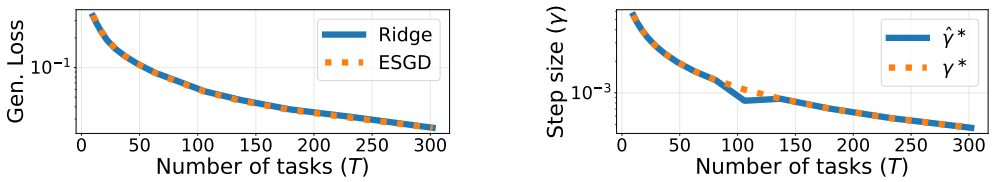
where $\Gamma = \frac{4\alpha v_x \|\mathbf{w}^\star - \mathbf{w}_0\|^2}{v_z + v_x \mathrm{Tr}(\mathbf{\Sigma})(\alpha+1)}$. In the degenerate noiseless case of $\mathrm{Tr}(\mathbf{\Sigma}) = v_z = 0$, the optimal choice is any $\gamma^\star \in \big(0, 2v_x^{-1}\big(1 + \alpha^{-1/2}\big)^{-2}\big)$ and $N \to \infty$.

[1] Having $\mathbf{w}^\star = \mathbf{w}_0$ would directly imply optimal step size $\gamma^\star \to 0$.

[2] We take the limits sequentially: first we take $n, d \to \infty$, and only then we let T grow.

Moreover, if $\gamma \asymp \frac{\ln T}{NT}$ the resulting generalization loss of the early-stopped GD matches that of optimally tuned L2 regularization up to $O(1/T)$ factor.

Experiments on Synthetic Regression Data. We empirically validate the theoretical predictions of Theorem 2 using finite dimensional data. For each task $i \in [1, T]$ we sample a teacher $\mathbf{w}_i^\star \sim \mathcal{N}(\mathbf{w}^\star, \frac{1}{d}\mathbf{I})$. We use $d = 80$ and $n = 50$ as the dimensions, and we sample $X_{i,j} \sim \mathcal{N}(0, 1)$. We fix the number of gradient steps as $N = 20$.



(a) Under optimal tuning, L2 regularization and early-stopped GD coincide as predicted by Theorem 2.

(b) Optimal step size strength decays as predicted by Theorem 2.

Figure 1: The plots provide empirical evidence for the Theorem 2. Here, we set $v_z = 1$; $\hat{\gamma}^\star$ denotes the empirically optimal value, while $\gamma^\star$ is the theoretical prediction by Theorem 2. We compute the empirical optimal $\hat{\gamma}^\star$, $\hat{\lambda}^\star$ with an exhaustive search. Empirical curves are averaged over 50 runs.

Up to this point, Theorem 2 yields a practical rule of thumb for choosing the number of gradient steps N and the step size γ in which we find that $\gamma^\star \asymp \frac{\ln T}{NT}$. In practice, this allows one to tune the parameter on tasks with smaller task horizons to identify this scaling, and then extrapolate it to larger task horizons, for which direct tuning would be computationally infeasible. For example, for $T = 1000$, one can tune at $T = 10$ and then set $\gamma(1000) = \frac{10\ln(1000)}{1000\ln(10)}\hat{\gamma}^\star(10) \approx 0.03\hat{\gamma}^\star(10)$. Moreover, under this choice, the resulting generalization loss remains close to the one attained with optimally tuned L2 regularization while achieving substantially lower computational cost. This result quantifies the close connection between early-stopped GD and optimally tuned isotropic regularization, which has been shown to perform comparably well to standard baselines in continual learning (Hsu et al., 2018; Smith et al., 2023). At the same time, many continual learning methods introduce more expressive diagonal-based regularization terms to selectively penalize deviations from previously learned predictors (Kirkpatrick et al., 2017; Schwarz et al., 2018; Aljundi et al., 2018), this naturally raises the question of when such simple isotropic regularization is sufficient. To capture this class of methods, we next study ridge regression with a diagonal regularization matrix.

5 Explicit Diagonal Regularization: When Is It Better?

In this section, we consider the generalized ridge regression algorithm (Algorithm 3) with arbitrary positive-definite diagonal regularization matrices $\{\mathbf{M}_t\}_{t=1}^T$ (spectrally bounded[3]). Importantly, this formulation captures the core structure of the algorithms that are widely used in practice (Kirkpatrick et al., 2017; Zenke et al., 2017; Chaudhry et al., 2018). Our goal is to derive an *exact* generalization analysis, under the following technical assumption.

Assumption 2 (Realizability)**.** *Considering the statistical setting of Section 3, we assume a single teacher across all tasks, that is,*

$$\mathbf{w}_1^\star = \mathbf{w}_2^\star = \ldots = \mathbf{w}_T^\star = \mathbf{w}^\star,$$

drawn independently of all random variables (features and labels).

The single-teacher assumption is standard in prior theoretical work (Li et al., 2023; Ding et al., 2024; Zhao et al., 2025). In our setting, it is primarily adopted for analytical simplicity rather than as a restriction, and our framework naturally extends to general (task-varying) teachers. This leads to the following closed-form characterization of the generalization loss.

[3]i.e., there exist constants $0 < l \leq u$ independent of n, d such that $l\mathbf{I} \preceq \mathbf{M}_t \preceq u\mathbf{I}$ for all $1 \leq t \leq T$.

Theorem 3 (Generalization loss)**.** *Under Assumption 2, the generalization loss after T tasks of the generalized ridge algorithm (Algorithm 3), specialized to a fixed regularization matrix* $\mathbf{M}$*, satisfies, as* $n, d \to \infty$*,*

$$G_T^{\text{gen}} \xrightarrow{P} (\mathbf{w}_0 - \mathbf{w}^\star)^\top \left(\bar{\mathbf{Q}}^{2T} + v_x \sum_{k=1}^{T} \bar{\mathbf{Q}}^{2k} \tilde{\mu} \tilde{\rho}_k \right) (\mathbf{w}_0 - \mathbf{w}^\star)$$

$$+ \frac{v_z}{n} \mathrm{Tr} \left[\sum_{k=1}^{T} \left[\mathbf{I} - \bar{\mathbf{Q}} \left[\mathbf{I} + \frac{v_x^2}{n} \mathrm{Tr} \left[\mathbf{M}^{-1} \bar{\mathbf{Q}}^2 \right] \tilde{\mu} \mathbf{M}^{-1} \right] \right] \left[\bar{\mathbf{Q}}^{2(T-k)} + v_x \sum_{i=k+1}^{T} \bar{\mathbf{Q}}^{2(i-k)} \tilde{\mu} \tilde{\rho}_i \right] \mathbf{M}^{-1} \bar{\mathbf{Q}} \right]$$

where $\bar{\mathbf{Q}} = \left(\mathbf{I} + v_x \tilde{m}_p(-1) \mathbf{M}^{-1} \right)^{-1}$*, with* $\tilde{m}_p(-1)$ *being the unique positive solution of*

$$\frac{1}{\tilde{m}_p(-1)} = 1 + \frac{1}{n} \sum_{i=1}^{d} \frac{v_x M_{i,i}^{-1}}{1 + \tilde{m}_p(-1) v_x M_{i,i}^{-1}},$$

and,

$$\tilde{\rho}_k = \frac{v_x}{n} \mathrm{Tr} \Big[\mathbf{M}^{-2} \Big(\bar{\mathbf{Q}}^{2(T-k+1)} + v_x \tilde{\mu} \sum_{i=k+1}^{T} \tilde{\rho}_i \bar{\mathbf{Q}}^{2(i-k+1)} \Big) \Big],$$

$$\tilde{\mu} = \left[\left[1 + \frac{v_x}{n} \mathrm{Tr} \left(\mathbf{M}^{-1} \bar{\mathbf{Q}} \right) \right]^2 - \frac{v_x^2}{n} \mathrm{Tr} \left(\mathbf{M}^{-1} \bar{\mathbf{Q}} \right)^2 \right]^{-1}.$$

The closed-form expression to task dependent diagonal regularization matrices is given in Equation (52), along with the remaining proofs of this section, which appear in Appendix D. To aid interpretation, we include a numerical validation of the predictions of Theorem 3 in Appendix A.2.

We now exhibit a simple setting in which scalar (isotropic) regularization provably fails.

Theorem 4 (Failure of isotropic ridge under sparse teachers)**.** *Assume that* $\alpha := \frac{n}{d} < 1$*. Let* $\mathbf{w}_1^\star - \mathbf{w}_0$ *be* k*-sparse with support* S *so that* $\mathbf{w}_1^\star \neq \mathbf{w}_0$*, and assume that* $\mathbf{X}_S$ *(the submatrix of* $\mathbf{X}$ *consisting of the columns indexed by* S*) has full column rank, then as* $n, d \to \infty$ *after a single task with no noise, it holds almost surely that,*

$$\sup_{\mathbf{M} \in \mathcal{M}} \inf_{\lambda > 0} \frac{G_1^{\text{ridge}}}{G_1^{\text{gen}}} = \infty.$$

where $\mathcal{M} = \bigcup_{0 < l \leq u < \infty} \{ \mathbf{M} : l \mathbf{I}_d \preceq \mathbf{M} \preceq u \mathbf{I}_d \}$ *and* u, l *are constants independent of* n, d*, i.e.* $\mathcal{M}$ *is the family of all spectrally bounded matrices (which may depend on the support* S*).*

The assumption that $\mathbf{X}_S$ has a full rank column is easy to satisfy: if k is fixed (or $k = o(d)$) and the entries $X_{i,j}$ are drawn from a continuous distribution, then almost surely $\mathbf{X}_S$ has a full rank column. Yet, it remains unclear under what conditions isotropic regularization remains optimal. Interestingly, in Appendix D.5 we find that the *necessary and sufficient* condition is that, for each coordinate, the teacher is equally far from the initialization, either deterministically, or, for a random teacher, with probability tending to 1 as $n, d \to \infty$ (sufficiency is straightforward, necessity is less trivial).

So far, our analysis characterizes how to optimally tune the parameters of early-stopped GD under the i.i.d. teacher setting, showing that the resulting generalization loss is asymptotically equivalent to that of optimally tuned isotropic regularization. Additionally, in the single teacher setting, our analysis identifies when such isotropic regularization remains optimal within the broader class of diagonal regularization schemes which are widely used in practice. Beyond the i.i.d. setting, a natural question is whether these conclusions persist under more structured, non-stationary task sequences.

6 Beyond i.i.d. teachers: optimal step-size need not scale with T

Finding the optimal parameter tuning in early-stopped GD had to make the simplifying assumption of i.i.d. teachers (Assumption 1). We now ask whether the optimal scaling identified by Theorem 2 which states that $\gamma^\star \asymp \frac{\ln T}{NT}$ persists beyond the i.i.d. teacher regime. We address it via a counterexample.

Assumption 3. *We assume a sinusoidal model across the teachers, that is, for all* $i \in [1, T]$*,*

$$\mathbf{w}_i^\star = \mathbf{A} + \sin(2\pi f i) \cdot \mathbf{B},$$

for some $\mathbf{A}, \mathbf{B} \in \mathbb{R}^d$ *and frequency* $f \in \mathbb{R}$*.*

This type of model is a fundamental building block for representing and analyzing periodic structure in learning (Vaswani et al., 2017), control systems (Astrom & Murray, 2008), and signal processing (Cooley & Tukey, 1965). In continual learning, recent works have been studying the effect of repetition. A few studies have proposed solutions that aim to exploit the repetition in the stream to learn continually (Liu et al., 2023). This has been demonstrated to significantly improve performance over standard task-split methods, emphasizing the effectiveness of cyclic strategies in continual learning settings.

Under the sinusoidal model (Assumption 3), the sequence of tasks is non-stationary, and the teachers do not concentrate around a single mean. As a result, the generalization measure as defined in Definition 2, which averages the loss with respect to all past teachers, is no longer informative.

Instead, we evaluate generalization with respect to teachers belonging to the same class (or phase) as the final task. Let $A \subseteq \{1, \dots, T\}$ denote the index set of such teachers. We define the generalization error at time T as

$$G_T \triangleq \frac{1}{|A|} \sum_{i \in A} \|\mathbf{w}_T - \mathbf{w}_i^\star\|^2 .$$

However, when the task horizon grows to infinity, the generalization error may be asymptotically oscillatory. We therefore analyze the *worst-case* generalization loss, defined as

$$G_\infty \triangleq \limsup_{T \to \infty} \frac{1}{|A_T|} \sum_{i \in A_T} \|\mathbf{w}_T - \mathbf{w}_i^\star\|^2 ,$$

where $A_T \subseteq \{1, \dots, T\}$ denotes the set of teachers belonging to the same class as the task at time T.

Theorem 5 (Generalization with infinite task horizon)**.** *Under Assumption 3, the generalization loss after T tasks of the early-stopped GD algorithm (Algorithm 2) satisfies, as $n, d \to \infty$,*

$$G_\infty \xrightarrow{P} C_0 + \sqrt{C_1^2 + C_2^2} \sup_{T \geq 1} \cos\left(4\pi f T + \varphi\right)$$

The proof, together with the explicit expressions for C_0, C_1, C_2 and φ, appears in Appendix E.1 (see Equation (61)). We further note that when f is irrational, then $\sup_{T \geq 1} \cos\left(4\pi f T + \varphi\right) = 1$.

We empirically evaluate the optimal step size under a sinusoidal model.

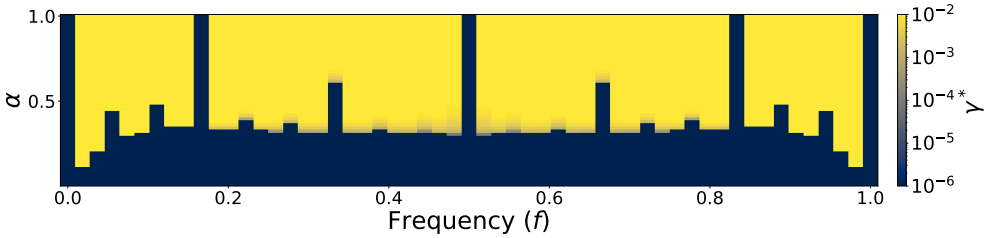
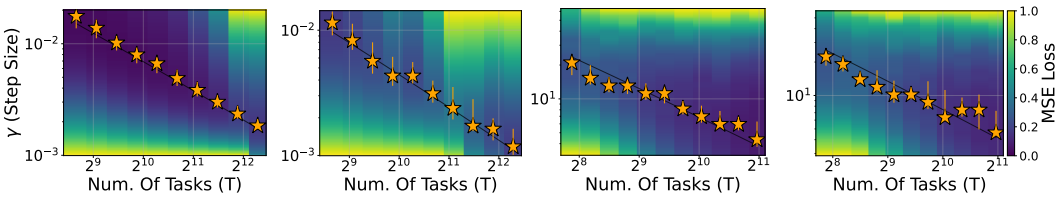


Figure 2: **The optimal step size remains finite as $T \to \infty$.** We plot the optimal step size $\gamma^\star$ of the objective G_∞ as a function of the ratio (α) and the frequency (f), fixing $\|\mathbf{B}\| = 1$, $v_z = 1$, $v_x = 1$, $d = 100$ and $N = 5$. We compute $\gamma^\star$ with an exhaustive search.

Under the i.i.d. teacher setting, our theoretical analysis (Theorem 2) shows that the optimal step size scales as $\ln T/T$, so we would expect $\gamma^\star \to 0$ in this regime. By contrast, as illustrated in Figure 2, for the sinusoidal teacher model the optimal step size also remains finite as $T \to \infty$.

7 Empirical experiments on MNIST-data

While our analysis assumes i.i.d. features and a linear model, we empirically validate our theoretical predictions on a finite-dimensional MNIST-data, rather than high-dimensional i.i.d. features, where the assumptions of Section 3 do not hold. We evaluate both linear models and ReLU networks. In all cases, we observe that the scaling $\gamma^\star \asymp \ln T/T$ as predicted by Theorem 2 is preserved.



(a) Single teacher (Linear) (b) i.i.d. teachers (Linear) (c) Single teacher (NN) (d) i.i.d. teachers (NN)

Figure 3: **Optimal step size scaling in linear models and neural networks (NN).** We plot the effect of step size on the mean squared error (MSE) after learning T tasks. Curves are averaged over 5 and 15 random seeds for the linear model and neural network, respectively. The star marks the minimizer of the mean curve; vertical bars indicate the 95% credible interval for the minimizer, and the black line indicates a fitted $\ln T/T$ scaling. The empirical MSE is normalized by the maximum MSE across all step sizes for the specific task horizon T.

Experimental Setup. We convert the `MNIST` data set labels into a binary classification problem where the digits 0-4 are converted to class 0 and 5-9 are converted to class 1. Moreover, we standardize the data to zero mean and unit variance.

Teacher models. In the experimental setup, we consider two types of teacher models.

1. **A single teacher.** We train a single teacher using the same architecture as the evaluated model—in the linear regression, the single teacher is trained over a subset of 500 samples (to fit a perfect linear teacher), and in the neural networks the teacher is trained until convergence over all `MNIST` data. In this case, each task suffers from label noise.
2. **i.i.d. noisy teachers.** Each task draws a teacher by adding Gaussian noise to the single teacher's weights. In addition to teacher noise, the labels also suffer from label noise.

Noisy Labels. Each task is affected by additive Gaussian noise with variance v_z applied to the binary labels. Specifically, we consider noise variances of $v_z = 0.1$ and, more challengingly, $v_z = 0.4$ for the linear and neural-network experiments, respectively.

Evaluation Metric. We evaluate generalization performance using the Mean Squared Error (MSE) computed over an independent test set.

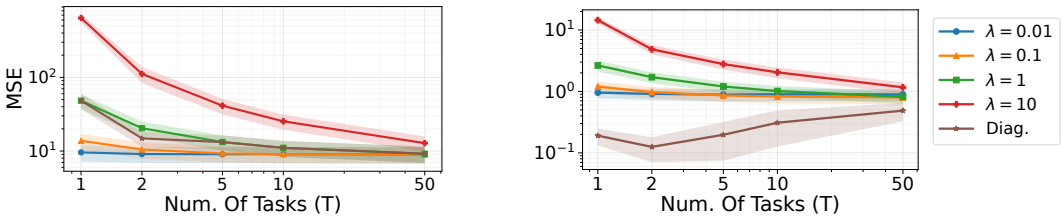
Linear Regression Experiments. We employ the early-stopped GD algorithm (Algorithm 2), trained by performing $N = 10$ gradient steps on the mean-squared error loss, to predict *noisy* labels from a single teacher (Figure 3a) or multiple i.i.d. teachers (Figure 3b). Evidently, the predicted $\ln T/T$ scaling is preserved, consistent with Theorem 2, even in the absence of i.i.d. features.

Neural Networks Experiments. We also consider a two-layer ReLU neural network ($784 \to 200 \to 1$), with a sigmoid output for the teacher and a linear output for the student (to predict noisy labels), trained by performing $N = 10$ gradient steps per task on the mean-squared error loss. Similarly, we repeat the previous experiments under both single-teacher (Figure 3c) and multiple-teachers (Figure 3d) settings. Notably, we again observe the same $\ln T/T$ scaling, which persists even in the substantially more complex NN model, matching the predicted scaling from Theorem 2.

7.1 Diagonal regularization experiments

We additionally experiment with the generalized L2 algorithm (Algorithm 3) in the linear regression setting. We consider two artificial teachers: a *sparse* teacher with 100 randomly selected coordinates equal to one and the remaining entries zero, and an *isotropic* teacher with all coordinates equal to one.

We use the optimal diagonal regularizer obtained by optimizing the theoretical objective in Theorem 3. Throughout this experiment, we set $v_z = 0.4$, initialize $\mathbf{w}_0 = \mathbf{0}$ and narrow the `MNIST` training data to 100 samples for each task. As observed, in the sparse-teacher setting the optimal diagonal regularizer substantially outperforms the scalar regularizer (Figure 4b), in agreement with the prediction of Theorem 4. In contrast, for an isotropic teacher (Figure 4a), the scalar regularizer is optimal, as anticipated by the condition found in Section 5, which requires the teacher vector to have coordinates equally far from the initialization.



(a) Isotropic teacher

(b) Sparse teacher

Figure 4: **Optimal isotropic regularization conditions.** We plot the mean MSE with varying regularization strengths and the optimal diagonal regularizer obtained by optimizing the theoretical objective in Theorem 3. Error bars denote the standard deviation across 50 random seeds.

8 Related work and discussion

Summary. In this paper, we studied continual linear regression under a simplified yet analytically tractable setting, assuming isotropic feature covariance, i.i.d. features, and either i.i.d. or single teacher. Despite these simplifying assumptions, our analysis yields meaningful theoretical insights that appear to persist beyond the linear regime. In particular, we identify a simple scaling law, $\gamma^\star \asymp \ln T/T$, which is supported both theoretically and empirically. This scaling provides a practical rule of thumb and significantly reduces the need for extensive hyperparameter tuning. Notably, we observe that this behavior extends beyond linear models, holding empirically even for neural networks. Next, we compare our results with existing work and discuss potential future directions.

Implicit regularization. Jung et al. (2025) study continual learning on multiple linear *classification* tasks, where each task is trained with gradient descent for a fixed budget of iterations per task. They consider both cyclic and random task ordering, for which they proved the convergence of the training loss to zero. For cyclic ordering, they provided the non-asymptotic rate convergence, whereas for random task ordering, they provided the asymptotic loss convergence. Levinstein et al. (2025) studied the optimal scaling of ridge regression and early-stopped GD in continual linear regression with arbitrary fixed features and a single noiseless teacher. Their analysis provides near-optimal *worst-case* convergence guarantees. In particular, they consider a fixed step-size $\gamma \in (0, 1/R^2)$ (where R is the data radius) and choose the iteration budget as $N = \frac{\ln\left(1-\frac{1}{\ln T}\right)}{\ln(1-\gamma R^2)}$, though it still remains unclear whether other values could perform better. Interestingly, in a similar single teacher noiseless setting our analysis dictates that the optimal number of gradient steps is $N \to \infty$. Only in the presence of label noise or teacher noise a finite optimal N emerge. Indeed, in our setting, early-stopping slows convergence, yet it mitigates the asymptotic error arising from label noise or variation in the teachers.

Regularization matrices. Zhao et al. (2024); Zhu et al. (2025) construct a task-dependent regularization matrix from the empirical covariance at each task, which requires storing a $d \times d$ matrix. However, maintaining such a matrix is computationally expensive. In principle, this can eliminate forgetting; see Evron et al. (2023, Proposition 5.5). Moreover, their analysis is restricted to a specific design of regularization matrices. In contrast, we study arbitrary diagonal regularization matrix which is considered a typical choice in continual learning (Kirkpatrick et al., 2017; Schwarz et al., 2018).

Optimal scaling. Karpel et al. (2026) showed that, under isotropic data and i.i.d. teachers, the optimal L2 regularization minimizing the expected generalization loss in the overparameterized regime scales nearly linearly with the number of tasks. In this paper, we significantly strengthen and extend their results by considering both overparameterized and underparameterized regimes, and providing almost-sure or high-probability guarantees. Furthermore, our analysis shows that even when each task is trained for many gradient steps, the predicted scaling of the optimal L2 regularization emerges only after sufficiently many tasks (see Figure 9, in Appendix B.1). Finally, we show that isotropic regularization can fail altogether even under a simple setting and we derive necessary and sufficient conditions for its optimality within the broader class of diagonal regularizers that are practically used.

Future directions. Several natural directions remain open. Future work may relax the isotropy and independence assumptions by considering non-isotropic covariances, temporally correlated features, or non-i.i.d. teacher sequences. Additional extensions include classification problems (e.g., logistic regression), nonlinear models, and more general non-stationary environments.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing within some of the proofs write-up. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and generating synthetic data sets, cleaning and reformatting dataset, assisting with translation, supporting qualitative and thematic data analysis, are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code. We have reviewed all AI-assisted work. Any LLM-generated code or text was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The required details to reproduce the main experimental results of the paper appear in Section 7. For theoretical results, we provide clear explanations of any assumptions and a complete proof for each one of the claims, see Sections 3 to 6.

REFERENCES

Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1370–1378. PMLR, 2019.

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)*, pp. 139–154, 2018.

Karl Johan Astrom and Richard M. Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, 2008.

Frederik Benzing. Unifying importance based regularisation methods for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2372–2396. PMLR, 2022.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 532–547, 2018.

James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

Romain Couillet and Zhenyu Liao. *Random matrix methods for machine learning*. Cambridge University Press, 2022.

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.

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)*, pp. 4028–4079. PMLR, 2022.

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*, pp. 9440–9484. PMLR, 23–29 Jul 2023.

Itay Evron, Ran Levinstein, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Nathan Srebro. From continual learning to SGD and back: Better rates for continual linear models. In *37th International Conference on Algorithmic Learning Theory*, 2026.

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3 (4):128–135, 1999.

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.

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.

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.

Gilad Karpel, Edward Moroshko, Ran Levinstein, Ron Meir, Daniel Soudry, and Itay Evron. Optimal l2 regularization in high-dimensional continual linear regression. In *Proceedings of The 37th International Conference on Algorithmic Learning Theory*, 2026.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Ran Levinstein, Amit Attia, Matan Schliserman, Uri Sherman, Daniel Soudry, Tomer Koren, and Itay Evron. Optimal rates in continual linear regression via increasing regularization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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 (eds.), *Proceedings of The 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pp. 513–533. PMLR, 22–25 Aug 2023.

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*, pp. 21078–21100. PMLR, 23–29 Jul 2023.

Qipeng Liu, Luojun Lin, Zhifeng Shen, and Zhifeng Yang. Periodically Exchange Teacher-Student for Source-Free Object Detection . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6391–6401. IEEE Computer Society, 2023.

Ekdeep Singh Lubana, Puja Trivedi, Danai Koutra, and Robert Dick. How do quadratic regularizers prevent catastrophic forgetting: The role of interpolation. In *Conference on Lifelong Learning Agents*, pp. 819–837. PMLR, 2022.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Pratik Patil, Arun Kumar Kuchibhotla, Yuting Wei, and Alessandro Rinaldo. Mitigating multiple descents: A model-agnostic framework for risk monotonization. *arXiv preprint arXiv:2205.12937*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4528–4537, 2018.

Jack W Silverstein and Zhi Dong Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192, 1995.

James Seale Smith, Junjiao Tian, Shaunak Halbe, Yen-Chang Hsu, and Zsolt Kira. A closer look at rehearsal-free continual learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2410–2420. IEEE, 2023.

Rishi Sonthalia, Jackie Lok, and Elizaveta Rebrova. On regularization via early stopping for least squares regression. *arXiv preprint arXiv:2406.04425*, 2024.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017.

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.

Yihan Zhao, Wenqing Su, and Ying Yang. High-dimensional asymptotics of generalization performance in continual ridge regression. *arXiv preprint arXiv:2508.15494*, 2025.

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

A NUMERICAL VALIDATIONS OF THE EXACT GENERALIZATION PREDICTIONS

In this Appendix, we perform a numerical validations of the *exact* generalization prediction produces by both Theorems 1 and 3.

A.1 NUMERICAL VALIDATION OF THE PREDICTION IN THEOREM 1

We verify the theoretical prediction from Theorem 1.

In this experiment, we set $d = 100$ and $v_x = 1$, with $N = 3$ and $\gamma = 0.2$. We examine both i.i.d. Gaussian-perturbed teachers (where the perturbation has zero mean and standard deviation of $1/2$) and sinusoidal teachers as defined in Assumption 3 (using $\mathbf{B} = \mathbf{1}_d$ and $f = 1/10$). In both configurations, the base teacher is initialized from a Bernoulli distribution with parameter $p = 0.7$. All reported results are averaged over 50 random seeds.

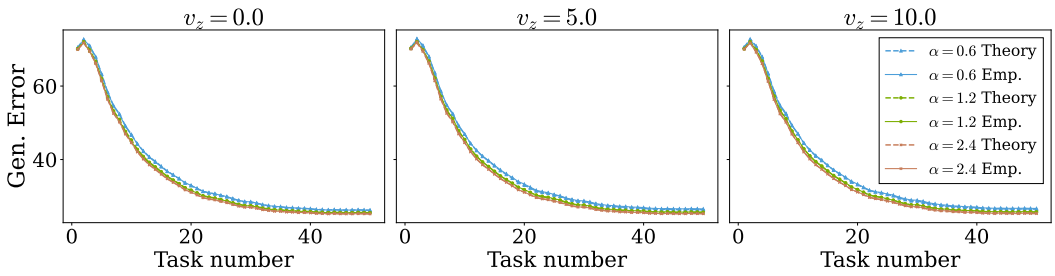


Figure 5: Continual linear regression with early-stopped GD (i.i.d. teachers).

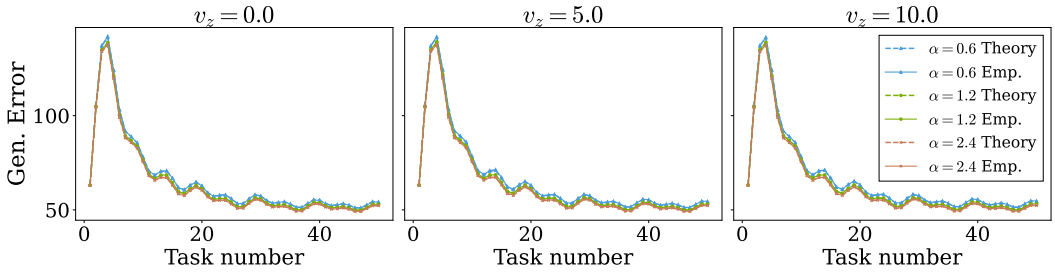


Figure 6: Continual linear regression with early-stopped GD (sinusoidal teachers).

In both Figures 5 and 6 we observe a near identical matching between the theoretical and empirical generalization loss.

A.2 Numerical validation of the prediction in Theorem 3

We verify the theoretical prediction from Theorem 3.

In this experiment, we set $d = 100$, $v_x = 1$ and use two diagonal regularizers. Ascending diagonal with entries $\mathrm{diag}(1/d, 2/d, \ldots, 1)$ and descending diagonal with entries $\mathrm{diag}(1, (d-1)/d, \ldots, 1/d)$[4], in both configurations, the teacher is defined by setting the first $d/10$ components to 1 and all remaining components to 0. All results are averaged over 50 random seeds.

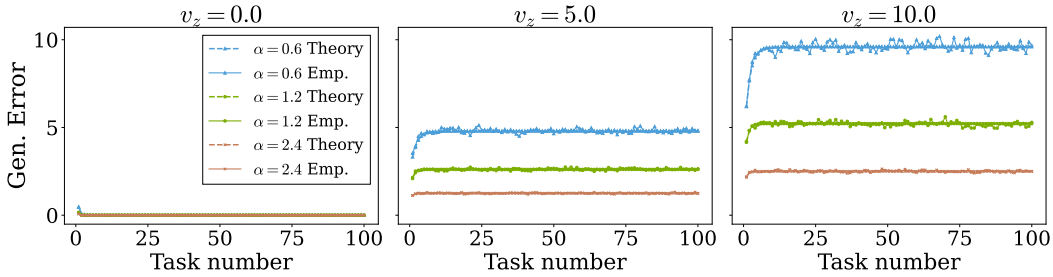


Figure 7: Continual Generalized L2 Regularization (ascending entries).

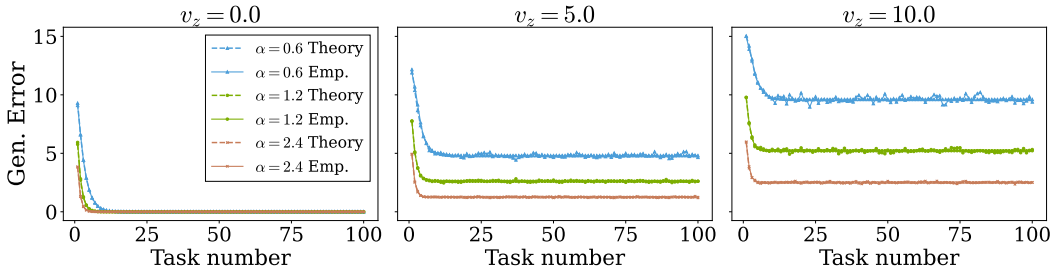


Figure 8: Continual Generalized L2 Regularization (descending entries).

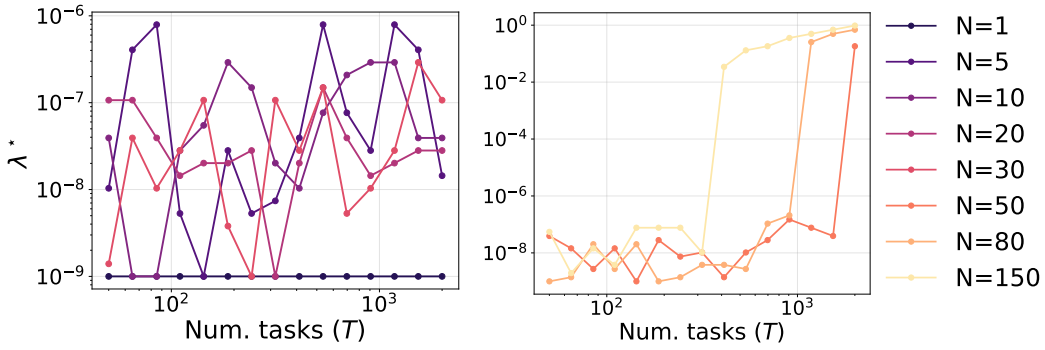
In both Figures 7 and 8 we observe a near identical matching between the theoretical and empirical generalization loss.

[4]Since d is finite-dimensional, in this situation the smallest eigenvalue is bounded from below by $1/100$. In contrast, when $d \to \infty$, such diagonal regularization matrices lie outside the class of matrices examined in the paper, as the smallest eigenvalue is not bounded below by any positive scalar.

B Additional Continual Linear Regression Experiments

B.1 Explicit Regularization Via an early-stopped GD On The Regularized Loss

We experiment with gradient descent on `MNIST`-based data (Section 7) run for N steps of over the *regularized* objective $\frac{1}{2n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2+\frac{\lambda}{2}\|\mathbf{w}-\mathbf{w}_{t-1}\|^2$. We visualize the empirical optimal regularizer over different numbers of steps and number of tasks. Evidently, even for large N (corresponding to near-convergent optimization within each task), the resulting $\lambda^\star(T)$ does not exhibit the predicted near linear scaling as predicted by Karpel et al. (2026); such behavior only begins to emerge after a large number of tasks.



(a) Effect of small steps (N) on the optimal regularizer.

(b) Effect of large steps (N) on the optimal regularizer.

Figure 9: **Explicit regularization does not replicate implicit regularization in continual learning.** We plot the optimal regularization parameter $\lambda^\star$ as a function of the number of tasks T for different numbers of gradient steps N. We compute $\lambda^\star$ with an exhaustive search.

C Proofs for Section 4: Early-Stopped GD

C.1 Auxiliary Lemmas and definitions

We begin by providing a few Lemmas, the following Lemma provides a closed-form solution to the iterates produced by early-stopped GD algorithm (Algorithm 2).

Lemma 1. *At task t the predictor update under Algorithm 2 after k gradient steps is given by,*

$$\mathbf{w}_t^k = \left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^k \mathbf{w}_{t-1} + \gamma\frac{1}{n}\sum_{i=0}^{k-1}\left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^i \mathbf{X}_t^\top\mathbf{y}_t.$$

Proof. Let $t \in [1, T]$, and let $k \in [1, N]$, then,

$$\begin{aligned}
\mathbf{w}_t^k - \mathbf{w}_t^\star &= \mathbf{w}_t^{k-1} - \gamma\nabla\frac{1}{2n}\left\|\mathbf{X}_t\mathbf{w}_t^{k-1} - \mathbf{y}_t\right\|^2 - \mathbf{w}_t^\star \\
&= \mathbf{w}_t^{k-1} - \gamma\frac{1}{n}\mathbf{X}_t^\top\left(\mathbf{X}_t\mathbf{w}_t^{k-1} - \mathbf{y}_t\right) - \mathbf{w}_t^\star \\
&= \left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)\mathbf{w}_t^{k-1} + \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{y}_t - \left[\left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)\mathbf{w}_t^\star + \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\mathbf{w}_t^\star\right] \\
&= \left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)\left(\mathbf{w}_t^{k-1} - \mathbf{w}_t^\star\right) + \gamma\frac{1}{n}\mathbf{X}_t^\top\left(\mathbf{y}_t - \mathbf{X}_t\mathbf{w}_t^\star\right) \\
&= \dots \\
&= \left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^k\left(\mathbf{w}_{t-1} - \mathbf{w}_t^\star\right) + \gamma\frac{1}{n}\sum_{i=0}^{k-1}\left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^i\mathbf{X}_t^\top\left(\mathbf{y}_t - \mathbf{X}_t\mathbf{w}_t^\star\right) \\
&= \left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^k\left(\mathbf{w}_{t-1} - \mathbf{w}_t^\star\right) + \gamma\frac{1}{n}\sum_{i=0}^{k-1}\left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^i\mathbf{X}_t^\top\mathbf{y}_t \\
&\qquad \underbrace{- \gamma\frac{1}{n}\sum_{i=0}^{k-1}\left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^i\mathbf{X}_t^\top\mathbf{X}_t\mathbf{w}_t^\star}_{*},
\end{aligned} \tag{3}$$

we continue with developing the last term in Equation (3) (denoted by *),

$$\begin{aligned}
\gamma\frac{1}{n}\sum_{i=0}^{k-1}\left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^i\mathbf{X}_t^\top\mathbf{X}_t\mathbf{w}_t^\star &= \gamma\frac{1}{n}\sum_{s=0}^{k-1}\sum_{p=0}^{s}\binom{s}{p}\left(-\gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^p\mathbf{X}_t^\top\mathbf{X}_t\mathbf{w}_t^\star \\
&= \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\sum_{s=0}^{k-1}\sum_{p=0}^{s}\binom{s}{p}\left(-\gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^p\mathbf{w}_t^\star \\
&= \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\sum_{s=0}^{k-1}\left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^s\mathbf{w}_t^\star \\
&= \left(\mathbf{I} - \left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^k\right)\mathbf{w}_t^\star,
\end{aligned} \tag{4}$$

Then by Equations (3) and (4),

$$\begin{aligned}
\mathbf{w}_t^k - \mathbf{w}_t^\star &= \left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^k\left(\mathbf{w}_{t-1} - \mathbf{w}_t^\star\right) \\
&\qquad + \gamma\frac{1}{n}\sum_{i=0}^{k-1}\left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^i\mathbf{X}_t^\top\mathbf{y}_t - \left(\mathbf{I} - \left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^k\right)\mathbf{w}_t^\star \\
&= \left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^k\mathbf{w}_{t-1} + \gamma\frac{1}{n}\sum_{i=0}^{k-1}\left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t\right)^i\mathbf{X}_t^\top\mathbf{y}_t - \mathbf{w}_t^\star.
\end{aligned}$$

The lemma is an immediate consequence of the equality stated above. □

The following Lemma provides the closed-from solution to the iterates produced by ridge-regression algorithm (Algorithm 1).

Lemma 2. *At each iteration t the predictor update under Algorithm 1 is given by*

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

Proof. Let

$$J_t(\mathbf{w}) = \frac{1}{2n}\left\|\mathbf{X}_t\mathbf{w} - \mathbf{y}_t\right\|^2 + \frac{\lambda}{2}\left\|\mathbf{w} - \mathbf{w}_{t-1}\right\|^2.$$

this function is convex, and its gradient is given by,

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

At an optimal point $\mathbf{w}_t$, the first-order optimality (stationarity) condition must hold, hence,

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

If $\lambda > 0$, then $\frac{1}{n}\mathbf{X}_t^\top \mathbf{X}_t + \lambda \mathbf{I}$ is symmetric positive definite and invertible, so

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

□

Lemma 3. *For any $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ and $\eta > 0$ we have*

$$\left\|\mathbf{u} + \mathbf{v} + \mathbf{w}\right\|^2 \le \left\|\mathbf{v}\right\|^2 (1 + \eta) + 2\left(\left\|\mathbf{u}\right\|^2 + \left\|\mathbf{w}\right\|^2\right)\left(1 + \frac{1}{\eta}\right).$$

Proof.

$$\begin{aligned}
\left\|\mathbf{u} + \mathbf{v} + \mathbf{w}\right\|^2 &= \left\|\mathbf{v}\right\|^2 + 2\langle \mathbf{v}, \mathbf{u} + \mathbf{w}\rangle + \left\|\mathbf{u} + \mathbf{w}\right\|^2 \\
&\overset{[1]}{\le} \left\|\mathbf{v}\right\|^2 + 2\left\|\mathbf{v}\right\|\left\|\mathbf{u} + \mathbf{w}\right\| + \left\|\mathbf{u} + \mathbf{w}\right\|^2 \\
&\overset{[2]}{\le} \left\|\mathbf{v}\right\|^2 + \eta\left\|\mathbf{v}\right\|^2 + \frac{1}{\eta}\left\|\mathbf{u} + \mathbf{w}\right\|^2 + \left\|\mathbf{u} + \mathbf{w}\right\|^2 \\
&= (1 + \eta)\left\|\mathbf{v}\right\|^2 + \left(1 + \frac{1}{\eta}\right)\left\|\mathbf{u} + \mathbf{w}\right\|^2 \\
&\le (1 + \eta)\left\|\mathbf{v}\right\|^2 + 2\left(1 + \frac{1}{\eta}\right)\left(\left\|\mathbf{u}\right\|^2 + \left\|\mathbf{w}\right\|^2\right),
\end{aligned}$$

where [1] follows by Cauchy-Schwarz inequality, and [2] follows by Young inequality for all $\eta > 0$. □

Lemma 4. *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ satisfy the high-dimensional setting of Section 3, and define the sample covariance $\mathbf{S} = \frac{1}{n}\mathbf{X}^\top \mathbf{X}$ with the Marchenko-Pastur distribution $\mu_{\mathrm{MP},\frac{1}{\alpha}}$. Let f be a fixed polynomial, and let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ be deterministic (or independent) vectors with bounded norm, i.e., $\|\mathbf{a}\| = \|\mathbf{b}\| = O(1)$ (where for random quantities it holds almost surely). Then as $n, d \to \infty$, we have almost surely that,*

$$\mathbf{a}^\top f(\mathbf{S})\mathbf{b} \to \Big(\int f(\lambda)\,\mathrm{d}\mu_{\mathrm{MP},\frac{1}{\alpha}}(\lambda)\Big)\mathbf{a}^\top \mathbf{b}. \tag{5}$$

Proof. Let

$$g_d(z) := \mathbf{a}^\top (\mathbf{S}_d - z\mathbf{I})^{-1}\mathbf{b}.$$

For each d, the function g_d is analytic on the domain $D = \mathbb{C} \setminus \mathrm{supp}(\mu_{\mathrm{MP},1/\alpha})$, and by spectral confinement, almost surely for all sufficiently large d, $\mathrm{spec}(\mathbf{S}_d) \subset \mathrm{supp}(\mu_{\mathrm{MP},1/\alpha}) + o(1)$, so that g_d is analytic on D for all large d. Moreover, for any compact set $K \subset D$, there exists $c_K > 0$ such that $\mathrm{dist}(z, \mathrm{spec}(\mathbf{S}_d)) \geq c_K$ for all $z \in K$ and all large d, almost surely. Hence, $|g_d(z)| \leq \|\mathbf{a}\|\|\mathbf{b}\|\|(\mathbf{S}_d - z\mathbf{I})^{-1}\| \leq \frac{\|\mathbf{a}\|\|\mathbf{b}\|}{c_K}$, so that (g_d) is uniformly bounded on compact subsets of D almost surely.

Let $E \subset D$ be a countable dense subset. By the isotropic resolvent law (see, e.g. Couillet & Liao (2022, Theorem 2.6)), for each fixed $z \in E$,

$$g_d(z) = \mathbf{a}^\top (\mathbf{S}_d - z\mathbf{I})^{-1}\mathbf{b} \to m_{1/\alpha}(z)\mathbf{a}^\top\mathbf{b},$$

almost surely. By taking a countable intersection over $z \in E$, this convergence holds simultaneously for all $z \in E$ on an event of probability one.

Therefore, by Vitali's convergence theorem, it follows that $g_d(z) \to m_{1/\alpha}(z)\mathbf{a}^\top\mathbf{b}$, uniformly on compact subsets of D, almost surely. Let $\Gamma \subset D$ be a contour enclosing $\mathrm{supp}(\mu_{\mathrm{MP},1/\alpha})$ with positive margin. By the holomorphic functional calculus,

$$\mathbf{a}^\top f(\mathbf{S}_d)\mathbf{b} = -\frac{1}{2\pi i}\oint_\Gamma f(z) g_d(z)\,\mathrm{d}z.$$

By uniform convergence of g_d on Γ, we may pass the limit inside the integral:

$$\mathbf{a}^\top f(\mathbf{S}_d)\mathbf{b} \to -\frac{1}{2\pi i}\oint_\Gamma f(z) m_{1/\alpha}(z)\,\mathrm{d}z\mathbf{a}^\top\mathbf{b},$$

almost surely. Finally, using the Stieltjes transform representation $m_{1/\alpha}(z) = \int \frac{1}{\lambda - z}\,\mathrm{d}\mu_{\mathrm{MP},1/\alpha}(\lambda)$, and Cauchy's integral formula, we obtain $-\frac{1}{2\pi i}\oint_\Gamma f(z) m_{1/\alpha}(z)\,\mathrm{d}z = \int f(\lambda)\,\mathrm{d}\mu_{\mathrm{MP},1/\alpha}(\lambda)$. This completes the proof. □

We aim to establish the tight relationship between ridge regression and early-stopped gradient descent. To do so, we partially rely on the results developed in Karpel et al. (2026). Note that our algorithmic formulation of ridge regression (Algorithm 1) differs slightly from the one used in Karpel et al. (2026); nonetheless, the following lemma establishes the connection between the two formulations.

Lemma 5. *Let $t \in [1, T]$ and consider the update*

$$\mathbf{w}_t \leftarrow \mathrm{argmin}_{\mathbf{w}}\Big\{\frac{1}{n}\|\mathbf{X}_t\mathbf{w} - \mathbf{y}_t\|^2 + \lambda\|\mathbf{w} - \mathbf{w}_{t-1}\|^2\Big\}$$

and the update

$$\mathbf{w}'_t \leftarrow \mathrm{argmin}_{\mathbf{w}}\Big\{\|\mathbf{X}_t\mathbf{w} - \mathbf{y}_t\|^2 + \tilde{\lambda}d\|\mathbf{w} - \mathbf{w}_{t-1}\|^2\Big\}$$

then if $\tilde{\lambda} = \frac{n}{d}\lambda$ so $\mathbf{w}'_t = \mathbf{w}_t$.

Proof. Using Lemma 2 we have,

$$\begin{aligned}\mathbf{w}_t &= \left(\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t + \lambda\mathbf{I}\right)^{-1}\left(\frac{1}{n}\mathbf{X}_t^\top\mathbf{y}_t + \lambda\mathbf{w}_{t-1}\right)\\ &= \left(\mathbf{X}_t^\top\mathbf{X}_t + \lambda n\mathbf{I}\right)^{-1}\left(\mathbf{X}_t^\top\mathbf{y}_t + \lambda n\mathbf{w}_{t-1}\right) = \left(\mathbf{X}_t^\top\mathbf{X}_t + \tilde{\lambda}d\mathbf{I}\right)^{-1}\left(\mathbf{X}_t^\top\mathbf{y}_t + \tilde{\lambda}d\mathbf{w}_{t-1}\right) = \mathbf{w}'_t.\end{aligned}$$

which is the exact update solution for $\mathbf{w}'_t$ after redefining $\tilde{\lambda} = \frac{n}{d}\lambda$. □

By Lemma 5, if we adopt the algorithmic formulation of ridge regression from Karpel et al. (2026) and simply replace $\tilde{\lambda}$ with $\tilde{\lambda} := \frac{n}{d}\lambda = \alpha\lambda$, where λ denotes the regularization parameter used in ridge-regression (Algorithm 1), then the iterates coincide at every step. Consequently, the expression in Corollary 1 also remains unchanged under our algorithmic formulation (after substituting $\tilde{\lambda}$).

Lemma 6. *Let h be analytic at ∞, and define $h(x) = g(1/x)$, where g is analytic in a neighborhood of 0. Let the Taylor expansion of g at 0 be*

$$g(x) = P_k(x) + R(x), \qquad R(x) = O(x^{k+1}),$$

where P_k is a polynomial of degree at most k. Then, as $x \to \infty$,

$$\frac{\mathrm{d}}{\mathrm{d}x} h(x) = -\frac{1}{x^2} P_k'\left(\frac{1}{x}\right) + O\left(\frac{1}{x^{k+2}}\right).$$

Proof. Since g is analytic at 0, there exists $\rho > 0$ such that for all $|x| < \rho$,

$$g(x) = \sum_{n=0}^{k} \frac{g^{(n)}(0)}{n!} x^n + \sum_{n=k+1}^{\infty} \frac{g^{(n)}(0)}{n!} x^n.$$

Power series may be differentiated term wise within their radius of convergence, and the resulting series converges uniformly on any closed interval $|x| \le r < \rho$. Differentiating, we obtain

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

For $|x| < r$, write $|x| = r\theta$ with $\theta \in [0, 1]$. Then

$$\begin{aligned} |g'(x) - P_k'(x)| &= \sum_{n=k+1}^{\infty} \frac{|g^{(n)}(0)|}{(n-1)!} |x|^{n-1} = \sum_{n=k+1}^{\infty} \frac{|g^{(n)}(0)|}{(n-1)!} (r\theta)^{n-1} \\ &\le \theta^k \sum_{n=k+1}^{\infty} \frac{|g^{(n)}(0)|}{(n-1)!} r^{n-1} = C\theta^k = \frac{C}{r^k} |x|^k, \end{aligned}$$

where the series is convergent since $r < \rho$. Thus,

$$g'(x) = P_k'(x) + O(x^k), \qquad |x| \le r.$$

Finally, using $h(x) = g(1/x)$ and the chain rule, for $|x| \ge \frac{1}{r}$,

$$\frac{\mathrm{d}}{\mathrm{d}x} h(x) = g'\left(\frac{1}{x}\right)\left(-\frac{1}{x^2}\right) = \left[P_k'\left(\frac{1}{x}\right) + O\left(\frac{1}{x^k}\right)\right]\left(-\frac{1}{x^2}\right) = -\frac{1}{x^2} P_k'\left(\frac{1}{x}\right) + O\left(\frac{1}{x^{k+2}}\right).$$

□

Lemma 7. *Let $f, g : \mathbb{R} \to \mathbb{R}$, be analytic functions at ∞, assume that as $x \to \infty$, $f(x) = \sum_{s=0}^{k} a_s x^{-s} + O(x^{-k-1})$ and $g(x) = \sum_{s=0}^{k} a_s x^{-s} + O(x^{-k-1})$ with some non-zero coefficient which is not a_0, then $\left|\frac{\mathrm{d}}{\mathrm{d}x} f(x) - \frac{\mathrm{d}}{\mathrm{d}x} g(x)\right| = o\left(\frac{\mathrm{d}}{\mathrm{d}x} g(x)\right)$.*

Proof. Since f, g are analytic function, we use Lemma 6 to obtain,

$$\left|\frac{\mathrm{d}}{\mathrm{d}x} f(x) - \frac{\mathrm{d}}{\mathrm{d}x} g(x)\right| = O(x^{-k-2}).$$

Since $\frac{\mathrm{d}}{\mathrm{d}x} g(x) = \Omega(x^{-k-1})$ we get $\left|\frac{\mathrm{d}}{\mathrm{d}x} f(x) - \frac{\mathrm{d}}{\mathrm{d}x} g(x)\right| = o\left(\frac{\mathrm{d}}{\mathrm{d}x} g(x)\right)$. □

C.2 Proof of Theorem 1: closed-form expression for the generalization loss from the early-stopped GD algorithm

Recall Theorem 1. Recall the high-dimensional setting where the number of samples and the dimension grow at proportional rate $\alpha := \frac{n}{d}$. Moreover, recall that v_x, v_z are the variances of features and labels, respectively. Then, the generalization loss after T tasks of the early-stopped GD algorithm (Algorithm 2) satisfies, as $n, d \to \infty$,

$$G_T^{\mathrm{esgd}} \xrightarrow{P} v_z c' \frac{1-a'^T}{1-a'} + \frac{1}{T}\sum_{i=1}^{T} \|\mathbf{w}_T^\star - \mathbf{w}_i^\star\|^2 + 2\big(\mathbf{w}_T^\star - \bar{\mathbf{w}}^\star\big)^\top \sum_{i=1}^{T} b'^{T-i+1}\big(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\big)$$
$$+ \sum_{i=1}^{T}\sum_{i'=1}^{T} a'^{T-\max(i,i')+1} b'^{|i-i'|}\big(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\big)^\top\big(\mathbf{w}_{i'-1}^\star - \mathbf{w}_{i'}^\star\big),$$

where we denote $\mathbf{w}_0^\star = \mathbf{w}_0$ and $\bar{\mathbf{w}}^\star = \frac{1}{T}\sum_i \mathbf{w}_i^\star$,

$$a' = 1 - 2Nv_x\gamma + O\big((N\gamma)^2\big), \qquad b' = 1 - Nv_x\gamma + O\big((N\gamma)^2\big),$$
$$c' = \frac{N^2}{\alpha} v_x\gamma^2 + O\big((N\gamma)^3\big). \tag{6}$$

Recall Lemma 1,

$$\mathbf{w}_T^{\mathrm{esgd}} = \left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_T^\top\mathbf{X}_T\right)^N \mathbf{w}_{T-1}^{\mathrm{esgd}} + \gamma\frac{1}{n}\sum_{s=0}^{N-1}\left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_T^\top\mathbf{X}_T\right)^s \mathbf{X}_T^\top\mathbf{y}_T,$$

and let $\mathbf{P}_k = \big(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_k^\top\mathbf{X}_k\big)$, with the convention $\mathbf{w}_0^\star = \mathbf{w}_0^{\mathrm{esgd}}$.

Then,

$$\begin{aligned}
\mathbf{w}_T^{\mathrm{esgd}} &= \left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_T^\top\mathbf{X}_T\right)^N \mathbf{w}_{T-1}^{\mathrm{esgd}} + \gamma\frac{1}{n}\sum_{s=0}^{N-1}\left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}_T^\top\mathbf{X}_T\right)^s \mathbf{X}_T^\top\mathbf{y}_T \\
&= \mathbf{P}_T^N\mathbf{w}_{T-1}^{\mathrm{esgd}} + \gamma\frac{1}{n}\sum_{s=0}^{N-1}\mathbf{P}_T^s\mathbf{X}_T^\top\left(\mathbf{X}_T\mathbf{w}_T^\star + \mathbf{z}_T\right) \\
&= \mathbf{P}_T^N\mathbf{w}_{T-1}^{\mathrm{esgd}} + \gamma\frac{1}{n}\sum_{s=0}^{N-1}\mathbf{P}_T^s\mathbf{X}_T^\top\mathbf{X}_T\mathbf{w}_T^\star + \gamma\frac{1}{n}\sum_{s=0}^{N-1}\mathbf{P}_T^s\mathbf{X}_T^\top\mathbf{z}_T.
\end{aligned}$$

Using $\mathbf{I} - \mathbf{P}_T = \gamma\frac{1}{n}\mathbf{X}_T^\top\mathbf{X}_T$, we obtain

$$\gamma\frac{1}{n}\sum_{s=0}^{N-1}\mathbf{P}_T^s\mathbf{X}_T^\top\mathbf{X}_T = \sum_{s=0}^{N-1}\mathbf{P}_T^s(\mathbf{I} - \mathbf{P}_T) = \mathbf{I} - \mathbf{P}_T^N.$$

Therefore,

$$\mathbf{w}_T^{\mathrm{esgd}} = \mathbf{P}_T^N\mathbf{w}_{T-1}^{\mathrm{esgd}} + \left(\mathbf{I} - \mathbf{P}_T^N\right)\mathbf{w}_T^\star + \gamma\frac{1}{n}\sum_{s=0}^{N-1}\mathbf{P}_T^s\mathbf{X}_T^\top\mathbf{z}_T. \tag{7}$$

Since $G_T^{\mathrm{esgd}} = \frac{1}{T}\sum_{i=1}^{T}\|\mathbf{w}_T^{\mathrm{esgd}} - \mathbf{w}_i^\star\|^2$, we therefore develop the corresponding expression for $\mathbf{w}_T^{\mathrm{esgd}} - \mathbf{w}_i^\star$, where $i \in [1, T]$. Using Equation (7),

$$\mathbf{w}_T^{\mathrm{esgd}} - \mathbf{w}_i^\star = \mathbf{P}_T^N\left(\mathbf{w}_{T-1}^{\mathrm{esgd}} - \mathbf{w}_i^\star\right) + \left(\mathbf{I} - \mathbf{P}_T^N\right)\left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right) + \gamma\frac{1}{n}\sum_{s=0}^{N-1}\mathbf{P}_T^s\mathbf{X}_T^\top\mathbf{z}_T.$$

To rewrite this term note that for all $k \in [1, T]$,

$$\left(\mathbf{I} - \mathbf{P}_k^N\right)\left(\mathbf{w}_k^\star - \mathbf{w}_i^\star\right) = \left[\left(\mathbf{w}_k^\star - \mathbf{w}_i^\star\right) - \mathbf{P}_k^N\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_i^\star\right)\right] + \mathbf{P}_k^N\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right).$$

Unrolling the recursion and using the telescoping of the bracketed term yields

$$\mathbf{w}_T^{\mathrm{esgd}} - \mathbf{w}_i^\star = \sum_{k=1}^{T} \left[\prod_{m=T}^{k+1} \mathbf{P}_m^N\right] \boldsymbol{\gamma} \frac{1}{n} \sum_{s=0}^{N-1} \mathbf{P}_k^s \mathbf{X}_k^\top \mathbf{z}_k + \sum_{k=1}^{T} \left[\prod_{m=T}^{k} \mathbf{P}_m^N\right] \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right),$$

so,

$$\begin{aligned}
&\left\|\mathbf{w}_T^{\mathrm{esgd}} - \mathbf{w}_i^\star\right\|^2 \\
&= \left(\sum_{k=1}^{T} \left[\prod_{m=T}^{k+1} \mathbf{P}_m^N\right] \boldsymbol{\gamma} \frac{1}{n} \sum_{s=0}^{N-1} \mathbf{P}_k^s \mathbf{X}_k^\top \mathbf{z}_k + \sum_{k=1}^{T} \left[\prod_{m=T}^{k} \mathbf{P}_m^N\right] \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right)\right)^\top \\
&\qquad \left(\sum_{k=1}^{T} \left[\prod_{m=T}^{k+1} \mathbf{P}_m^N\right] \boldsymbol{\gamma} \frac{1}{n} \sum_{s=0}^{N-1} \mathbf{P}_k^s \mathbf{X}_k^\top \mathbf{z}_k + \sum_{k=1}^{T} \left[\prod_{m=T}^{k} \mathbf{P}_m^N\right] \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right)\right) \\
&\xrightarrow{P}^{[1]} \underbrace{\sum_{j=1}^{T} \gamma \frac{1}{n} \sum_{s=0}^{N-1} \mathbf{z}_j^\top \mathbf{X}_j \mathbf{P}_j^s \left(\prod_{k=j+1}^{T} \mathbf{P}_k^N\right)^\top \sum_{j'=1}^{T} \gamma \frac{1}{n} \left(\prod_{k=j'+1}^{T} \mathbf{P}_k^N\right) \sum_{s=0}^{N-1} \mathbf{P}_{j'}^s \mathbf{X}_{j'}^\top \mathbf{z}_{j'}}_{\text{Term 1}} \\
&+ \underbrace{\left(\sum_{k=1}^{T} \left[\prod_{m=k}^{T} \mathbf{P}_m^N\right] \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right)\right)^\top \left(\sum_{k=1}^{T} \left[\prod_{m=k}^{T} \mathbf{P}_m^N\right] \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right)\right)}_{\text{Term 2}},
\end{aligned} \tag{8}$$

where [1] follows since $\mathbb{E}[\mathbf{z}_k] = \mathbf{0}$ for all $k \in [1, T]$, and the noise variables are sampled independently of all tasks, so by the weak law of large numbers cross terms vanish as $n \to \infty$.

Let $\boldsymbol{\xi}_r^l = \left(\prod_{k=r+1}^{l} \mathbf{P}_k^N\right) \left[\sum_{s=0}^{N-1} \mathbf{P}_r^s \mathbf{X}_r^\top\right] \mathbf{z}_r$. Then, the first term becomes,

$$\begin{aligned}
&\sum_{j=1}^{T} \left[\gamma \frac{1}{n} \sum_{s=0}^{N-1} \mathbf{z}_j^\top \mathbf{X}_j \mathbf{P}_j^s\right] \left(\prod_{k=j+1}^{T} \mathbf{P}_k^N\right)^\top \sum_{j'=1}^{T} \left(\prod_{k=j'+1}^{T} \mathbf{P}_k^N\right) \left[\gamma \frac{1}{n} \sum_{s=0}^{N-1} \mathbf{P}_{j'}^s \mathbf{X}_{j'}^\top \mathbf{z}_{j'}\right] \\
&= \sum_{j=1}^{T} \sum_{j'=1}^{T} \left[\gamma \frac{1}{n} \sum_{s=0}^{N-1} \mathbf{z}_j^\top \mathbf{X}_j \mathbf{P}_j^s\right] \left(\prod_{k=j+1}^{T} \mathbf{P}_k^N\right)^\top \left(\prod_{k=j'+1}^{T} \mathbf{P}_k^N\right) \left[\gamma \frac{1}{n} \sum_{s=0}^{N-1} \mathbf{P}_{j'}^s \mathbf{X}_{j'}^\top \mathbf{z}_{j'}\right] \\
&\xrightarrow{P}^{[1]} \sum_{j=1}^{T} \mathbf{z}_j^\top \left[\gamma \frac{1}{n} \sum_{s=0}^{N-1} \mathbf{X}_j \mathbf{P}_j^s\right] \left(\prod_{k=j+1}^{T} \mathbf{P}_k^N\right)^\top \left(\prod_{k=j+1}^{T} \mathbf{P}_k^N\right) \left[\gamma \frac{1}{n} \sum_{s=0}^{N-1} \mathbf{P}_j^s \mathbf{X}_j^\top\right] \mathbf{z}_j \\
&= \gamma^2 \sum_{j=1}^{T} \left[\frac{\boldsymbol{\xi}_j^{(T-1)^\top}}{n} \mathbf{P}_T^{2N} \frac{\boldsymbol{\xi}_j^{(T-1)}}{n}\right] \\
&\xrightarrow{P}^{[2]} \frac{\gamma^2}{n^2} \sum_{j=1}^{T} a' \boldsymbol{\xi}_j^{(T-1)^\top} \boldsymbol{\xi}_j^{(T-1)} \\
&= \frac{\gamma^2}{n^2} \sum_{j=1}^{T} a' \mathbf{z}_j^\top \left[\sum_{s=0}^{N-1} \mathbf{X}_j \mathbf{P}_j^s\right] \left(\prod_{k=j+1}^{T-1} \mathbf{P}_k^N\right)^\top \left(\prod_{k=j+1}^{T-1} \mathbf{P}_k^N\right) \left[\sum_{s=0}^{N-1} \mathbf{P}_j^s \mathbf{X}_j^\top\right] \mathbf{z}_j \\
&= \ldots \xrightarrow{P} \frac{\gamma^2}{n^2} \sum_{j=1}^{T} a'^{T-j} \mathbf{z}_j^\top \sum_{s=0}^{N-1} \mathbf{X}_j \mathbf{P}_j^s \sum_{s=0}^{N-1} \mathbf{P}_j^s \mathbf{X}_j^\top \mathbf{z}_j \\
&= \frac{\gamma^2}{n^2} \sum_{j=1}^{T} a'^{T-j} \mathbf{z}_j^\top \mathbf{X}_j \left(\sum_{s,s'=0}^{N-1} \mathbf{P}_j^{s+s'}\right) \mathbf{X}_j^\top \mathbf{z}_j.
\end{aligned} \tag{9}$$

where [1] follows since under the statistical design (Section 3) in which $\mathbb{E}[\mathbf{z}_k] = \mathbf{0}$ for all $k \in [1,T]$, and the noise variables are sampled independently of all tasks, by the weak law of large numbers cross terms vanish as $n \to \infty$. And [2] follows from Lemma 4, by the independence of the random variable $\boldsymbol{\xi}_j^{t-1}$ with $\mathbf{P}_t$ and since $\left\|\boldsymbol{\xi}_j^t\right\|^2/n < \infty$ almost surely for all $t \in [1,T]$. Also, note that we substituted a' as the $2N$-th moment of the Marchenko-Pastur law.

Using the closed-form expression for the k-th moment of the Marchenko-Pastur law, we have (for all $j \in [1,T]$),

$$\begin{aligned}
&\frac{\gamma^2}{n^2}\mathbf{z}_j^\top \mathbf{X}_j \left(\sum_{s,s'=0}^{N-1} \mathbf{P}_j^{s+s'}\right) \mathbf{X}_j^\top \mathbf{z}_j = \frac{\gamma^2}{n^2}\mathrm{Tr}\left(\mathbf{z}_j^\top \mathbf{X}_j \sum_{s=0}^{2N-2} m_s \mathbf{P}_j^s \mathbf{X}_j^\top \mathbf{z}_j\right) \\
&\xrightarrow{P}^{[1]} v_z \frac{\gamma^2}{n^2}\mathrm{Tr}\left(\mathbf{X}_j^\top \mathbf{X}_j \sum_{s=0}^{2N-2} m_s \mathbf{P}_j^s\right) = \frac{1}{n}\sum_{s=0}^{2N-2} m_s \sum_{l=0}^{s}\binom{s}{l}(-\gamma)^{l+2}\frac{1}{n}\mathrm{Tr}\left[\left(\frac{1}{n}\mathbf{X}_j^\top \mathbf{X}_j\right)^{l+1}\right] \\
&\xrightarrow{P}^{[2]} v_z \sum_{s=0}^{2N-2} m_s \sum_{l=0}^{s}\binom{s}{l}(-\gamma)^{l+2}\sum_{r=0}^{l}\frac{v_x^{l+1}}{r+1}\binom{l+1}{r}\binom{l}{r}\frac{1}{\alpha^{r+1}} := v_z c'(\gamma, N) := v_z c', \qquad (10)
\end{aligned}$$

where,

$$m_s := \#\left\{(u,v) \in \{0,\ldots,N-1\}^2 : u+v=s\right\} = \begin{cases} s+1, & 0 \le s \le N-1, \\ 2N-1-s, & N \le s \le 2N-2, \\ 0, & \text{otherwise.} \end{cases},$$

and [1] follows by the weak law of large numbers. Moreover, [2] follows by Lemma 4 with the closed-form expression. Similarly,

$$a' = a'(\gamma, N) = 1 + \sum_{k=0}^{2N}\binom{2N}{k}(-\gamma)^k \sum_{r=0}^{k-1}\frac{v_x^k}{r+1}\binom{k}{r}\binom{k-1}{r}\frac{1}{\alpha^r} \qquad (11)$$

Then the first term in Equations (9) and (10) becomes,

$$\sum_{j=1}^{T} a'^{T-j}\frac{\gamma^2}{n}\mathbf{z}_j^\top \sum_{s,s'=0}^{N-1}\mathbf{P}_j^{s+s'}\frac{1}{n}\mathbf{X}_j^\top \mathbf{X}_j \mathbf{z}_j = v_z c' \sum_{j=1}^{T} a'^{T-j} = v_z c' \frac{1-a'^T}{1-a'}$$

And the second term in Equation (8) becomes,

$$\begin{aligned}
&\left(\sum_{k=1}^{T}\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top \left[\prod_{m=T}^{k}\mathbf{P}_m^N\right] + \left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right)^\top\right)\left(\sum_{k=1}^{T}\left[\prod_{m=k}^{T}\mathbf{P}_m^N\right]\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right)\right) \\
&= \sum_{k=1}^{T}\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top \left[\prod_{m=T}^{k}\mathbf{P}_m^N\right]\sum_{k=1}^{T}\left[\prod_{m=k}^{T}\mathbf{P}_m^N\right]\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) \\
&\qquad + 2\sum_{k=1}^{T}\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top \left[\prod_{m=T}^{k}\mathbf{P}_m^N\right]\left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right) + \left\|\mathbf{w}_T^\star - \mathbf{w}_i^\star\right\|^2. \qquad (12)
\end{aligned}$$

The first term in Equation (12) becomes,

$$\begin{aligned}
&\sum_{k=1}^{T}\sum_{k'=1}^{T}\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top \left[\prod_{m=k}^{T}\mathbf{P}_m^N \prod_{m=T}^{k'}\mathbf{P}_m^N\right]\left(\mathbf{w}_{k'-1}^\star - \mathbf{w}_{k'}^\star\right) \\
&= \sum_{k=1}^{T}\sum_{k'=1}^{T}\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top \left[\prod_{m=(k\vee k')}^{T}\mathbf{P}_m^N \prod_{m=T}^{(k\vee k')}\mathbf{P}_m^N \prod_{m=(k\vee k')-1}^{(k'\wedge k)}\mathbf{P}_m^N\right]\left(\mathbf{w}_{k'-1}^\star - \mathbf{w}_{k'}^\star\right) \\
&\xrightarrow{P}^{[1]} \sum_{k=1}^{T}\sum_{k'=1}^{T} a'^{T\left(k\vee k'\right)+1} b'^{|k-k'|}\left(\mathbf{w}_{k'-1}^\star - \mathbf{w}_{k'}^\star\right)^\top \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right),
\end{aligned}$$

To justify [1] we similarly define $\boldsymbol{\xi}_r^i = \prod_{m=i}^{r} \mathbf{P}_m^N \left(\mathbf{w}_{r-1}^\star - \mathbf{w}_r^\star\right)$ and $\tilde{\boldsymbol{\xi}}_r^i = \prod_{m=r}^{i} \mathbf{P}_m^N \left(\mathbf{w}_{r-1}^\star - \mathbf{w}_r^\star\right)$ (where products over empty index sets are defined to be the identity). Then for all $k, k' \in [1, T]$ we have,

$$
\begin{aligned}
&\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top \left[\prod_{m=(k\vee k')}^{T} \mathbf{P}_m^N \prod_{m=T}^{(k\vee k')} \mathbf{P}_m^N \prod_{m=(k\vee k')-1}^{(k'\wedge k)} \mathbf{P}_m^N\right] \left(\mathbf{w}_{k'-1}^\star - \mathbf{w}_{k'}^\star\right) \\
&= \left(\boldsymbol{\xi}_{T-1}^{(k\vee k')}\right)^\top \mathbf{P}_T^{2N} \tilde{\boldsymbol{\xi}}_{T-1}^{(k\wedge k')} \xrightarrow{P}{}^{[1]} a' \left(\boldsymbol{\xi}_{T-1}^{(k\vee k')}\right)^\top \tilde{\boldsymbol{\xi}}_{T-1}^{(k\wedge k')} = a' \left(\boldsymbol{\xi}_{T-2}^{(k\vee k')}\right)^\top \mathbf{P}_{T-1}^{2N} \tilde{\boldsymbol{\xi}}_{T-2}^{(k\wedge k')} \\
&= \dots \\
&\xrightarrow{P} a'^{T-(k\vee k')+1} \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top \prod_{m=(k\vee k')-1}^{(k'\wedge k)} \mathbf{P}_m^N \left(\mathbf{w}_{k'-1}^\star - \mathbf{w}_{k'}^\star\right) \\
&= a'^{T-(k\vee k')+1} \left[\left(\boldsymbol{\xi}_{(k\vee k')-1}^{(k\vee k')}\right)^\top \mathbf{P}_{(k\vee k')-1}^N \tilde{\boldsymbol{\xi}}_{(k\vee k')-1}^{(k\wedge k')}\right] \\
&\xrightarrow{P}{}^{[1]} a'^{T-(k\vee k')+1} b' \left(\boldsymbol{\xi}_{(k\vee k')-1}^{(k\vee k')}\right)^\top \tilde{\boldsymbol{\xi}}_{(k\vee k')-1}^{(k\wedge k')} \\
&= \dots \\
&\xrightarrow{P} a'^{T-(k\vee k')+1} b'^{|k-k'|} \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top \left(\mathbf{w}_{k'-1}^\star - \mathbf{w}_{k'}^\star\right),
\end{aligned}
\tag{13}
$$

Here, [1] follows from Lemma 4 and from the independence of $\boldsymbol{\xi}_{t-1}^{(k\vee k')}$ and $\tilde{\boldsymbol{\xi}}_{t-1}^{(k\wedge k')}$ from $\mathbf{P}_t$, together with the facts that $\left\|\boldsymbol{\xi}_{t-1}^{(k\vee k')}\right\| \in O(1)$ and $\left\|\tilde{\boldsymbol{\xi}}_{t-1}^{(k\wedge k')}\right\| \in O(1)$ almost surely, which follows since $\left\|\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right\| \in O(1)$ (for all $k \in [1, T]$, see Section 3) and $\|\mathbf{P}_t\|_2 < 1$ for all $t \in [1, T]$ almost surely, as $\gamma < \frac{2}{v_x\left(1+\sqrt{\frac{1}{\alpha}}\right)^2}$. Furthermore, using the closed expression for the N-th moment in the Marchenko-Pastur law,

$$
b' = b'(\gamma, N) = 1 + \sum_{k=0}^{N} \binom{N}{k} (-\gamma)^k \sum_{r=0}^{k-1} \frac{v_x^k}{r+1} \binom{k}{r} \binom{k-1}{r} \frac{1}{\alpha^r}
\tag{14}
$$

Since $\left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top \left[\prod_{m=T}^{k} \mathbf{P}_m^N\right] \left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right) = b'^{T-k+1} \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top \left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right)$ which similarly follows by the identity in Equation (13). Then, substituting back to Equation (8), yields,

$$
\begin{aligned}
\left\|\mathbf{w}_T^{\text{esgd}} - \mathbf{w}_i^\star\right\|^2 \xrightarrow{P} v_z c' \frac{1 - a'^T}{1 - a'} &+ \sum_{k=1}^{T} \sum_{k'=1}^{T} a'^{T-(k\vee k')+1} b'^{|k-k'|} \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top \left(\mathbf{w}_{k'-1}^\star - \mathbf{w}_{k'}^\star\right) \\
&+ 2 \sum_{k=1}^{T} b'^{T-k+1} \left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right)^\top \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right) + \left\|\mathbf{w}_T^\star - \mathbf{w}_i^\star\right\|^2.
\end{aligned}
\tag{15}
$$

And therefore,

$$
\begin{aligned}
G_T^{\text{esgd}} &= \frac{1}{T} \sum_{i=1}^{T} \|\mathbf{w}_T - \mathbf{w}_i^\star\|^2 \\
&\xrightarrow{P} v_z c' \frac{1 - a'^T}{1 - a'} + \frac{1}{T} \sum_{i=1}^{T} \|\mathbf{w}_T^\star - \mathbf{w}_i^\star\|^2 + 2\left(\mathbf{w}_T^\star - \bar{\mathbf{w}}^\star\right)^\top \sum_{i=1}^{T} b'^{T-i+1} \left(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\right) \\
&\quad + \sum_{i=1}^{T} \sum_{i'=1}^{T} a'^{T-\max(i,i')+1} b'^{|i-i'|} \left(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\right)^\top \left(\mathbf{w}_{i'-1}^\star - \mathbf{w}_{i'}^\star\right),
\end{aligned}
$$

Here, $\bar{\mathbf{w}}^\star$ denotes the average teacher, i.e., $\bar{\mathbf{w}}^\star = \frac{1}{T}\sum_{i=1}^{T} \mathbf{w}_i^\star$.

By taking the Taylor expansion around $N\gamma \to 0$ for a', b', c', we get,

$$
\begin{aligned}
a'(\gamma, N) &= 1 - 2Nv_x\gamma + N(2N-1)v_x^2\left(1+\frac{1}{\alpha}\right)\gamma^2 + O\big((N\gamma)^3\big), \\
b'(\gamma, N) &= 1 - Nv_x\gamma + \frac{N(N-1)}{2}v_x^2\left(1+\frac{1}{\alpha}\right)\gamma^2 + O\big((N\gamma)^3\big), \\
c'(\gamma, N) &= \frac{N^2}{\alpha}v_x\gamma^2 + O\big((N\gamma)^3\big).
\end{aligned}
\tag{16}
$$

Before deriving the result of the optimal step size in early-stopped GD we simplify the expression of the generalization loss given in Theorem 1 under the assumption of the i.i.d. teachers (Assumption 1).

Corollary 1. *Under the i.i.d. setting (Assumption 1), and for all $T > 1$, the generalization loss of Algorithm 1 becomes, as $n, d \to \infty$*

$$G_T^{\text{ridge}} \xrightarrow{P} v_z c' \frac{1-a'^T}{1-a'} + a'^T \left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2 + 2\text{Tr}(\mathbf{\Sigma}) \left(\frac{1-b'}{1-a'} + \frac{2b'^T - 1}{T} + \frac{a'^T\left(-a' + 2b' - 1\right)}{2(1-a')} \right). \tag{17}$$

where

$$\begin{aligned}
a' &= 1 + \sum_{k=0}^{2N} \binom{2N}{k} (-\gamma)^k \sum_{r=0}^{k-1} \frac{v_x^k}{r+1} \binom{k}{r} \binom{k-1}{r} \frac{1}{\alpha^r}, \\
b' &= 1 + \sum_{k=0}^{N} \binom{N}{k} (-\gamma)^k \sum_{r=0}^{k-1} \frac{v_x^k}{r+1} \binom{k}{r} \binom{k-1}{r} \frac{1}{\alpha^r}, \\
c' &= \sum_{s=0}^{2N-2} m_s \sum_{l=0}^{s} \binom{s}{l} (-\gamma)^{l+2} \sum_{r=0}^{l} \frac{v_x^{l+1}}{r+1} \binom{l+1}{r} \binom{l}{r} \frac{1}{\alpha^{r+1}}.
\end{aligned}$$

Proof. Under the i.i.d. teacher assumption 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 $\mathbf{0}_d$ and covariance $\text{Cov}(\boldsymbol{\xi}_i) = \mathbf{\Sigma}$. Substituting this representation into the general expression derived in Equation (15), yields the following explicit decomposition of the squared error,

$$\begin{aligned}
\left\| \mathbf{w}_T^{\text{esgd}} - \mathbf{w}_i^\star \right\|^2 \xrightarrow{P} & a'^T \left\|\mathbf{w}^\star - \mathbf{w}_0 + \boldsymbol{\xi}_1\right\|^2 + 2b'^T \left(\boldsymbol{\xi}_i - \boldsymbol{\xi}_T\right)^\top \left(\mathbf{w}^\star - \mathbf{w}_0 + \boldsymbol{\xi}_1\right) \\
& + v_z c' \frac{1-a'^T}{1-a'} + 2\sum_{k=2}^{T} a'^{T-k+1} b'^{k-1} \left(\boldsymbol{\xi}_k - \boldsymbol{\xi}_{k-1}\right)^\top \left(\mathbf{w}^\star - \mathbf{w}_0 + \boldsymbol{\xi}_1\right) \\
& + \sum_{k=2}^{T}\sum_{k'=2}^{T} a'^{T-(k\vee k')+1} b'^{(k\vee k')-(k\wedge k')} \left(\boldsymbol{\xi}_{k-1} - \boldsymbol{\xi}_k\right)^\top \left(\boldsymbol{\xi}_{k'-1} - \boldsymbol{\xi}_{k'}\right) \\
& + 2\sum_{k=2}^{T} b'^{T-k+1} \left(\boldsymbol{\xi}_T - \boldsymbol{\xi}_i\right)^\top \left(\boldsymbol{\xi}_{k-1} - \boldsymbol{\xi}_k\right) + \left\|\boldsymbol{\xi}_T - \boldsymbol{\xi}_i\right\|^2.
\end{aligned}$$

Additionally, for all $i, k \in [T]$ so that $i \neq k$ we have,

$$\begin{aligned}
\boldsymbol{\xi}_i^\top \boldsymbol{\xi}_i = \sum_{j=1}^{d} \xi_{i,j}^2 \xrightarrow{P} \text{Tr}\left(\mathbf{\Sigma}\right), \quad \boldsymbol{\xi}_i^\top \boldsymbol{\xi}_k = \sum_{j=1}^{d} \xi_{i,j}\xi_{k,j} \xrightarrow{P} 0, \\
\boldsymbol{\xi}_i^\top \mathbf{w}^\star = \sum_{j=1}^{d} \xi_{i,j} w_j^\star \xrightarrow{P} 0, \quad \boldsymbol{\xi}_i^\top \mathbf{w}_0 = \sum_{j=1}^{d} \xi_{i,j} w_{0j} \xrightarrow{P} 0,
\end{aligned} \tag{18}$$

where the convergence in probability in each one of the cases follows by the uniform integrability of teacher correlations and from independence across coordinates with $\mathbb{E}[\boldsymbol{\xi}_i] = \mathbb{E}[\boldsymbol{\xi}_k] = 0$, and $\text{Cov}(\boldsymbol{\xi}_i) = \mathbf{\Sigma}$. Using the identities in Equation (18), we evaluate the resulting expression by separating the analysis into three cases.

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

$$\left\|\mathbf{w}_T^{\text{esgd}} - \mathbf{w}_i^\star\right\|^2 \xrightarrow{P} \left(a'^T - 2a'^{T-1}b'\right)\text{Tr}\left(\mathbf{\Sigma}\right) + a'^T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2 + c'v_z\frac{1-a'^T}{1-a'} + 2\text{Tr}\left(\mathbf{\Sigma}\right)\left(\sum_{k=2}^{T} a'^{T-k+1} - b'\sum_{k=3}^{T} a'^{T-k+1}\right)$$

$$= \left(a'^T - 2a'^{T-1}b'\right)\text{Tr}\left(\mathbf{\Sigma}\right) + a'^T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2 + c'v_z\frac{1-a'^T}{1-a'} + 2\text{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{a'\left(1-a'^{T-1}\right)}{1-a'} - b'\frac{a'\left(1-a'^{T-2}\right)}{1-a'}\right). \tag{19}$$

When $1 < i < T$:

$$\left\|\mathbf{w}_T^{\text{esgd}} - \mathbf{w}_i^\star\right\|^2 \xrightarrow{P} \left(a'^T - 2a'^{T-1}b'\right)\text{Tr}\left(\mathbf{\Sigma}\right) + a'^T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2 + c'v_z\frac{1-a'^T}{1-a'} + 2\text{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{a'\left(1-a^{T-1}\right)}{1-a'} - b'\frac{a'\left(1-a^{T-2}\right)}{1-a'} + 1\right) + 2\left(b'^{T-i+1} - b'^{T-i} - b'\right)\text{Tr}\left(\mathbf{\Sigma}\right). \tag{20}$$

When $i = 1$:

$$\left\|\mathbf{w}_T^{\text{esgd}} - \mathbf{w}_i^\star\right\|^2 \xrightarrow{P} \left(a'^T - 2a'^{T-1}b'\right)\text{Tr}\left(\mathbf{\Sigma}\right) + a'^T\|\mathbf{w}^\star - \mathbf{w}_0\|^2 + 2b'^T\text{Tr}\left(\mathbf{\Sigma}\right) + c'v_z\frac{1-a'^T}{1-a'} + 2\text{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{a'\left(1-a'^{T-1}\right)}{1-a'} - b'\frac{a'\left(1-a'^{T-2}\right)}{1-a'} + 1\right) + 2\left(b'^T - b'^{T-1} - b'\right)\text{Tr}\left(\mathbf{\Sigma}\right). \tag{21}$$

Using Equations (19) to (21) when summing over all cases,

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

$$\left(a'^T - 2a'^{T-1}b'\right)\text{Tr}\left(\mathbf{\Sigma}\right) + a^T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2 + 2\text{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{a\left(1-a'^{T-1}\right)}{1-a'} - b'\frac{a'\left(1-a'^{T-2}\right)}{1-a'}\right) +$$

$$cv_z\frac{1-a'^T}{1-a'} + \frac{2}{T}\left(2b'^T + T - 1 - Tb'\right)\text{Tr}\left(\mathbf{\Sigma}\right) \tag{22}$$

$$= c'v_z\frac{1-a'^T}{1-a'} + a'^T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2 + 2\text{Tr}(\mathbf{\Sigma})\left(1 - b' + \frac{2b'^T - 1}{T} + \frac{a'}{1-a'}\left(1 - a'^{T-1} - b\left(1 - a'^{T-2}\right)\right) + \frac{a^{T-1}\left(a' - 2b\right)}{2}\right)$$

$$= c'v_z\frac{1-a^T}{1-a} + a^T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2 + 2\text{Tr}(\mathbf{\Sigma})\left(1 - b' + \frac{2b'^T - 1}{T} + \frac{a'}{1-a'}\left(1 - b' + a'^{T-2}\left(b' - a'\right)\right) + \frac{a'^{T-1}\left(a' - 2b'\right)}{2}\right)$$

$$= cv_z\frac{1-a'^T}{1-a'} + a'^T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2 + 2\text{Tr}(\mathbf{\Sigma})\left(\frac{(1-b')(1-a')}{1-a'} + \frac{a'(1-b')}{1-a'} + \frac{2b'^T - 1}{T} + \frac{a'^{T-1}}{1-a'}\left(b' - a'\right) + \frac{a'^{T-1}\left(a' - 2b'\right)}{2}\right)$$

$$= cv_z\frac{1-a'^T}{1-a'} + a'^T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2 + 2\text{Tr}(\mathbf{\Sigma})\left(\frac{1-b'}{1-a'} + \frac{2b'^T - 1}{T} + \frac{a'^T\left(-a' + 2b' - 1\right)}{2\left(1-a'\right)}\right).$$

□

C.3 Proof of Theorem 2: Optimal Step Size in Early-Stopped GD

Recall Theorem 2. Under the i.i.d. teacher setting (Assumption 1) with a nontrivial mean teacher $\mathbf{w}^\star \neq \mathbf{w}_0$[5], as $n, d \to \infty$ the optimal fixed step size $\gamma^\star$ that minimizes the generalization loss of the early-stopped GD algorithm (Algorithm 2) after T tasks satisfies $\gamma^\star \asymp \frac{\ln T}{NT}$. Specifically, for all sufficiently large $T > T_0 = \Theta\big(\frac{1}{\epsilon^2 \ln^2(1/\epsilon)}\big)$, the optimal parameters satisfy

$$\frac{1}{1+\epsilon}\frac{\ln(T)+\ln(\Gamma)}{2Nv_xT} < \gamma^\star < \frac{1}{1-\epsilon}\frac{\ln(T)+\ln(\Gamma)}{2Nv_xT},$$

where $\Gamma = \frac{4\alpha v_x \|\mathbf{w}^\star - \mathbf{w}_0\|^2}{v_z + v_x \mathrm{Tr}(\mathbf{\Sigma})(\alpha+1)}$. In the degenerate noiseless case of $\mathrm{Tr}(\mathbf{\Sigma}) = v_z = 0$, the optimal choice is any $\gamma^\star \in \left(0, 2v_x^{-1}\left(1+\alpha^{-1/2}\right)^{-2}\right)$ and $N \to \infty$.

Moreover, if $\gamma \asymp \frac{\ln T}{NT}$ the resulting generalization loss of the early-stopped GD matches that of optimally tuned L2 regularization up to $O(1/T)$ factor.

Proof. The proof of the main theorem proceeds as follows. We first show that no optimal regularizer can be obtained in the regime where T is 'significantly' larger than λ. Next, by employing exact Laurent analysis in the regime where T increases more slowly than λ, we derive the approximately optimal value for the step size. Finally, in a suitable regime, we find a close connection between the ridge objective and the objective of early-stopped gradient descent.

Recall Theorem 1,

$$G_T^{\mathrm{esgd}} \xrightarrow{P} v_z c' \frac{1-a'^T}{1-a'} + \frac{1}{T}\sum_{i=1}^{T} \|\mathbf{w}_T^\star - \mathbf{w}_i^\star\|^2 + 2\big(\mathbf{w}_T^\star - \bar{\mathbf{w}}^\star\big)^\top \sum_{i=1}^{T} b'^{T-i+1}\big(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\big)$$
$$+ \sum_{i=1}^{T}\sum_{i'=1}^{T} a'^{T-\max(i,i')+1} b'^{|i-i'|}\big(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\big)^\top\big(\mathbf{w}_{i'-1}^\star - \mathbf{w}_{i'}^\star\big),$$

For convenience, throughout the proof we use $G_T^{\mathrm{esgd}}, G_T^{\mathrm{ridge}}$ to denote their corresponding asymptotic deterministic equivalent, to which they converge in probability as $n, d \to \infty$.

To further simplify this expression in the i.i.d. teacher setting we use Corollary 1. We conclude that under early-stopped GD (Algorithm 2) the expected generalization after $T > 1$ steps satisfies,

$$G_T^{\mathrm{esgd}} = v_z c' \frac{1-a'^T}{1-a'} + a'^T \|\mathbf{w}^\star - \mathbf{w}_0\|^2 + 2\mathrm{Tr}(\mathbf{\Sigma})\left(\frac{1-b'}{1-a'} + \frac{2b'^T - 1}{T} + \frac{a'^T(-a'+2b'-1)}{2(1-a')}\right). \tag{23}$$

By taking the Taylor expansion around $N\gamma \to 0$ for a', b', c', as in Equation (16), we obtain,

$$\begin{aligned} a'(\gamma, N) &= 1 - 2Nv_x\gamma + N(2N-1)v_x^2\left(1+\frac{1}{\alpha}\right)\gamma^2 + O\big((N\gamma)^3\big), \\ b'(\gamma, N) &= 1 - Nv_x\gamma + \frac{N(N-1)}{2}v_x^2\left(1+\frac{1}{\alpha}\right)\gamma^2 + O\big((N\gamma)^3\big), \\ c'(\gamma, N) &= \frac{N^2}{\alpha}v_x\gamma^2 + O\big((N\gamma)^3\big). \end{aligned} \tag{24}$$

We first show that the optimal step size γ and number of gradient steps N, can not be in any region where $\frac{1}{N\gamma} := \lambda \in o(T^\rho)$, for all $\rho < 1$. First, we consider the limit of the objective (Equation (23)), i.e.

$$\lim_{T\to\infty} G_T^{\mathrm{esgd}} = c' v_z \frac{1}{1-a'} + 2\mathrm{Tr}(\mathbf{\Sigma})\left(\frac{1-b'}{1-a'}\right),$$

[5] Having $\mathbf{w}^\star = \mathbf{w}_0$ would directly imply optimal step size $\gamma^\star \to 0$.

Taking the derivative of the limiting expression,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\lim_{T\to\infty} G_T^{\mathrm{esgd}} = \underbrace{\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(c'\right)v_z\frac{1}{1-a'} + \frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{1}{1-a'}c'v_z}_{\text{term 1}} + \underbrace{2\mathrm{Tr}\left(\mathbf{\Sigma}\right)\left(\frac{1-b'}{(1-a')^2}\frac{\mathrm{d}}{\mathrm{d}\lambda}a' - \frac{1}{1-a'}\frac{\mathrm{d}}{\mathrm{d}\lambda}b'\right)}_{\text{term 2}}. \tag{25}$$

Using the Taylor expansion in Equation (24), and since a', b', c' are analytic at $\gamma = 0$ (all are polynomial), then using Lemma 6 and the relation $\lambda = \frac{1}{N\gamma}$, term 1 becomes,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(c'\right)v_z\frac{1}{1-a'} + \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{1-a'}\right)c'v_z \\
&= \left(\frac{-2v_x}{\lambda^3\alpha} + O(\lambda^{-4})\right)\frac{v_z}{\frac{2v_x}{\lambda}+O(\lambda^{-2})} + \frac{\mathrm{d}a'}{\mathrm{d}\lambda}\frac{\mathrm{d}}{\mathrm{d}a'}\left(\frac{1}{1-a'}\right)c'v_z \\
&= \left(\frac{-2v_x}{\lambda^3\alpha} + O(\lambda^{-4})\right)\frac{v_z}{\frac{2v_x}{\lambda}+O(\lambda^{-2})} + \left(\frac{2v_x}{\lambda^2}+O(\lambda^{-3})\right)\frac{1}{\left(\frac{2v_x}{\lambda}+O(\lambda^{-2})\right)^2}\left(\frac{v_x}{\lambda^2\alpha}+O(\lambda^{-3})\right)v_z \\
&= \left(\frac{-v_z}{\lambda^2\alpha}+O(\lambda^{-3})\right) + \left(\frac{1}{2v_x}+O(\lambda^{-2})\right)\left(\frac{v_x}{\lambda^2\alpha}+O(\lambda^{-3})\right)v_z \\
&= \frac{-v_z}{\lambda^2\alpha} + \frac{v_z}{2\lambda^2\alpha} + O\left(\lambda^{-3}\right) = \frac{-v_z}{2\lambda^2\alpha} + O\left(\lambda^{-3}\right).
\end{aligned} \tag{26}$$

And term 2 becomes,

$$\begin{aligned}
&2\mathrm{Tr}(\mathbf{\Sigma})\left(\frac{1-b'}{(1-a')^2}\frac{\mathrm{d}a'}{\mathrm{d}\lambda} - \left(\frac{1}{1-a'}\right)\frac{\mathrm{d}b'}{\mathrm{d}\lambda}\right) \\
&= 2\mathrm{Tr}(\mathbf{\Sigma})\Bigg(\left[\frac{v_x}{\lambda} - \frac{N-1}{2N}v_x^2\left(1+\frac{1}{\alpha}\right)\frac{1}{\lambda^2} + O(\lambda^{-3})\right] \\
&\qquad \left[\frac{\lambda^2}{4v_x^2} + \frac{2N-1}{4N}\left(1+\frac{1}{\alpha}\right)\frac{\lambda}{v_x} + O(1)\right]\left[\frac{2v_x}{\lambda^2} - \frac{2(2N-1)}{N}v_x^2\left(1+\frac{1}{\alpha}\right)\frac{1}{\lambda^3} + O(\lambda^{-4})\right] \\
&\qquad - \left[\frac{\lambda}{2v_x} + \frac{2N-1}{4N}\left(1+\frac{1}{\alpha}\right) + O(\lambda^{-1})\right]\left[\frac{v_x}{\lambda^2} - \frac{N-1}{N}v_x^2\left(1+\frac{1}{\alpha}\right)\frac{1}{\lambda^3} + O(\lambda^{-4})\right]\Bigg) \\
&= 2\mathrm{Tr}(\mathbf{\Sigma}) \\
&\quad \Bigg(\left[\frac{1}{2\lambda} + \frac{1}{\lambda^2}\left(-\frac{N-1}{4N}v_x\left(1+\frac{1}{\alpha}\right) + \frac{2N-1}{4N}v_x\left(1+\frac{1}{\alpha}\right) - \frac{2N-1}{2N}v_x\left(1+\frac{1}{\alpha}\right)\right)\right] \\
&\qquad - \left[\frac{1}{2\lambda} + \frac{1}{\lambda^2}\left(\frac{2N-1}{2N}v_x\left(1+\frac{1}{\alpha}\right) - \frac{N-1}{2N}v_x\left(1+\frac{1}{\alpha}\right)\right)\right] + O(\lambda^{-3})\Bigg) \\
&= 2\mathrm{Tr}(\mathbf{\Sigma})\left(\cancel{\frac{1}{2\lambda}} - \cancel{\frac{1}{2\lambda}} + \frac{1}{\lambda^2}\left(-\frac{N-1}{4N} + \frac{2N-1}{2N} - \frac{2N-1}{2N} - \frac{2N-1}{4N} + \frac{N-1}{2N}\right)v_x\left(1+\frac{1}{\alpha}\right) + O(\lambda^{-3})\right) \\
&= 2\mathrm{Tr}(\mathbf{\Sigma})\left(-\frac{1}{4}v_x\left(1+\frac{1}{\alpha}\right)\frac{1}{\lambda^2} + O(\lambda^{-3})\right) = -\frac{\mathrm{Tr}(\mathbf{\Sigma})}{2}v_x\left(1+\frac{1}{\alpha}\right)\frac{1}{\lambda^2} + O(\lambda^{-3}).
\end{aligned} \tag{27}$$

Therefore, the derivative is strictly negative (for all sufficiently large λ) and is of order $\Theta(\lambda^{-2})$. We aim to show that the distance between $\frac{\mathrm{d}}{\mathrm{d}\lambda}G_T^{\mathrm{esgd}}$ is exponentially small in the aforementioned regime, and thereby showing that $\frac{\mathrm{d}}{\mathrm{d}\lambda}G_T^{\mathrm{esgd}}$ is of the same order as $\frac{\mathrm{d}}{\mathrm{d}\lambda}\lim_{T\to\infty} G_T^{\mathrm{esgd}}$. First note that,

$$\begin{aligned}
&G_T^{\mathrm{esgd}} - \lim_{T\to\infty} G_T^{\mathrm{esgd}} \\
&= v_z c'\frac{-a'^T}{1-a'} + a'^T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2 + 2\mathrm{Tr}(\mathbf{\Sigma})\left(\frac{2b'^T-1}{T} + \frac{a'^T\left(-a'+2b'-1\right)}{2(1-a')}\right).
\end{aligned} \tag{28}$$

Then by taking the derivative,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[G_T^{\mathrm{esgd}} - \lim_{T\to\infty} G_T^{\mathrm{esgd}}\right] = -\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{a'^T}{1-a'}v_z c' - \frac{a'^T}{1-a'}v_z\frac{\mathrm{d}}{\mathrm{d}\lambda}c' + \frac{\mathrm{d}a}{\mathrm{d}\lambda}Ta'^{T-1}\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2$$
$$= \frac{Ta'^{T-1}(1-a') + a'^T}{(1-a')^2}\frac{\mathrm{d}a'}{\mathrm{d}\lambda}v_z c' - \frac{a'^T}{1-a'}v_z\frac{\mathrm{d}c'}{\mathrm{d}\lambda} + \frac{\mathrm{d}a'}{\mathrm{d}\lambda}Ta'^{T-1}\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2$$

$$+ 2\mathrm{Tr}(\mathbf{\Sigma})\left(2b'^{T-1}\frac{\mathrm{d}b'}{\mathrm{d}\lambda} + \frac{1}{2(1-a')}\left[Ta'^{T-1}\frac{\mathrm{d}a'}{\mathrm{d}\lambda}\big(-a'+2b'-1\big) + a'^T\left(-\frac{\mathrm{d}a'}{\mathrm{d}\lambda} + 2\frac{\mathrm{d}b'}{\mathrm{d}\lambda}\right)\right]\right.$$
$$\left. + \frac{a'^T\big(-a'+2b'-1\big)}{2(1-a')^2}\frac{\mathrm{d}a'}{\mathrm{d}\lambda}\right)$$

Since a', b', c' and their corresponding derivatives are bounded (for all sufficiently large λ) by Equation (24) and Lemma 6, and $a', b' = 1 - \Theta(\frac{1}{\lambda})$ by Equation (24), then $a'^T, b'^T = \left(1 - \Theta\left(\frac{1}{\lambda}\right)\right)^T \leq \exp\left(-\beta\frac{T}{\lambda}\right)$, for some $\beta > 0$ (uniformly). So,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[G_T^{\mathrm{esgd}} - \lim_{T\to\infty} G_T^{\mathrm{esgd}}\right] \leq \exp\left(-\beta\frac{T}{\lambda}\right)\cdot \mathrm{Poly}(\lambda^{-1}, T).$$

where $\mathrm{Poly}(\lambda^{-1}, T)$ denotes a polynomial in (λ^{-1}, T) (with bounded degree). Assume $\lambda = O(T^\rho)$ for some $\rho < 1$. Then $\exp(-\beta T/\lambda)\mathrm{Poly}(\lambda^{-1}, T) = o(\lambda^{-2})$. Since $\frac{\mathrm{d}}{\mathrm{d}\lambda}\lim_{T\to\infty} G_T^{\mathrm{esgd}} = \Omega(\lambda^{-2})$, by Equation (26), it follows that $\frac{\mathrm{d}}{\mathrm{d}\lambda}G_T^{\mathrm{esgd}} = \frac{\mathrm{d}}{\mathrm{d}\lambda}\lim_{T\to\infty} G_T^{\mathrm{esgd}} + o(\lambda^{-2})$, and in particular the finite-T derivative has the same leading-order behavior.

Now we show, that if $\frac{1}{N\gamma} \in \omega\left(T^{\frac{1}{2}}\right)$, then the difference between the objectives' derivative is not 'large', i.e. $\left|\frac{\mathrm{d}}{\mathrm{d}\lambda}G_T^{\mathrm{esgd}} - \frac{\mathrm{d}}{\mathrm{d}\lambda}G_T^{\mathrm{ridge}}\right| = o\left(\frac{\mathrm{d}}{\mathrm{d}\lambda}G_T^{\mathrm{ridge}}\right)$ (almost everywhere). We develop the corresponding Laurent expansion using the relation of $\frac{1}{N\gamma} = \lambda$ and Karpel et al. (2026, Lemma D.18),

$$\begin{aligned}
&G_T^{\mathrm{esgd}} - \lim_{T\to\infty} G_T^{\mathrm{esgd}} \\
&= a'^T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2 + v_z c'\frac{-a'^T}{1-a'} + 2\mathrm{Tr}(\mathbf{\Sigma})\left(\frac{2b'^T - 1}{T} + \frac{a'^T(-a'+2b'-1)}{2(1-a')}\right) \\
&= \exp\left(-\frac{2Tv_x}{\lambda}\right)\left(1 + O\left(T\lambda^{-2}\right)\right)\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2 \\
&\quad - v_z\left(\frac{\lambda}{2v_x} + O(1)\right)\left(\frac{v_x}{\alpha\lambda^2} + O\left(\lambda^{-3}\right)\right)\exp\left(-\frac{2Tv_x}{\lambda}\right)\left(1 + O\left(T\lambda^{-2}\right)\right) \\
&\quad + \frac{2\mathrm{Tr}(\mathbf{\Sigma})}{T}\left(2\exp\left(-\frac{Tv_x}{\lambda}\right)\left(1 + O\left(T\lambda^{-2}\right)\right) - 1\right) \\
&\quad + \mathrm{Tr}(\mathbf{\Sigma})\left(\exp\left(-\frac{2Tv_x}{\lambda}\right)\left(1 + O\left(T\lambda^{-2}\right)\right)\right)O\left(\lambda^{-1}\right).
\end{aligned} \tag{29}$$

An analogous statement holds for $G_T^{\mathrm{ridge}} - \lim_{T\to\infty} G_T^{\mathrm{ridge}}$. In this case, one replaces a', b', c' by a, b, c, respectively, because the corresponding coefficients admit identical expansions[6]. Furthermore, the limiting objective's derivative attains the following expansion

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\lim_{T\to\infty} G_T^{\mathrm{ridge}} = -\frac{v_x}{2}\left(\frac{v_z}{v_x} + \mathrm{Tr}(\mathbf{\Sigma})(\alpha+1)\right)\lambda^{-2} + O(\lambda^{-3}),$$

[6]While their expression is stated for the expectation (including expectation on the features and labels), their analysis can be naturally extended to yield the corresponding convergence-in-probability statement under the same asymptotic regime.

by Karpel et al. (2026, Lemma E.3), which coincides with the expansion we had for $\frac{\mathrm{d}}{\mathrm{d}\lambda}\lim_{T\to\infty} G_T^{\mathrm{esgd}}$ (under the substitution of $\lambda\alpha$, see Lemma 5 and Equation (26)). Therefore, G_T^{ridge} admits the same expansion as G_T^{esgd}, under the substitution of $N\gamma = \frac{1}{\lambda}$.

So far our analysis accommodated large enough λ (or equivalently, small enough $N\gamma$), to deal with the case of $\gamma \in \Theta(1)$ but is small enough to ensure convergence (i.e. $\gamma \in \left(0, 2v_x^{-1}\left(1+\alpha^{-1/2}\right)^{-2}\right)$), it is enough to show that these points can not be the optimal of the objective $\lim_{T\to\infty} G_T^{\mathrm{esgd}}$, that is, there exists a point which attains a strictly smaller error, then for all sufficiently large T, it will remain non-optimal for the finite T objective G_T^{esgd}, since the two objectives can be made arbitrary close (see Equation (28)). Let $\gamma \in \Theta(1)$ where $\gamma \in \left(0, 2v_x^{-1}\left(1+\alpha^{-1/2}\right)^{-2}\right)$.

Recall that a' and b' are the corresponding $2N$'th and N'th moment of the MP law. Therefore, by Jensen's inequality $a' \geq b'^2$, and,

$$\frac{1-b'}{1-a'} \geq \frac{1-b'}{1-b'^2} = \frac{1}{1+b'}.$$

This yields,

$$\lim_{T\to\infty} G_T^{\mathrm{esgd}} \geq v_z c' \frac{1}{1-a'} + \frac{2}{1+b'}\mathrm{Tr}(\mathbf{\Sigma}) > \mathrm{Tr}(\mathbf{\Sigma}).$$

This follows since $a', b' < 1$ in the range $\gamma \in \left(0, 2v_x^{-1}\left(1+\alpha^{-1/2}\right)^{-2}\right)$ (and $v_z, c' > 0$ always).

Denote by $\Delta > 0$ as $\lim_{T\to\infty} \mathbb{E}G_T^{\mathrm{esgd}} - \mathrm{Tr}(\mathbf{\Sigma}) = \Delta$. As $\lambda \to \infty$ we have $\lim_{T\to\infty} G_T^{\mathrm{esgd}} = \mathrm{Tr}\left(\mathbf{\Sigma}\right)$ (as $\frac{1-b'}{1-a'} \to \frac{1}{2}$, and $\frac{c'}{1-a'} \to 0$ which follows immediately by the Taylor approximations of a', b', c'), take $\tilde{\lambda}$ large enough (or $N\tilde{\gamma}$ small enough) such that $\left|\lim_{T\to\infty} G_T^{\mathrm{esgd}} - \mathrm{Tr}\left(\mathbf{\Sigma}\right)\right| < \frac{\Delta}{2}$, then

$$G_T^{\mathrm{esgd}}(\gamma) - G_T^{\mathrm{esgd}}(\tilde{\gamma}) > \left(\lim_{T\to\infty} G_T^{\mathrm{esgd}}(\gamma) - \mathrm{Tr}\left(\mathbf{\Sigma}\right)\right) - \left(\lim_{T\to\infty} G_T^{\mathrm{esgd}}(\tilde{\gamma}) - \mathrm{Tr}\left(\mathbf{\Sigma}\right)\right) \geq \frac{\Delta}{2} > 0$$

therefore there exists λ sufficiently large that attains a smaller error.

If both $\mathrm{Tr}(\mathbf{\Sigma}) = 0$ and $v_z = 0$, then by Theorem 1,

$$G_T^{\mathrm{esgd}} = \left\|\mathbf{w}_T - \mathbf{w}^\star\right\|^2 = (a')^T \left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2.$$

When $N \to \infty$ and γ is finite (but ensures the convergence of early-stopped GD algorithm, i.e. any $\gamma \in \left(0, 2v_x^{-1}\left(1+\alpha^{-1/2}\right)^{-2}\right)$) then $a' = \lim_{N\to\infty} \frac{1}{d}\mathrm{Tr}\left(\mathbf{I} - \gamma\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)^{2N} = (1-\alpha)_+$ since for every nonzero eigenvalue λ_i of $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$, we have $(1-\gamma\lambda_i)^{2N} \to 0$, while zero eigenvalues remain equal to one. Hence a' attains its minimal possible value, corresponding to the irreducible nullspace component.

Otherwise, the optimal step size $\frac{1}{N\gamma^\star} = \Theta(T/\ln T)$. Recall Equation (28),

$$G_T^{\mathrm{esgd}} - \lim_{T\to\infty} G_T^{\mathrm{esgd}} = \underbrace{\frac{2\mathrm{Tr}\left(\mathbf{\Sigma}\right)}{T}\left(2(b')^T - 1\right)}_{\text{term 1}} + \underbrace{\mathrm{Tr}\left(\mathbf{\Sigma}\right)(a')^T\left(1 - 2\frac{1-b'}{1-a'}\right)}_{\text{term 2}}$$
$$+ \underbrace{(a')^T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2 + v_z c'\frac{-(a')^T}{1-a'}}_{\text{term 3}}.$$

We develop a Laurent expansion for each term. Term 1 becomes,

$$\frac{2\mathrm{Tr}(\mathbf{\Sigma})}{T}\left(2(b')^T - 1\right) = \frac{2\mathrm{Tr}(\mathbf{\Sigma})}{T}\left(2\exp\left(-\frac{v_x T}{\lambda}\right)\left(1 + O(T\lambda^{-2})\right) - 1\right). \tag{30}$$

Term 2 is,

$$\begin{aligned}
&\operatorname{Tr}(\boldsymbol{\Sigma})\,(a')^T\left[1-2\left(\frac{1-b'}{1-a'}\right)\right]\\
&=\operatorname{Tr}(\boldsymbol{\Sigma})\left(\exp\left(-\frac{2v_xT}{\lambda}\right)\left(1+O\left(T\lambda^{-2}\right)\right)\right)O\left(\lambda^{-1}\right).
\end{aligned}\tag{31}$$

Term 3 is,

$$\begin{aligned}
&(a')^T\|\mathbf{w}^\star-\mathbf{w}_0\|^2+v_zc'\frac{-(a')^T}{1-a'}\\
&=\exp\left(-\frac{2v_xT}{\lambda}+O\left(T\lambda^{-2}\right)\right)\|\mathbf{w}^\star-\mathbf{w}_0\|^2\\
&\quad-\frac{\lambda v_z}{4\alpha v_x^2}\left(\frac{2\alpha}{\left(\frac{\lambda\alpha}{v_x}+\alpha+1\right)^2}+O\left(\lambda^{-4}\right)\right)\exp\left(-\frac{2v_xT}{\lambda}+O\left(T\lambda^{-2}\right)\right)\\
&=\exp\left(-\frac{2v_xT}{\lambda}\right)\left(1+O\left(T\lambda^{-2}\right)\right)\|\mathbf{w}^\star-\mathbf{w}_0\|^2\\
&\quad-\frac{v_z}{2\alpha\lambda}\left(1+O(\lambda^{-1})\right)\exp\left(-\frac{2v_xT}{\lambda}\right)\left(1+O\left(T\lambda^{-2}\right)\right).
\end{aligned}\tag{32}$$

Using Lemma 6, we differentiate with respect to λ, term by term.

Differentiating Equation (30),

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{2\mathrm{Tr}(\boldsymbol{\Sigma})}{T}\left(2(b')^T-1\right)\\
&=\frac{2\mathrm{Tr}(\boldsymbol{\Sigma})}{T}\left(2\frac{v_xT}{\lambda^2}\exp\left(-\frac{v_xT}{\lambda}\right)\left(1+O\left(T\lambda^{-2}\right)\right)\right)\\
&\quad+\frac{2\mathrm{Tr}(\boldsymbol{\Sigma})}{T}\left(2\exp\left(-\frac{v_xT}{\lambda}\right)O\left(T\lambda^{-3}\right)\right).
\end{aligned}\tag{33}$$

Differentiating Equation (31),

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\lambda}\operatorname{Tr}(\boldsymbol{\Sigma})\,(a')^T\left[1-2\left(\frac{1-b'}{1-a'}\right)\right]\\
&=\operatorname{Tr}(\boldsymbol{\Sigma})\left(\frac{2v_xT}{\lambda^2}\exp\left(-\frac{2v_xT}{\lambda}\right)\left(1+O\left(T\lambda^{-2}\right)\right)\right.\\
&\qquad\left.+\exp\left(-\frac{2v_xT}{\lambda}\right)O\left(T\lambda^{-3}\right)\right)O(\lambda^{-1})\\
&\quad+\operatorname{Tr}(\boldsymbol{\Sigma})\left(\exp\left(-\frac{2v_xT}{\lambda}\right)\left(1+O\left(T\lambda^{-2}\right)\right)\right)O(\lambda^{-2}).
\end{aligned}\tag{34}$$

Differentiating Equation (32),

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[(a')^T\|\mathbf{w}^\star-\mathbf{w}_0\|^2+f\frac{-(a')^T}{1-a'}\right]\\
&=\frac{v_z}{2\alpha\lambda^2}\left(1+O(\lambda^{-1})\right)\exp\left(-\frac{2v_xT}{\lambda}\right)\left(1+O\left(T\lambda^{-2}\right)\right)\\
&\quad-\frac{v_z}{2\alpha\lambda}\left(1+O(\lambda^{-1})\right)\left(\frac{2v_xT}{\lambda^2}\exp\left(-\frac{2v_xT}{\lambda}\right)\left(1+O\left(T\lambda^{-2}\right)\right)+\exp\left(-\frac{2v_xT}{\lambda}\right)O\left(T\lambda^{-3}\right)\right)\\
&\quad+\frac{2v_xT}{\lambda^2}\exp\left(-\frac{2v_xT}{\lambda}\right)\left(1+O\left(T\lambda^{-2}\right)\right)\|\mathbf{w}^\star-\mathbf{w}_0\|^2+\exp\left(-\frac{2v_xT}{\lambda}\right)O\left(T\lambda^{-3}\right)\|\mathbf{w}^\star-\mathbf{w}_0\|^2.
\end{aligned}\tag{35}$$

Retaining only the leading-order terms,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\mathbb{E}\left[G_T^{\mathrm{esgd}}\right]-\lim_{T\to\infty}\mathbb{E}\left[G_T^{\mathrm{esgd}}\right]\right]=\frac{2v_xT}{\lambda^2}\exp\left(-q\right)\left\|\mathbf{w}^\star-\mathbf{w}_0\right\|^2+\frac{4\mathrm{Tr}(\mathbf{\Sigma})}{\lambda^2}v_x\exp\left(-\frac{q}{2}\right),$$

where $q=\frac{2v_xT}{\lambda}$. Taking the Laurent expansion in Equation (25), yields

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

Therefore, the stationary point satisfies

$$\begin{aligned}\frac{2v_xT}{\lambda^2}\exp\left(-q\right)\left\|\mathbf{w}^\star-\mathbf{w}_0\right\|^2+\frac{4\mathrm{Tr}(\mathbf{\Sigma})}{\lambda^2}v_x\exp\left(-\frac{q}{2}\right)&\\ =\frac{1}{2}\left(\frac{v_z}{\alpha}+v_x\mathrm{Tr}(\mathbf{\Sigma})\left(1+\frac{1}{\alpha}\right)\right)\lambda^{-2}+O\left(\lambda^{-3}\right).&\end{aligned}$$

Multiplying by $\frac{2\lambda^2}{v_xv_z}$, we define

$$\widetilde{H}(q)=\frac{4T\|\mathbf{w}^\star-\mathbf{w}_0\|^2}{v_z}e^{-q}+\frac{8\mathrm{Tr}(\mathbf{\Sigma})}{v_z}e^{-q/2}+z,$$

where $z=-\frac{1}{v_z}\left(\frac{v_z}{v_x\alpha}+\mathrm{Tr}(\mathbf{\Sigma})\left(1+\frac{1}{\alpha}\right)\right)$. The derivative satisfies

$$\widetilde{H}'(q)=-\frac{4T\|\mathbf{w}^\star-\mathbf{w}_0\|^2}{v_z}e^{-q}-\frac{4\mathrm{Tr}(\mathbf{\Sigma})}{v_z}e^{-q/2}<0.$$

Hence $\widetilde{H}$ is strictly decreasing. Moreover, if $q\in o(\ln T)$, then

$$\widetilde{H}(q)=\frac{4T\|\mathbf{w}^\star-\mathbf{w}_0\|^2}{v_z}e^{-q}+z+o(1),$$

which is positive for all sufficiently large T. On the other hand, if $q\in\omega(\ln T)$, then $\widetilde{H}(q)=z+o(1)<0$. Therefore, it suffices to identify $q^\star$ satisfying $\widetilde{H}(q^\star)\approx 0$. Retaining only the leading-order term in the regime $q^\star\in\Theta(\ln T)$, we obtain $\frac{4\alpha T\|\mathbf{w}^\star-\mathbf{w}_0\|^2}{v_z}e^{-q^\star}\approx -z$, which gives $q^\star\approx\ln\frac{4T\|\mathbf{w}^\star-\mathbf{w}_0\|^2}{-v_zz}$. Let $\epsilon>0$. Then,

$$\widetilde{H}\left(\frac{q^\star}{1+\epsilon}\right)=-z\left(\frac{4T\|\mathbf{w}^\star-\mathbf{w}_0\|^2}{-v_zz}\right)^{\frac{\epsilon}{1+\epsilon}}+z+o(1)\xrightarrow[T\to\infty]{}+\infty,\tag{36}$$

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

$$\widetilde{H}\left(\frac{q^\star}{1-\epsilon}\right)=-z\left(\frac{4T\|\mathbf{w}^\star-\mathbf{w}_0\|^2}{-v_zz}\right)^{-\frac{\epsilon}{1-\epsilon}}+z+o(1)\xrightarrow[T\to\infty]{}z<0,\tag{37}$$

and therefore $\widetilde{H}\left(\frac{q^\star}{1-\epsilon}\right)<0$ for all sufficiently large T. This proves that the optimal parameters satisfy

$$\frac{1}{1+\epsilon}\frac{\ln\left(T\right)+\ln\left(\frac{4\alpha v_x\|\mathbf{w}^\star-\mathbf{w}_0\|^2}{v_z+v_x\mathrm{Tr}(\mathbf{\Sigma})(\alpha+1)}\right)}{2Nv_xT}<\gamma^\star<\frac{1}{1-\epsilon}\frac{\ln\left(T\right)+\ln\left(\frac{4\alpha v_x\|\mathbf{w}^\star-\mathbf{w}_0\|^2}{v_z+v_x\mathrm{Tr}(\mathbf{\Sigma})(\alpha+1)}\right)}{2Nv_xT}.$$

To satisfy $\tilde{H}\left(\frac{1}{1+\epsilon}q^\star\right) > 0$, we begin with Equation (36). In this case, the non-leading terms contribute $O\left(T^{-\frac{1}{2(1+\epsilon)}}\right)$ — as $\exp(-q/2)$, the dominating term becomes $O\left(T^{-\frac{1}{2(1+\epsilon)}}\right)$.

$$
\begin{aligned}
&- z \cdot \left(\left(\frac{4T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2}{-v_z z}\right)^{\frac{\epsilon}{1+\epsilon}} - 1\right) + O\left(T^{-\frac{1}{2(1+\epsilon)}}\right) \\
&= -z \cdot \left(\exp\left(\frac{\epsilon}{1+\epsilon}\ln\left(\frac{4T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2}{-v_z z}\right)\right) - 1\right) + O\left(T^{-\frac{1}{2(1+\epsilon)}}\right) \\
\left[e^x - 1 \geq x\right] &\geq -z \cdot \frac{\epsilon}{1+\epsilon}\ln\left(\frac{4T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2}{-v_z z}\right) + O\left(T^{-\frac{1}{2(1+\epsilon)}}\right) \\
&\gtrsim -z \cdot \frac{\epsilon}{1+\epsilon}\ln(T) + O\left(T^{-\frac{1}{2(1+\epsilon)}}\right).
\end{aligned}
$$

There exists a constant C_1 such that for all sufficiently large T we have $O\left(T^{-\frac{1}{2(1+\epsilon)}}\right) \leq C_1 T^{-\frac{1}{2(1+\epsilon)}}$, and hence it suffices to require $-z\frac{\epsilon}{1+\epsilon}\ln T \gtrsim C_1 T^{-\frac{1}{2(1+\epsilon)}}$. This holds for $T > T_0 = \Theta\left(\frac{1}{\epsilon^2 \ln^2(1/\epsilon)}\right)$, with a sufficiently large constant, since at this scale $\ln T_0 \asymp \ln(1/\epsilon)$ and $T_0^{-\frac{1}{2(1+\epsilon)}} \asymp \epsilon \ln(1/\epsilon)$. Note that $\frac{4T\|\mathbf{w}^\star - \mathbf{w}_0\|^2}{-v_z z} > 0$. We proceed with Equation (37), in this case, the non-leading terms contribute $O\left(T^{-\frac{1}{2(1-\epsilon)}}\right)$ — as $\exp(-q/2)$, the dominating term becomes $O\left(T^{-\frac{1}{2(1-\epsilon)}}\right)$). Splitting it to cases.

Case (1) When $\frac{\epsilon}{1-\epsilon}\ln\left(\frac{4T\|\mathbf{w}^\star - \mathbf{w}_0\|^2}{-v_z z}\right) \in (0,1)$,

$$
\begin{aligned}
&- z \cdot \left(\exp\left(-\frac{\epsilon}{1-\epsilon}\ln\left(\frac{4T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2}{-v_z z}\right)\right) - 1\right) + O\left(T^{-\frac{1}{2(1-\epsilon)}}\right) \\
\left[e^{-x} - 1 \leq -\tfrac{1}{2}x,\ \forall x \in (0,1)\right] &\leq z \cdot \frac{\epsilon}{2(1-\epsilon)}\ln\left(\frac{4T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2}{-v_z z}\right) + O\left(T^{-\frac{1}{2(1-\epsilon)}}\right) \\
&\lesssim z \cdot \frac{\epsilon}{1-\epsilon}\ln(T) + O\left(T^{-\frac{1}{2(1-\epsilon)}}\right).
\end{aligned}
$$

As before, there exists a constant C_2 such that for all sufficiently large T we have $O\left(T^{-\frac{1}{2(1-\epsilon)}}\right) \leq C_2 T^{-\frac{1}{2(1-\epsilon)}}$, and hence we require $-z\frac{\epsilon}{1-\epsilon}\ln T \gtrsim C_2 T^{-\frac{1}{2(1-\epsilon)}}$. Although this equation does not admit a closed-form solution for T, choosing T_0 so that $T > T_0 = \Theta\left(\frac{1}{\epsilon^2 \ln^2(1/\epsilon)}\right)$, with a sufficiently large implicit constant, ensures that $\ln T_0 \asymp \ln(1/\epsilon)$ and $T_0^{-\frac{1}{2(1+\epsilon)}} \asymp \epsilon \ln(1/\epsilon)$.

Case (2) When $\frac{\epsilon}{1-\epsilon}\ln\left(\frac{4T\|\mathbf{w}^\star - \mathbf{w}_0\|^2}{-v_z z}\right) \in [1,\infty)$,

$$
\begin{aligned}
&- z\left(\exp\left(-\frac{\epsilon}{1-\epsilon}\ln\left(\frac{4T\left\|\mathbf{w}^\star - \mathbf{w}_0\right\|^2}{-v_z z}\right)\right) - 1\right) + O\left(T^{-\frac{1}{2(1-\epsilon)}}\right) \\
&\leq -z\left(e^{-1} - 1\right) + O\left(T^{-\frac{1}{2(1-\epsilon)}}\right).
\end{aligned}
$$

Therefore, we require,

$$
-z\left(e^{-1} - 1\right) \gtrsim K_2 T^{-\frac{1}{2(1-\epsilon)}}
$$

which holds for all sufficiently large T independent of ϵ.

In this regime ($\gamma \asymp \frac{\ln T}{NT}$), the maximal gap between G_T^{esgd} and G_T^{ridge} is bounded by $O(\frac{\ln T}{T^{1+\beta}})$ (for some $\beta > 0$, since they share the same expansion in Equation (29)). Consequently, we have $G_T^{\mathrm{esgd}} - G_T^{\mathrm{ridge}} = O\left(\frac{1}{T}\right)$.

□

D Proofs for Section 5: Generalized Ridge Regression

D.1 Auxiliary Lemmas and definitions

Definition 3 (Zhao et al. (2025), Definition 7)**.** *Let $\mathbf{A}_n, \mathbf{B}_n \in \mathbb{R}^{n\times n}$ be sequences of random or deterministic symmetric matrices. We say that $\mathbf{A}_n$ and $\mathbf{B}_n$ are equivalent if, for any sequence of deterministic matrices $\mathbf{C}_n \in \mathbb{R}^{n\times n}$ such that $\limsup_{n\to\infty} \|\mathbf{C}_n\|_{\mathrm{op}} < \infty$, and for any sequence of deterministic vectors $\mathbf{v}_n \in \mathbb{R}^n$ such that $\limsup_{n\to\infty} \|\mathbf{v}_n\| < \infty$, we have, almost surely,*

$$\frac{1}{n}\mathrm{Tr}\left[\mathbf{C}_n(\mathbf{A}_n - \mathbf{B}_n)\right] \to 0, \qquad \mathbf{v}_n^\top(\mathbf{A}_n - \mathbf{B}_n)\mathbf{v}_n \to 0.$$

Here, the operator norm is defined by $\|\mathbf{A}\|_{\mathrm{op}} := \sqrt{\lambda_{\max}(\mathbf{A}^\top\mathbf{A})}$ for any real matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$.

We write $\mathbf{A}_n \asymp \mathbf{B}_n$ if the sequences of matrices $\{\mathbf{A}_n\}$ and $\{\mathbf{B}_n\}$ are equivalent. For a random matrix sequence $\{\mathbf{A}_n\}$, if there exists a sequence of deterministic matrices $\{\bar{\mathbf{A}}_n\}$ such that $\mathbf{A}_n \asymp \bar{\mathbf{A}}_n$, then $\bar{\mathbf{A}}_n$ is referred to as a *deterministic equivalent* of $\mathbf{A}_n$.

Definition 4 (Couillet & Liao (2022), Notation 2)**.** *For $\mathcal{A} \subset \mathbb{C}$, $z \in \mathcal{A}$ and $m \in \mathbb{C}$, we denote $\mathcal{Z}(\mathcal{A})$ the set of scalar pairs*

$$\mathcal{Z}(A) = \Big\{(z,m) \in \mathcal{A}\times\mathbb{C},\ \textit{such that } \Im[z]\cdot\Im[m] > 0 \ \textit{ if } \Im[z] \neq 0$$
$$\textit{or } (m > 0 \textit{ if } z < \inf\mathcal{A}\cap\mathbb{R}) \textit{ or } (m < 0 \textit{ if } z > \sup\mathcal{A}\cap\mathbb{R})\Big\}.$$

Proposition 1 (Zhao et al. (2025), Proposition 8)**.**

1. *If $\mathbf{A}_n \asymp \mathbf{B}_n$ and $\mathbf{C}_n \asymp \mathbf{D}_n$, then*
$$\mathbf{A}_n + \mathbf{C}_n \asymp \mathbf{B}_n + \mathbf{D}_n.$$
2. *If $\mathbf{A}_n \asymp \mathbf{B}_n$ and $\{\mathbf{E}_n\}$ is a sequence of deterministic matrices such that $\|\mathbf{E}_n\|_{\mathrm{op}} < \infty$, then*
$$\mathbf{E}_n\mathbf{A}_n\mathbf{E}_n^\top \asymp \mathbf{E}_n\mathbf{B}_n\mathbf{E}_n^\top.$$
3. *If $\mathbf{A}_n \asymp \mathbf{B}_n$, then*
$$\frac{1}{n}\mathrm{Tr}[\mathbf{A}_n - \mathbf{B}_n] \xrightarrow{a.s.} 0.$$

Theorem 6 (Sample covariance matrix, Silverstein & Bai (1995); Couillet & Liao (2022), Theorem 2.6)**.** *Let $\mathbf{X} = \mathbf{C}^{1/2}\mathbf{Z} \in \mathbb{R}^{d\times n}$ with symmetric nonnegative definite $\mathbf{C} \in \mathbb{R}^{d\times d}$ of bounded operator norm (i.e., $\limsup_{d\to\infty}\|\mathbf{C}\|_{\mathrm{op}} < \infty$), and $\mathbf{Z} \in \mathbb{R}^{d\times n}$ having independent zero-mean and unit-variance entries with $4+\epsilon$ bounded moment for some $\epsilon > 0$. Then, as $n, d \to \infty$ with $d/n \to c \in (0,\infty)$, letting*

$$\mathbf{Q}(z) = \left(\frac{1}{n}\mathbf{X}\mathbf{X}^\top - z\mathbf{I}_d\right)^{-1}, \qquad \widetilde{\mathbf{Q}}(z) = \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} - z\mathbf{I}_n\right)^{-1},$$

we have

$$\mathbf{Q}(z) \asymp \bar{\mathbf{Q}}(z) = -\frac{1}{z}\big(\mathbf{I}_d + \widetilde{m}_d(z)\mathbf{C}\big)^{-1},$$
$$\widetilde{\mathbf{Q}}(z) \asymp \overline{\widetilde{\mathbf{Q}}}(z) = \widetilde{m}_d(z)\mathbf{I}_n,$$

where $(z, \widetilde{m}_p(z))$ is the unique solution in $\mathcal{Z}\left(\mathbb{C}\setminus\mathbb{R}_+\right)$ (see Definition 4) of

$$\widetilde{m}_d(z) = \left(-z + \frac{1}{n}\mathrm{Tr}\mathbf{C}\big(\mathbf{I}_d + \widetilde{m}_d(z)\mathbf{C}\big)^{-1}\right)^{-1}. \tag{38}$$

In particular, if the empirical spectral measure of $\mathbf{C}$ converges, i.e., $\mu_{\mathbf{C}} \Rightarrow \nu$ as $d\to\infty$, then

$$\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\top} \xrightarrow{a.s.} \mu, \qquad \mu_{\frac{1}{n}\mathbf{X}^\top\mathbf{X}} \xrightarrow{a.s.} \widetilde{\mu} \quad \textit{as } d, n \to \infty,$$

where μ and $\widetilde{\mu}$ are the unique measures having Stieltjes transforms $m(z)$ and $\widetilde{m}(z)$, respectively, with

$$m(z) = \frac{1}{c}\widetilde{m}(z) + \frac{1-c}{cz}, \qquad \widetilde{m}(z) = \left(-z + c\int\frac{t\nu(\mathrm{d}t)}{1+\widetilde{m}(z)t}\right)^{-1}.$$

Theorem 7. *Assume $\mathbf{X} = \mathbf{Z}\mathbf{\Sigma}^{\frac{1}{2}}$, where $\mathbf{Z} \in \mathbb{R}^{n\times d}$ has i.i.d. entries with zero mean, unit variance and finite $4+\epsilon$ moment. Let $\lambda > 0$ be a fixed real number, $\mathbf{Q}(-\lambda) = \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d\right)^{-1}$ be the resolvent matrix, $\mathbf{A} \in \mathbb{R}^{d\times d}$ is a symmetric deterministic matrix or symmetric random matrix independent of $\mathbf{Q}(-\lambda)$, and having bounded operator norm. Suppose $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ is a deterministic positive semi–definite matrix with bounded operator norm, then as $n, d \to \infty$ with $n/d \to \alpha \in (0,\infty)$ we have*

$$\mathbf{Q}(-\lambda)\,\mathbf{A}\mathbf{Q}(-\lambda) \asymp \bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}} + \frac{1}{n}\mathrm{Tr}\left[\mathbf{\Sigma}\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}\right]\left[\left[1+\frac{1}{n}\mathrm{Tr}\left(\mathbf{\Sigma}\bar{\mathbf{Q}}\right)\right]^2 - \frac{1}{n}\mathrm{Tr}\left(\mathbf{\Sigma}\bar{\mathbf{Q}}\right)^2\right]^{-1}\bar{\mathbf{Q}}\mathbf{\Sigma}\bar{\mathbf{Q}}.$$

Proof. If $\mathbf{A}$ is random, we condition throughout on $\mathbf{A}$. Since $\mathbf{A}$ is independent of $\mathbf{Q}(-\lambda)$ and has bounded operator norm, it may then be treated as deterministic. For $|\rho|$ sufficiently small, $\mathbf{I}_p + \rho\mathbf{A}$ is positive definite.

The key observation (as also identified in Patil et al. (2022), Lemma S.6.10) is that

$$\lambda^2\mathbf{Q}(-\lambda)\mathbf{A}\mathbf{Q}(-\lambda) = -\frac{\partial}{\partial\rho}\left[\lambda\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d + \lambda\rho\mathbf{A}\right)^{-1}\right]\Bigg|_{\rho=0}. \tag{39}$$

Thus, it suffices to obtain a deterministic equivalent for the parameterized resolvent in Equation (39), differentiate it with respect to ρ, and set $\rho = 0$.

We would like to make use of Theorem 6, note that,

$$\begin{aligned}&\lambda\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d + \lambda\rho\mathbf{A}\right)^{-1} = \lambda\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\left(\mathbf{I}_d + \rho\mathbf{A}\right)\right)^{-1}\\ &= (\mathbf{I}_d+\rho\mathbf{A})^{-1/2}\lambda\left[(\mathbf{I}_d+\rho\mathbf{A})^{-1/2}\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,(\mathbf{I}_d+\rho\mathbf{A})^{-1/2} + \lambda\mathbf{I}_d\right]^{-1}(\mathbf{I}_d+\rho\mathbf{A})^{-1/2}.\end{aligned} \tag{40}$$

The population covariance corresponding to the resolvent in the middle of Equation (40) is

$$(\mathbf{I}_p+\rho\mathbf{A})^{-1/2}\mathbf{\Sigma}(\mathbf{I}_p+\rho\mathbf{A})^{-1/2}.$$

Since its operator norm remains uniformly bounded for ρ in a sufficiently small neighborhood of zero, and since $\mathbb{E}|Z_{ij}|^{4+\epsilon} < \infty$, we may apply Theorem 6. It follows that

$$\begin{aligned}&\lambda\left[(\mathbf{I}_d+\rho\mathbf{A})^{-1/2}\frac{1}{n}\mathbf{X}^\top\mathbf{X}(\mathbf{I}_d+\rho\mathbf{A})^{-1/2} + \lambda\mathbf{I}_d\right]^{-1}\\ &\qquad\asymp\left[\widetilde{m}_p(-\lambda,\rho)(\mathbf{I}_d+\rho\mathbf{A})^{-1/2}\mathbf{\Sigma}(\mathbf{I}_d+\rho\mathbf{A})^{-1/2} + \mathbf{I}_d\right]^{-1},\end{aligned} \tag{41}$$

where $\widetilde{m}_p(-\lambda,\rho)$ is the unique positive solution of the fixed-point equation from Theorem 6,

$$\frac{1}{\widetilde{m}_p(-\lambda,\rho)} = \lambda + \frac{1}{n}\mathrm{Tr}\left[\mathbf{\Sigma}\left(\widetilde{m}_p(-\lambda,\rho)\mathbf{\Sigma} + \mathbf{I}_d + \rho\mathbf{A}\right)^{-1}\right]. \tag{42}$$

Combining Equations (40) and (41) and using the product rule for deterministic equivalents (Proposition 1) gives

$$\begin{aligned}&\lambda\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d + \lambda\rho\mathbf{A}\right)^{-1}\\ &= (\mathbf{I}_d+\rho\mathbf{A})^{-1/2}\lambda\left[(\mathbf{I}_d+\rho\mathbf{A})^{-1/2}\frac{1}{n}\mathbf{X}^\top\mathbf{X}\,(\mathbf{I}_d+\rho\mathbf{A})^{-1/2} + \lambda\mathbf{I}_d\right]^{-1}(\mathbf{I}_d+\rho\mathbf{A})^{-1/2}.\\ &\asymp(\mathbf{I}_d+\rho\mathbf{A})^{-1/2}\left[\widetilde{m}_p(-\lambda,\rho)(\mathbf{I}_d+\rho\mathbf{A})^{-1/2}\mathbf{\Sigma}(\mathbf{I}_d+\rho\mathbf{A})^{-1/2} + \mathbf{I}_d\right]^{-1}(\mathbf{I}_d+\rho\mathbf{A})^{-1/2}\\ &= \left(\widetilde{m}_p(-\lambda,\rho)\mathbf{\Sigma} + \mathbf{I}_d + \rho\mathbf{A}\right)^{-1}\end{aligned} \tag{43}$$

At $\rho = 0$,

$$(\widetilde{m}_p(-\lambda, 0)\mathbf{\Sigma} + \mathbf{I}_d)^{-1} = \lambda\bar{\mathbf{Q}}. \tag{44}$$

Setting $\rho = 0$ in Equation (42) and using Equation (44) gives

$$\frac{1}{\widetilde{m}_p(-\lambda, 0)} = \lambda + \frac{\lambda}{n}\mathrm{Tr}\left(\mathbf{\Sigma}\bar{\mathbf{Q}}\right) = \lambda\left[1 + \frac{1}{n}\mathrm{Tr}\left(\mathbf{\Sigma}\bar{\mathbf{Q}}\right)\right]. \tag{45}$$

Both sides of Equation (43) are analytic in ρ in a neighborhood of zero and are uniformly bounded there, since $\lambda > 0$ and $\|\mathbf{A}\|_{\mathrm{op}} = O(1)$. We use the differentiation technique for deterministic equivalents (see Patil et al. (2022), Lemma S.7.2 (5)). Let $\mathbf{T} \in \mathbb{R}^{d\times d}$ have trace norm uniformly bounded in d. For $|\rho|$ sufficiently small such that $\mathbf{I}_d + \rho\mathbf{A} \succ 0$, we have

$$\left|\mathrm{Tr}\left[\mathbf{T}\lambda\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d + \lambda\rho\mathbf{A}\right)^{-1}\mathbf{A}\right]\right| \leq \|\mathbf{T}\|_* \left\|\lambda\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda(\mathbf{I}_d + \rho\mathbf{A})\right)^{-1}\right\|_{\mathrm{op}} \|\mathbf{A}\|_{\mathrm{op}}.$$

Since

$$\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda(\mathbf{I}_d + \rho\mathbf{A}) \succeq \lambda(\mathbf{I}_d + \rho\mathbf{A}),$$

we obtain

$$\left\|\lambda\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda(\mathbf{I}_d + \rho\mathbf{A})\right)^{-1}\right\|_{\mathrm{op}} \leq \|(\mathbf{I}_d + \rho\mathbf{A})^{-1}\|_{\mathrm{op}}.$$

Therefore, for ρ in a sufficiently small fixed neighborhood of zero,

$$\left|\mathrm{Tr}\left[\mathbf{T}\lambda\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d + \lambda\rho\mathbf{A}\right)^{-1}\mathbf{A}\right]\right| \leq M,$$

for some constant $M < \infty$ independent of n, d. Thus, we can apply the differentiation rule from Patil et al. (2022, Lemma S.7.2 (5)). Differentiating Equation (43) at $\rho = 0$ and using Equation (39) yields

$$\lambda^2\mathbf{Q}(-\lambda)\mathbf{A}\mathbf{Q}(-\lambda) \asymp (\widetilde{m}_p(-\lambda, 0)\mathbf{\Sigma} + \mathbf{I}_d)^{-1}\left[\frac{\partial}{\partial\rho}\widetilde{m}_p(-\lambda, \rho)\Big|_{\rho=0}\mathbf{\Sigma} + \mathbf{A}\right](\widetilde{m}_p(-\lambda, 0)\mathbf{\Sigma} + \mathbf{I}_d)^{-1}. \tag{46}$$

It remains to evaluate $\frac{\partial}{\partial\rho}\widetilde{m}_p(-\lambda, \rho)\Big|_{\rho=0}$. Differentiating Equation (42) with respect to ρ and then setting $\rho = 0$ gives

$$\begin{aligned}
&-\frac{\partial}{\partial\rho}\widetilde{m}_p(-\lambda, \rho)\Big|_{\rho=0}\frac{1}{\widetilde{m}_p(-\lambda, 0)^2} \\
&\quad = -\frac{1}{n}\mathrm{Tr}\left[\mathbf{\Sigma}\left(\widetilde{m}_p(-\lambda, 0)\mathbf{\Sigma} + \mathbf{I}_d\right)^{-1}\left(\frac{\partial}{\partial\rho}\widetilde{m}_p(-\lambda, \rho)\Big|_{\rho=0}\mathbf{\Sigma} + \mathbf{A}\right)\left(\widetilde{m}_p(-\lambda, 0)\mathbf{\Sigma} + \mathbf{I}_d\right)^{-1}\right].
\end{aligned} \tag{47}$$

Since $\mathbf{\Sigma}$ commutes with $(\widetilde{m}_p(-\lambda, 0)\mathbf{\Sigma} + \mathbf{I}_p)^{-1}$, solving Equation (47) gives

$$\frac{\partial}{\partial\rho}\widetilde{m}_p(-\lambda, \rho)\Big|_{\rho=0} = \frac{\frac{1}{n}\mathrm{Tr}\left[\mathbf{\Sigma}\left(\widetilde{m}_p(-\lambda, 0)\mathbf{\Sigma} + \mathbf{I}_d\right)^{-1}\mathbf{A}\left(\widetilde{m}_p(-\lambda, 0)\mathbf{\Sigma} + \mathbf{I}_d\right)^{-1}\right]}{\frac{1}{\widetilde{m}_p(-\lambda, 0)^2} - \frac{1}{n}\mathrm{Tr}\left[\mathbf{\Sigma}^2\left(\widetilde{m}_p(-\lambda, 0)\mathbf{\Sigma} + \mathbf{I}_d\right)^{-2}\right]}. \tag{48}$$

Using Equations (44) and (45), the numerator of Equation (48) becomes $\frac{\lambda^2}{n}\mathrm{Tr}\left[\mathbf{\Sigma}\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}\right]$, while its denominator becomes $\lambda^2\left[\left(1 + \frac{1}{n}\mathrm{Tr}\left(\mathbf{\Sigma}\bar{\mathbf{Q}}\right)\right)^2 - \frac{1}{n}\mathrm{Tr}\left[\left(\mathbf{\Sigma}\bar{\mathbf{Q}}\right)^2\right]\right]$. Therefore,

$$\frac{\partial}{\partial\rho}\widetilde{m}_p(-\lambda, \rho)\Big|_{\rho=0} = \frac{\frac{1}{n}\mathrm{Tr}\left[\mathbf{\Sigma}\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}\right]}{\left(1 + \frac{1}{n}\mathrm{Tr}\left(\mathbf{\Sigma}\bar{\mathbf{Q}}\right)\right)^2 - \frac{1}{n}\mathrm{Tr}\left[\left(\mathbf{\Sigma}\bar{\mathbf{Q}}\right)^2\right]}. \tag{49}$$

Finally, substituting Equations (44) and (49) into Equation (46) and dividing by λ^2 yields

$$\mathbf{Q}(-\lambda)\mathbf{A}\mathbf{Q}(-\lambda) \asymp \bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}} + \frac{\frac{1}{n}\mathrm{Tr}\left[\mathbf{\Sigma}\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}\right]}{\left(1+\frac{1}{n}\mathrm{Tr}\left(\mathbf{\Sigma}\bar{\mathbf{Q}}\right)\right)^2 - \frac{1}{n}\mathrm{Tr}\left[\left(\mathbf{\Sigma}\bar{\mathbf{Q}}\right)^2\right]}\bar{\mathbf{Q}}\mathbf{\Sigma}\bar{\mathbf{Q}},$$

as claimed. □

The following Lemma provides the closed-from solution to the iterates produced by the generalized ridge regression algorithm (Algorithm 3).

Lemma 8. *At each iteration t the predictor update under Algorithm 3 is*

$$\mathbf{w}_t = \left(\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t + \mathbf{M}\right)^{-1}\left(\frac{1}{n}\mathbf{X}_t^\top\mathbf{y}_t + \mathbf{M}\mathbf{w}_{t-1}\right).$$

Proof. Let $t \in [1, T]$ and consider the optimization problem

$$\mathbf{w}_t \in \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2n}\left\|\mathbf{X}_t\mathbf{w} - \mathbf{y}_t\right\|_2^2 + \frac{1}{2}\left\|\mathbf{w} - \mathbf{w}_{t-1}\right\|_{\mathbf{M}}^2,$$

where $\mathbf{M} \succ 0$. The objective is a convex quadratic function of $\mathbf{w}$, hence the minimizer is unique in that case.

Expand the objective and differentiate with respect to $\mathbf{w}$:

$$f_t(\mathbf{w}) := \frac{1}{2n}\|\mathbf{X}_t\mathbf{w} - \mathbf{y}_t\|_2^2 + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{t-1})^\top\mathbf{M}(\mathbf{w} - \mathbf{w}_{t-1}),$$
$$\nabla f_t(\mathbf{w}) = \frac{1}{n}\mathbf{X}_t^\top(\mathbf{X}_t\mathbf{w} - \mathbf{y}_t) + \mathbf{M}(\mathbf{w} - \mathbf{w}_{t-1}).$$

Setting $\nabla f_t(\mathbf{w}) = 0$ gives the normal equations

$$\left(\tfrac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t + \mathbf{M}\right)\mathbf{w} = \tfrac{1}{n}\mathbf{X}_t^\top\mathbf{y}_t + \mathbf{M}\mathbf{w}_{t-1}.$$

Since $\frac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t + \mathbf{M}$ is invertible, we obtain the closed-form solution

$$\mathbf{w}_t = \left(\tfrac{1}{n}\mathbf{X}_t^\top\mathbf{X}_t + \mathbf{M}\right)^{-1}\left(\tfrac{1}{n}\mathbf{X}_t^\top\mathbf{y}_t + \mathbf{M}\mathbf{w}_{t-1}\right).$$

□

D.2 Proof of Theorem 3: Closed-Form Generalization Loss with Diagonal Regularization

Recall Theorem 3. Under Assumption 2, the generalization loss after T tasks of the generalized ridge algorithm (Algorithm 3), specialized to a fixed regularization matrix $\mathbf{M}$, as $n, d \to \infty$, satisfies

$$G_T^{\text{gen}} \xrightarrow{P} (\mathbf{w}_0 - \mathbf{w}^\star)^\top \left(\bar{\mathbf{Q}}^{2T} + v_x \sum_{k=1}^{T} \bar{\mathbf{Q}}^{2k} \tilde{\mu}\tilde{\rho}_k \right) (\mathbf{w}_0 - \mathbf{w}^\star)$$
$$+ \frac{v_z}{n} \mathrm{Tr}\left[\sum_{k=1}^{T} \left[\mathbf{I} - \bar{\mathbf{Q}} \left[\mathbf{I} + \frac{v_x^2}{n} \mathrm{Tr}\left[\mathbf{M}^{-1} \bar{\mathbf{Q}}^2 \right] \tilde{\mu} \mathbf{M}^{-1} \right] \right] \left[\bar{\mathbf{Q}}^{2(T-k)} + v_x \sum_{i=k+1}^{T} \bar{\mathbf{Q}}^{2(i-k)} \tilde{\mu}\tilde{\rho}_i \right] \mathbf{M}^{-1} \bar{\mathbf{Q}} \right]$$

where $\bar{\mathbf{Q}} = \left(\mathbf{I} + v_x \tilde{m}_p(-1) \mathbf{M}^{-1} \right)^{-1}$, with $\tilde{m}_p(-1)$ being the unique positive solution of

$$\frac{1}{\tilde{m}_p(-1)} = 1 + \frac{1}{n} \sum_{i=1}^{d} \frac{v_x M_{i,i}^{-1}}{1 + \tilde{m}_p(-1) v_x M_{i,i}^{-1}}.$$

And,

$$\tilde{\rho}_k = \frac{v_x}{n} \mathrm{Tr}\Big[\mathbf{M}^{-2} \Big(\bar{\mathbf{Q}}^{2(T-k+1)} + v_x \tilde{\mu} \sum_{i=k+1}^{T} \tilde{\rho}_i \bar{\mathbf{Q}}^{2(i-k+1)} \Big) \Big],$$
$$\tilde{\mu} = \left[\left[1 + \frac{v_x}{n} \mathrm{Tr}\left(\mathbf{M}^{-1} \bar{\mathbf{Q}} \right) \right]^2 - \frac{v_x^2}{n} \mathrm{Tr}\left(\mathbf{M}^{-1} \bar{\mathbf{Q}} \right)^2 \right]^{-1}.$$

Proof. We will develop a slightly more general result, the outcome would be able to accommodate general diagonal covariance matrix $\mathbf{\Sigma}$ (spectrally bounded), in contrast to the setting considered in Section 3. Specifically, here we may assume that the data for task t is generated as $\mathbf{X}_t = \mathbf{Z}_t \mathbf{\Sigma}^{1/2}$, where $\mathbf{Z}_t \in \mathbb{R}^{n \times d}$ has i.i.d. entries with zero mean and unit variance.

For each $t \in [1, T]$, we are interested in solving

$$\mathbf{w}_t^{\text{gen}} = \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \left\| \mathbf{X}_t \mathbf{w} - \mathbf{y} \right\|_2^2 + \left\| \mathbf{w} - \mathbf{w}_{t-1} \right\|_{\mathbf{M}}^2.$$

By Lemma 8,

$$\mathbf{w}_t^{\text{gen}} = \left(\frac{1}{n} \mathbf{X}_t^\top \mathbf{X}_t + \mathbf{M} \right)^{-1} \left(\frac{1}{n} \mathbf{X}_t^\top \mathbf{y}_t + \mathbf{M} \mathbf{w}_{t-1}^{\text{gen}} \right).$$

Therefore,

$$\begin{aligned}
\mathbf{w}_T^{\text{gen}} - \mathbf{w}^\star &= \left(\frac{1}{n} \mathbf{X}_T^\top \mathbf{X}_T + \mathbf{M} \right)^{-1} \left(\frac{1}{n} \mathbf{X}_T^\top \left(\mathbf{X}_T \mathbf{w}^\star + \mathbf{z}_T \right) + \mathbf{M} \mathbf{w}_{T-1}^{\text{gen}} \right) - \mathbf{w}^\star \\
&= \left(\frac{1}{n} \mathbf{X}_T^\top \mathbf{X}_T + \mathbf{M} \right)^{-1} \left(\frac{1}{n} \mathbf{X}_T^\top \left(\mathbf{X}_T \mathbf{w}^\star + \mathbf{z}_T \right) + \mathbf{M} \mathbf{w}_{T-1}^{\text{gen}} - \left(\frac{1}{n} \mathbf{X}_T^\top \mathbf{X}_T + \mathbf{M} \right) \mathbf{w}^\star \right) \\
&= \left(\frac{1}{n} \mathbf{X}_T^\top \mathbf{X}_T + \mathbf{M} \right)^{-1} \left(\frac{1}{n} \mathbf{X}_T^\top \mathbf{z}_T + \mathbf{M} \left(\mathbf{w}_{T-1}^{\text{gen}} - \mathbf{w}^\star \right) \right) \\
&= \mathbf{P}_T \mathbf{M} \left(\mathbf{w}_{T-1}^{\text{gen}} - \mathbf{w}^\star \right) + \mathbf{P}_T \frac{1}{n} \mathbf{X}_T^\top \mathbf{z}_T \\
&= \dots \\
&= \left[\prod_{m=T}^{1} \mathbf{P}_m \mathbf{M} \right] (\mathbf{w}_0 - \mathbf{w}^\star) + \sum_{k=1}^{T} \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M} \right] \mathbf{P}_k \frac{1}{n} \mathbf{X}_k^\top \mathbf{z}_k,
\end{aligned}$$

where $\mathbf{P}_t = \left(\frac{1}{n} \mathbf{X}_t^\top \mathbf{X}_t + \mathbf{M} \right)^{-1} = \mathbf{M}^{-\frac{1}{2}} \left(\frac{1}{n} \mathbf{M}^{-\frac{1}{2}} \mathbf{X}_t^\top \mathbf{X}_t \mathbf{M}^{-\frac{1}{2}} + \mathbf{I} \right)^{-1} \mathbf{M}^{-\frac{1}{2}}$ for all $t \in [1, T]$.

Then,

$$G_T^{\text{gen}} = \left[\left[\prod_{m=T}^{1}\mathbf{P}_m\mathbf{M}\right](\mathbf{w}_0-\mathbf{w}^\star)+\sum_{k=1}^{T}\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]\mathbf{P}_k\frac{1}{n}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z}_k^\top\mathbf{z}_k\right]^\top \left[\left[\prod_{m=T}^{1}\mathbf{P}_m\mathbf{M}\right](\mathbf{w}_0-\mathbf{w}^\star)+\sum_{k=1}^{T}\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]\mathbf{P}_k\frac{1}{n}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z}_k^\top\mathbf{z}_k\right]$$

$$\overset{[1]}{=}(\mathbf{w}_0-\mathbf{w}^\star)^\top\underbrace{\left[\prod_{m=T}^{1}\mathbf{P}_m\mathbf{M}\right]^\top\left[\prod_{m=T}^{1}\mathbf{P}_m\mathbf{M}\right]}_{\text{term 1}}(\mathbf{w}_0-\mathbf{w}^\star) + \underbrace{\left[\frac{1}{n^2}\sum_{k=1}^{T}\mathbf{z}_k^\top\mathbf{Z}_k\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]^\top\sum_{k'=1}^{T}\left[\prod_{m=T}^{k'+1}\mathbf{P}_m\mathbf{M}\right]\mathbf{P}_{k'}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z}_{k'}^\top\mathbf{z}_{k'}\right]}_{\text{term 2}}, \tag{50}$$

where [1] follows since under the statistical design (Section 3) in which $\mathbb{E}[\mathbf{z}_k]=\mathbf{0}$ for all $k\in[1,T]$, and the noise variables are sampled independently of all tasks, so by the weak law of large numbers cross terms vanish as $n\to\infty$.

We begin by analyzing the first term. Let $\mathbf{A}_i=\mathbf{M}^{\frac{1}{2}}\mathbf{P}_i\left[\prod_{m=T}^{i+1}\mathbf{P}_m\mathbf{M}\right]^\top\left[\prod_{m=T}^{i+1}\mathbf{P}_m\mathbf{M}\right]\mathbf{P}_i\mathbf{M}^{\frac{1}{2}}$. To build some intuition for the application of Theorem 7, note that in order to compute term 1 we would need to solve

$$\left[\prod_{m=T}^{1}\mathbf{P}_m\mathbf{M}\right]^\top\left[\prod_{m=T}^{1}\mathbf{P}_m\mathbf{M}\right]=\mathbf{M}\mathbf{P}_1\ldots\mathbf{M}\mathbf{P}_T\mathbf{P}_T\mathbf{M}\ldots\mathbf{P}_1\mathbf{M}$$
$$=\mathbf{M}^{\frac{1}{2}}\underbrace{\underbrace{\left[\mathbf{M}^{\frac{1}{2}}\mathbf{P}_1\mathbf{M}^{\frac{1}{2}}\right]}_{\mathbf{Q}(-1)}\underbrace{\ldots\mathbf{M}\mathbf{P}_T\mathbf{M}^{\frac{1}{2}}\mathbf{M}^{-1}\mathbf{M}^{\frac{1}{2}}\mathbf{P}_T\mathbf{M}\ldots}_{\text{independent}}\underbrace{\left[\mathbf{M}^{\frac{1}{2}}\mathbf{P}_1\mathbf{M}^{\frac{1}{2}}\right]}_{\mathbf{Q}(-1)}}_{\mathbf{Q}(-1)\mathbf{A}_2\mathbf{Q}(-1)}\mathbf{M}^{\frac{1}{2}}.$$

where $\mathbf{Q}(z):=\left(\frac{1}{n}\mathbf{M}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z}^\top\mathbf{Z}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{M}^{-\frac{1}{2}}-z\mathbf{I}\right)^{-1}$ is the resolvent matrix with $\mathbf{Z}$ having i.i.d. entries with zero mean and unit variance. The middle block is a symmetric random matrix independent of $\mathbf{Q}(-1)$ with bounded operator norm, therefore we use Theorem 7 to evaluate $\mathbf{Q}(-1)\mathbf{A}_2\mathbf{Q}(-1)$.

Using Theorem 6, $\mathbf{Q}(-1)$ admits a deterministic equivalent, where,

$$\bar{\mathbf{Q}}(-1)=\left(\mathbf{I}+\tilde{m}_p(-1)\boldsymbol{\Sigma}\mathbf{M}^{-1}\right)^{-1}\text{ and }\frac{1}{\tilde{m}_p(-1)}=1+\frac{1}{n}\sum_{i=1}^{d}\frac{\sigma_i m_i^{-1}}{1+\tilde{m}_p(-1)\sigma_i m_i^{-1}}.$$

Here σ_i, m_i denote respectively, the i'th eigenvalues of the matrices $\boldsymbol{\Sigma},\mathbf{M}$. For notational convenience, we set $\mathbf{Q}:=\mathbf{Q}(-1)$, and $\bar{\mathbf{Q}}:=\bar{\mathbf{Q}}(-1)$. Both matrices are diagonal since both $\boldsymbol{\Sigma}$ and $\mathbf{M}$ are diagonal. Moreover, we note that under our standing assumption, we have $\sigma_i=v_x$ for all i. Nonetheless, as previously mentioned, our analysis also allows for non-isotropic covariance matrices.

For all $k\geq 1$, using Theorem 7 we attain the following recurrence,

$$\mathbf{A}_k\asymp\bar{\mathbf{Q}}(-1)\mathbf{A}_{k+1}\bar{\mathbf{Q}}(-1)+\tilde{\mu}\tilde{\rho}_k\mathbf{D} \tag{51}$$

where

$$\tilde{\rho}_k=\frac{1}{n}\mathrm{Tr}\left[\boldsymbol{\Sigma}\mathbf{M}^{-1}\bar{\mathbf{Q}}\mathbf{A}_{k+1}\bar{\mathbf{Q}}\right],$$
$$\tilde{\mu}=\left[\left[1+\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}\mathbf{M}^{-1}\bar{\mathbf{Q}}\right)\right]^2-\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}\mathbf{M}^{-1}\bar{\mathbf{Q}}\right)^2\right]^{-1},$$
$$\mathbf{D}=\bar{\mathbf{Q}}\boldsymbol{\Sigma}\mathbf{M}^{-1}\bar{\mathbf{Q}}.$$

Using the recurrence from Equation (51),

$$\begin{aligned}\mathbf{A}_k &\asymp \bar{\mathbf{Q}}\mathbf{A}_{k+1}\bar{\mathbf{Q}} + \tilde{\mu}\tilde{\rho}_k\mathbf{D} \\ &\asymp \bar{\mathbf{Q}}\left[\bar{\mathbf{Q}}\mathbf{A}_{k+2}\bar{\mathbf{Q}} + \tilde{\mu}\tilde{\rho}_{k+1}\mathbf{D}\right]\bar{\mathbf{Q}} + \tilde{\mu}\tilde{\rho}_k\mathbf{D} \\ &\asymp \dots \\ &\asymp \bar{\mathbf{Q}}^{T-k+1}\underbrace{\mathbf{A}_{T+1}}_{\mathbf{M}^{-1}}\bar{\mathbf{Q}}^{T-k+1} + \sum_{i=k}^{T}\bar{\mathbf{Q}}^{i-k}\tilde{\mu}\tilde{\rho}_i\bar{\mathbf{Q}}^{i-k}\mathbf{D}.\end{aligned}$$

Therefore, using Proposition 1,

$$\begin{aligned}\tilde{\rho}_k &= \frac{1}{n}\mathrm{Tr}\left[\mathbf{\Sigma}\mathbf{M}^{-1}\bar{\mathbf{Q}}\mathbf{A}_{k+1}\bar{\mathbf{Q}}\right] = \frac{1}{n}\mathrm{Tr}\left[\bar{\mathbf{Q}}\mathbf{\Sigma}\mathbf{M}^{-1}\bar{\mathbf{Q}}\mathbf{A}_{k+1}\right] \\ &\to \frac{1}{n}\mathrm{Tr}\left[\bar{\mathbf{Q}}\mathbf{\Sigma}\mathbf{M}^{-1}\bar{\mathbf{Q}}\left[\bar{\mathbf{Q}}^{2(T-k)}\mathbf{M}^{-1} + \sum_{i=k+1}^{T}\bar{\mathbf{Q}}^{2(i-k-1)}\tilde{\mu}\tilde{\rho}_i\mathbf{D}\right]\right] \\ &= \frac{1}{n}\mathrm{Tr}\left[\bar{\mathbf{Q}}^{2(T-k+1)}\mathbf{\Sigma}\mathbf{M}^{-2}\right] + \frac{1}{n}\mathrm{Tr}\left[\sum_{i=k+1}^{T}\bar{\mathbf{Q}}^{2(i-k+1)}\mathbf{\Sigma}^2\mathbf{M}^{-2}\tilde{\mu}\tilde{\rho}_i\right],\end{aligned}$$

almost surely.

And term 2 in Equation (50), becomes,

$$\begin{aligned}&\frac{1}{n^2}\left[\sum_{k=1}^{T}\mathbf{z}_k^\top\mathbf{Z}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]^\top\sum_{k'=1}^{T}\left[\prod_{m=T}^{k'+1}\mathbf{P}_m\mathbf{M}\right]\mathbf{P}_{k'}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z}_{k'}^\top\mathbf{z}_{k'}\right] \\ &= \frac{1}{n^2}\left[\sum_{k,k'=1}^{T}\mathbf{z}_k^\top\mathbf{Z}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]^\top\left[\prod_{m=T}^{k'+1}\mathbf{P}_m\mathbf{M}\right]\mathbf{P}_{k'}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z}_{k'}^\top\mathbf{z}_{k'}\right] \\ &\xrightarrow[P]{[1]} \frac{1}{n^2}\left[\sum_{k=1}^{T}\mathbf{z}_k^\top\mathbf{Z}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]\mathbf{P}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z}_k^\top\mathbf{z}_k\right] \\ &= \frac{1}{n^2}\left[\sum_{k=1}^{T}\mathrm{Tr}\left[\mathbf{z}_k^\top\mathbf{Z}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]\mathbf{P}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z}_k^\top\mathbf{z}_k\right]\right] \\ &= \frac{1}{n^2}\left[\sum_{k=1}^{T}\mathrm{Tr}\left[\mathbf{z}_k\mathbf{z}_k^\top\mathbf{Z}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]\mathbf{P}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z}_k^\top\right]\right] \\ &= \frac{1}{n^2}\sum_{k=1}^{T}\mathrm{Tr}\left[\mathbf{z}_k\mathbf{z}_k^\top\left[\mathbf{Z}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]\mathbf{P}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z}_k^\top\right]\right] \\ &\xrightarrow[P]{[2]} \frac{v_z}{n^2}\left[\sum_{k=1}^{T}\mathrm{Tr}\left[\mathbf{Z}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]\mathbf{P}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z}_k^\top\right]\right] \\ &= \frac{v_z}{n^2}\left[\sum_{k=1}^{T}\mathrm{Tr}\left[\mathbf{P}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z}_k^\top\mathbf{Z}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]\right]\right] \\ &= \frac{v_z}{n^2}\mathrm{Tr}\left[\sum_{k=1}^{T}\left[\mathbf{P}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z}_k^\top\mathbf{Z}_k\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{P}_k\right]\left[\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}\right]\right]\right].\end{aligned}$$

Where [1] and [2] follows since the noise $\mathbf{z}_k$ is sample independently of the data matrices $\mathbf{X}_1, \ldots, \mathbf{X}_T$ and $\mathbb{E}[\mathbf{z}_k \mathbf{z}_k^\top] = v_z \mathbf{I}$, with the weak law of large numbers. Note that,

$$\begin{aligned}
&\mathbf{P}_k \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{Z}_k^\top \mathbf{Z}_k \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{P}_k \\
&= n^2 \left(\boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{Z}_k^\top \mathbf{Z}_k \boldsymbol{\Sigma}^{\frac{1}{2}} + n\mathbf{M} \right)^{-1} \left(\boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{Z}_k^\top \mathbf{Z}_k \boldsymbol{\Sigma}^{\frac{1}{2}} + n\mathbf{M} - n\mathbf{M} \right) \left(\boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{Z}_k^\top \mathbf{Z}_k \boldsymbol{\Sigma}^{\frac{1}{2}} + n\mathbf{M} \right)^{-1} \\
&= n \left(\mathbf{P}_k - \mathbf{P}_k \mathbf{M} \mathbf{P}_k \right)
\end{aligned}$$

Therefore using Theorem 7,

$$\begin{aligned}
&\overline{\frac{1}{n} \mathbf{P}_k \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{Z}_k^\top \mathbf{Z}_k \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{P}_k} = \overline{\mathbf{P}_k - \mathbf{P}_k \mathbf{M} \mathbf{P}_k} = \mathbf{M}^{-\frac{1}{2}} \bar{\mathbf{Q}} \mathbf{M}^{-\frac{1}{2}} - \overline{\mathbf{M}^{-\frac{1}{2}} \mathbf{Q}^2 \mathbf{M}^{-\frac{1}{2}}} \\
&= \mathbf{M}^{-\frac{1}{2}} \left(\bar{\mathbf{Q}} - \left(\bar{\mathbf{Q}}^2 + \frac{1}{n} \mathrm{Tr} \left[\boldsymbol{\Sigma} \mathbf{M}^{-1} \bar{\mathbf{Q}}^2 \right] \left[\left[1 + \frac{1}{n} \mathrm{Tr} \left(\boldsymbol{\Sigma} \mathbf{M}^{-1} \bar{\mathbf{Q}} \right) \right]^2 - \frac{1}{n} \mathrm{Tr} \left(\boldsymbol{\Sigma} \mathbf{M}^{-1} \bar{\mathbf{Q}} \right)^2 \right]^{-1} \boldsymbol{\Sigma} \mathbf{M}^{-1} \bar{\mathbf{Q}}^2 \right) \right) \mathbf{M}^{-\frac{1}{2}}.
\end{aligned}$$

Since both $\left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M} \right]^\top \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M} \right]$ and $\mathbf{P}_k \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{Z}_k^\top \mathbf{Z}_k \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{P}_k$ have almost surely finite operator norm, then by the properties of deterministic equivalent we have

$$\begin{aligned}
&\frac{v_z}{n^2} \mathrm{Tr} \left[\sum_{k=1}^{T} \mathbf{P}_k \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{Z}_k^\top \mathbf{Z}_k \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{P}_k \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M} \right]^\top \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M} \right] \right] \\
&\xrightarrow{P} \frac{v_z}{n} \mathrm{Tr} \left[\sum_{k=1}^{T} \overline{\frac{1}{n} \mathbf{P}_k \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{Z}_k^\top \mathbf{Z}_k \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{P}_k} \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M} \right]^\top \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M} \right] \right] \\
&\xrightarrow{P} \frac{v_z}{n} \mathrm{Tr} \left[\sum_{k=1}^{T} \overline{\frac{1}{n} \mathbf{P}_k \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{Z}_k^\top \mathbf{Z}_k \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{P}_k} \overline{\left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M} \right]^\top \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M} \right]} \right],
\end{aligned}$$

where in each step we condition on the event that each one of the matrices has finite operator norm and then use Definition 3. Using

$$\left[\prod_{m=T}^{k} \mathbf{P}_m \mathbf{M}_m \right]^\top \left[\prod_{m=T}^{k} \mathbf{P}_m \mathbf{M}_m \right] = \mathbf{M}^{\frac{1}{2}} \mathbf{A}_k \mathbf{M}^{\frac{1}{2}} \asymp \bar{\mathbf{Q}}^{2(T-k+1)} + \sum_{i=k}^{T} \bar{\mathbf{Q}}^{2(i-k+1)} \tilde{\mu} \tilde{\rho}_i,$$

and combining all the terms in Equation (50).

$$\begin{aligned}
&G_T^{\mathrm{gen}} \xrightarrow{P} (\mathbf{w}_0 - \mathbf{w}^\star)^\top \left(\bar{\mathbf{Q}}^{2T} + \sum_{k=1}^{T} \bar{\mathbf{Q}}^{2k} \tilde{\mu} \tilde{\rho}_k \right) (\mathbf{w}_0 - \mathbf{w}^\star) \\
&+ \frac{v_z}{n} \mathrm{Tr} \left[\sum_{k=1}^{T} \mathbf{M}^{-1} \bar{\mathbf{Q}} \left[\mathbf{I} - \bar{\mathbf{Q}} \left[\mathbf{I} + \frac{1}{n} \mathrm{Tr} \left[\boldsymbol{\Sigma} \mathbf{M}^{-1} \bar{\mathbf{Q}}^2 \right] \tilde{\mu} \boldsymbol{\Sigma} \mathbf{M}^{-1} \right] \right] \left[\bar{\mathbf{Q}}^{2(T-k)} + \sum_{i=k+1}^{T} \bar{\mathbf{Q}}^{2(i-k)} \tilde{\mu} \tilde{\rho}_i \right] \right].
\end{aligned}$$

□

D.3 Closed-Form Generalization Loss with Task-Dependent Diagonal Regularization

In this appendix, we provide an *exact* generalization analysis under arbitrary positive-definite diagonal regularization matrices $\{\mathbf{M}_t\}_{t=1}^T$ (spectrally bounded). The resulting expression accommodates general task-dependent diagonal covariance matrices $\mathbf{\Sigma}_t$ (spectrally bounded), in contrast to the isotropic setting considered in Section 3. Specifically, here we may assume that the data for task t is generated as $\mathbf{X}_t = \mathbf{Z}_t \mathbf{\Sigma}_t^{1/2}$, where $\mathbf{Z}_t \in \mathbb{R}^{n\times d}$ has i.i.d. entries with zero mean and unit variance.

Theorem 8. *Under Assumption 2, the generalization loss after T tasks of the generalized ridge algorithm (Algorithm 3) with task dependent diagonal regularization, satisfies, as $n, d \to \infty$,*

$$
\begin{aligned}
G_T^{\text{gen}} \xrightarrow{P} (\mathbf{w}_0 - \mathbf{w}^\star)^\top & \left[\left[\prod_{m=1}^{T} \bar{\mathbf{Q}}_m \right]^2 + \sum_{i=2}^{T} \mathbf{\Sigma}_i \left[\prod_{r=1}^{i} \bar{\mathbf{Q}}_r \right]^2 \tilde{\mu}_i \tilde{\rho}_i + \tilde{\rho}_1 \tilde{\mu}_1 \bar{\mathbf{Q}}_1^2 \right] (\mathbf{w}_0 - \mathbf{w}^\star) \\
& + \frac{v_z}{n} \mathrm{Tr} \left[\sum_{k=1}^{T} \bar{\mathbf{Q}}_k \mathbf{M}_k^{-1} \left[\mathbf{I} - \bar{\mathbf{Q}}_k \left[\mathbf{I} + \frac{1}{n} \mathrm{Tr} \left[\mathbf{\Sigma}_k^{-1} \mathbf{M}_k^{-1} \bar{\mathbf{Q}}_k^2 \right] \tilde{\mu}_k \mathbf{M}_k^{-1} \right] \right] \right. \\
& \left. \left[\left[\prod_{m=k+1}^{T} \bar{\mathbf{Q}}_m \right]^2 + \sum_{i=k+2}^{T} \mathbf{\Sigma}_i \left[\prod_{r=k+1}^{i} \bar{\mathbf{Q}}_r \right]^2 \tilde{\mu}_i \tilde{\rho}_i + \tilde{\rho}_{k+1} \tilde{\mu}_{k+1} \bar{\mathbf{Q}}_{k+1}^2 \right] \right].
\end{aligned} \tag{52}
$$

where $\bar{\mathbf{Q}}_j = \left(\mathbf{I} + \tilde{m}_p^j(-1)\, \mathbf{\Sigma}_j \mathbf{M}_j^{-1} \right)^{-1}$, *with* $\tilde{m}_p^j(-1)$ *being the unique positive solution of*

$$
\frac{1}{\tilde{m}_p^j(-1)} = 1 + \frac{1}{n} \sum_{i=1}^{d} \frac{\sigma_i^j (m_i^j)^{-1}}{1 + \tilde{m}_p^j(-1)\, \sigma_i^j (m_i^j)^{-1}},
$$

and,

$$
\begin{aligned}
\tilde{\rho}_j &= \frac{1}{n} \mathrm{Tr} \left[\mathbf{\Sigma}_j \mathbf{M}_j^{-2} \left[\left[\prod_{m=j}^{T} \bar{\mathbf{Q}}_m \right]^2 + \sum_{i=j+1}^{T} \mathbf{\Sigma}_i \left[\prod_{r=j}^{i} \bar{\mathbf{Q}}_r \right]^2 \tilde{\mu}_i \tilde{\rho}_i \right] \right] \\
\tilde{\mu}_j &= \left[\left[1 + \frac{1}{n} \mathrm{Tr} \left(\mathbf{\Sigma}_j \mathbf{M}_j^{-1} \bar{\mathbf{Q}}_j \right) \right]^2 - \frac{1}{n} \mathrm{Tr} \left(\mathbf{\Sigma}_j \mathbf{M}_j^{-1} \bar{\mathbf{Q}}_j \right)^2 \right]^{-1}.
\end{aligned}
$$

Here σ_i^j, m_i^j denote respectively, the i'th eigenvalues of the matrices $\mathbf{\Sigma}_j, \mathbf{M}_j$.

Proof. For each $t \in [1, T]$, we are interested in solving

$$
\mathbf{w}_t^{\text{gen}} = \min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}_t \mathbf{w} - \mathbf{y}\|_2^2 + \left\| \mathbf{w} - \mathbf{w}_{t-1}^{\text{gen}} \right\|_{\mathbf{M}_t}^2.
$$

By Lemma 8 we have,

$$
\mathbf{w}_t^{\text{gen}} = \left(\frac{1}{n} \mathbf{X}_t^\top \mathbf{X}_t + \mathbf{M}_t \right)^{-1} \left(\frac{1}{n} \mathbf{X}_t^\top \mathbf{y}_t + \mathbf{M}_t \mathbf{w}_{t-1}^{\text{gen}} \right)
$$

so,

$$
\begin{aligned}
\mathbf{w}_T^{\text{gen}} - \mathbf{w}^\star &= \left(\frac{1}{n} \mathbf{X}_T^\top \mathbf{X}_T + \mathbf{M}_T \right)^{-1} \left(\frac{1}{n} \mathbf{X}_T^\top (\mathbf{X}_T \mathbf{w}^\star + \mathbf{z}_T) + \mathbf{M}_T \mathbf{w}_{T-1}^{\text{gen}} \right) - \mathbf{w}^\star \\
&= \left(\frac{1}{n} \mathbf{X}_T^\top \mathbf{X}_T + \mathbf{M}_T \right)^{-1} \left(\frac{1}{n} \mathbf{X}_T^\top (\mathbf{X}_T \mathbf{w}^\star + \mathbf{z}_T) + \mathbf{M}_T \mathbf{w}_{T-1}^{\text{gen}} - \left(\frac{1}{n} \mathbf{X}_T^\top \mathbf{X}_T + \mathbf{M}_T \right) \mathbf{w}^\star \right) \\
&= \left(\frac{1}{n} \mathbf{X}_T^\top \mathbf{X}_T + \mathbf{M}_T \right)^{-1} \left(\frac{1}{n} \mathbf{X}_T^\top \mathbf{z}_T + \mathbf{M}_T \left(\mathbf{w}_{T-1}^{\text{gen}} - \mathbf{w}^\star \right) \right)
\end{aligned}
$$

$$= \mathbf{P}_T \mathbf{M}_T \left(\mathbf{w}_{T-1}^{\text{gen}} - \mathbf{w}^\star\right) + \mathbf{P}_T \frac{1}{n} \mathbf{X}_T^\top \mathbf{z}_T$$
$$= \dots$$
$$= \left[\prod_{m=T}^{1} \mathbf{P}_m \mathbf{M}_m\right] (\mathbf{w}_0 - \mathbf{w}^\star) + \sum_{k=1}^{T} \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M}_m\right] \mathbf{P}_k \frac{1}{n} \mathbf{X}_k^\top \mathbf{z}_k,$$

where $\mathbf{P}_t = \left(\frac{1}{n}\mathbf{X}_t^\top \mathbf{X}_t + \mathbf{M}_t\right)^{-1} = \mathbf{M}_t^{-\frac{1}{2}} \left(\mathbf{I} + \frac{1}{n}\mathbf{M}_t^{-\frac{1}{2}} \mathbf{X}_t^\top \mathbf{X}_t \mathbf{M}_t^{-\frac{1}{2}}\right)^{-1} \mathbf{M}_t^{-\frac{1}{2}}$ for all $t \in [1, T]$.

Note that $\mathbf{X}_t$ can be re-written as $\mathbf{X}_t = \mathbf{Z}_t \mathbf{\Sigma}_t^{1/2}$ where $\mathbf{Z}_t$ has i.i.d. entries with zero mean and unit variance. Substituting into $G_T^{\text{gen}} = \left\|\mathbf{w}_T^{\text{gen}} - \mathbf{w}^\star\right\|^2$,

$$\begin{aligned} G_T^{\text{gen}} = &\left[\left[\prod_{m=T}^{1} \mathbf{P}_m \mathbf{M}_m\right] (\mathbf{w}_0 - \mathbf{w}^\star) + \sum_{k=1}^{T} \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M}_m\right] \mathbf{P}_k \frac{1}{n} \mathbf{X}_k^\top \mathbf{z}_k\right]^\top \\ &\left[\left[\prod_{m=T}^{1} \mathbf{P}_m \mathbf{M}_m\right] (\mathbf{w}_0 - \mathbf{w}^\star) + \sum_{k=1}^{T} \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M}_m\right] \mathbf{P}_k \frac{1}{n} \mathbf{X}_k^\top \mathbf{z}_k\right] \\ = &\underbrace{(\mathbf{w}_0 - \mathbf{w}^\star)^\top \left[\prod_{m=T}^{1} \mathbf{P}_m \mathbf{M}_m\right]^\top \left[\prod_{m=T}^{1} \mathbf{P}_m \mathbf{M}_m\right] (\mathbf{w}_0 - \mathbf{w}^\star)}_{\text{term 1}} \\ &+ \underbrace{\frac{v_z}{n^2} \sum_{k=1}^{T} \mathbf{z}_k^\top \mathbf{X}_k \mathbf{P}_k \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M}_m\right]^\top \sum_{k'=1}^{T} \left[\prod_{m=T}^{k'+1} \mathbf{P}_m \mathbf{M}_m\right] \mathbf{P}_{k'} \mathbf{X}_{k'}^\top \mathbf{z}_{k'}}_{\text{term 2}}. \end{aligned} \tag{53}$$

Let $\mathbf{A}_k := \mathbf{M}_k^{-\frac{1}{2}} \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M}_m\right]^\top \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M}_m\right] \mathbf{M}_k^{-\frac{1}{2}}$. Using this definition we can find the equivalent to $\left[\prod_{m=T}^{1} \mathbf{P}_m \mathbf{M}_m\right]^\top \left[\prod_{m=T}^{1} \mathbf{P}_m \mathbf{M}_m\right]$, in the following manner,

$$\begin{aligned} &\left[\prod_{m=T}^{1} \mathbf{P}_m \mathbf{M}_m\right]^\top \left[\prod_{m=T}^{1} \mathbf{P}_m \mathbf{M}_m\right] = \mathbf{M}_1 \left[\mathbf{P}_1 \dots \mathbf{M}_T \mathbf{P}_T \mathbf{P}_T \dots \mathbf{P}_1\right] \mathbf{M}_1 \\ &= \mathbf{M}_1^{\frac{1}{2}} \left[\underbrace{\left[\mathbf{M}_1^{\frac{1}{2}} \mathbf{P}_1 \mathbf{M}_1^{\frac{1}{2}}\right]}_{\mathbf{Q}_1(-1)} \underbrace{\mathbf{M}_1^{-\frac{1}{2}} \mathbf{M}_2 \dots \mathbf{M}_T \mathbf{P}_T \mathbf{P}_T \mathbf{M}_T \dots \mathbf{M}_2 \mathbf{M}_1^{-\frac{1}{2}}}_{\mathbf{A}_1} \underbrace{\left[\mathbf{M}_1^{\frac{1}{2}} \mathbf{P}_1 \mathbf{M}_1^{\frac{1}{2}}\right]}_{\mathbf{Q}_1(-1)}\right] \mathbf{M}_1^{\frac{1}{2}} \end{aligned} \tag{54}$$

where we let $\mathbf{Q}_t(z) := \left(\frac{1}{n}\mathbf{M}_t^{-\frac{1}{2}} \mathbf{\Sigma}_t^{\frac{1}{2}} \mathbf{Z}^\top \mathbf{Z} \mathbf{\Sigma}_t^{\frac{1}{2}} \mathbf{M}_t^{-\frac{1}{2}} - z\mathbf{I}\right)^{-1}$ with $\mathbf{Z}$ having i.i.d. entries with zero mean and unit variance, then by Theorem 6, $\mathbf{Q}_t(-1)$ admits a deterministic equivalent, where,

$$\mathbf{Q}_t(-1) \asymp \bar{\mathbf{Q}}_t(-1) = \left(\mathbf{I} + \tilde{m}_p^t(-1)\, \mathbf{\Sigma}_t \mathbf{M}_t^{-1}\right)^{-1} \text{ and } \frac{1}{\tilde{m}_p^t(-1)} = 1 + \frac{1}{n} \sum_{i=1}^{d} \frac{\sigma_i^t (m_i^t)^{-1}}{1 + \tilde{m}_p^t(-1)\, \sigma_i^t (m_i^t)^{-1}}.$$

Here σ_i^t, m_i^t denote respectively, the i'th eigenvalues of the matrices $\mathbf{\Sigma}_t, \mathbf{M}_t$. For notational convenience, we set $\mathbf{Q}_t := \mathbf{Q}_t(-1)$, and $\bar{\mathbf{Q}}_t := \bar{\mathbf{Q}}_t(-1)$. Both matrices are diagonal since both $\mathbf{\Sigma}_t$ and $\mathbf{M}_t$ are diagonal for all $t \in [1, T]$. Moreover, as previously mentioned, we note that under our standing assumption, we have $\sigma_i^t = v_x$, for all i and t. Nevertheless, our analysis accommodates non-isotropic covariance matrices.

Using Theorem 7, we get,

$$\begin{aligned} \mathbf{A}_k &= \mathbf{M}_k^{-\frac{1}{2}} \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M}_m\right]^\top \left[\prod_{m=T}^{k+1} \mathbf{P}_m \mathbf{M}_m\right] \mathbf{M}_k^{-\frac{1}{2}} \\ &\asymp \mathbf{M}_k^{-\frac{1}{2}} \mathbf{M}_{k+1}^{\frac{1}{2}} \left(\bar{\mathbf{Q}}_{k+1} \mathbf{A}_{k+1} \bar{\mathbf{Q}}_{k+1} + \tilde{\mu}_{k+1} \tilde{\rho}_{k+1} \mathbf{D}_{k+1}\right) \mathbf{M}_{k+1}^{\frac{1}{2}} \mathbf{M}_k^{-\frac{1}{2}} \end{aligned}$$

$$
\begin{aligned}
\asymp & \mathbf{M}_k^{-\frac{1}{2}}\mathbf{M}_{k+1}^{\frac{1}{2}}\Big(\bar{\mathbf{Q}}_{k+1}\Big(\mathbf{M}_{k+1}^{-\frac{1}{2}}\mathbf{M}_{k+2}^{\frac{1}{2}}\Big(\bar{\mathbf{Q}}_{k+2}\mathbf{A}_{k+2}\bar{\mathbf{Q}}_{k+2} \\
& \qquad + \tilde{\mu}_{k+2}\tilde{\rho}_{k+2}\mathbf{D}_{k+2}\Big)\mathbf{M}_{k+2}^{\frac{1}{2}}\mathbf{M}_{k+1}^{-\frac{1}{2}}\Big)\bar{\mathbf{Q}}_{k+1} + \tilde{\mu}_{k+1}\tilde{\rho}_{k+1}\mathbf{D}_{k+1}\Big)\mathbf{M}_{k+1}^{\frac{1}{2}}\mathbf{M}_k^{-\frac{1}{2}} \\
\asymp & \mathbf{M}_k^{-\frac{1}{2}}\mathbf{M}_T^{\frac{1}{2}}\left[\prod_{m=k+1}^{T}\bar{\mathbf{Q}}_m\right]\underbrace{\mathbf{A}_T}_{\mathbf{M}_T^{-1}}\left[\prod_{m=k+1}^{T}\bar{\mathbf{Q}}_m\right]\mathbf{M}_T^{\frac{1}{2}}\mathbf{M}_k^{-\frac{1}{2}} \\
& \qquad + \mathbf{M}_k^{-\frac{1}{2}}\sum_{i=k+1}^{T}\mathbf{M}_i^{\frac{1}{2}}\left[\prod_{r=k+1}^{i-1}\bar{\mathbf{Q}}_r\right]\tilde{\mu}_i\tilde{\rho}_i\mathbf{D}_i\left[\prod_{r=k+1}^{i-1}\bar{\mathbf{Q}}_r\right]\mathbf{M}_i^{\frac{1}{2}}\mathbf{M}_k^{-\frac{1}{2}},
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{\rho}_k &= \frac{1}{n}\mathrm{Tr}\left[\mathbf{\Sigma}_k\mathbf{M}_k^{-1}\bar{\mathbf{Q}}_k\mathbf{A}_k\bar{\mathbf{Q}}_k\right], \\
\tilde{\mu}_k &= \left[\left[1+\frac{1}{n}\mathrm{Tr}\left(\mathbf{\Sigma}_k\mathbf{M}_k^{-1}\bar{\mathbf{Q}}_k\right)\right]^2 - \frac{1}{n}\mathrm{Tr}\left(\mathbf{\Sigma}_k\mathbf{M}_k^{-1}\bar{\mathbf{Q}}_k\right)^2\right]^{-1}, \\
\mathbf{D}_k &= \bar{\mathbf{Q}}_k\mathbf{\Sigma}_k\mathbf{M}_k^{-1}\bar{\mathbf{Q}}_k.
\end{aligned}
$$

Substituting back into Equation (54) and using Proposition 1,

$$
\begin{aligned}
\left[\prod_{m=T}^{k}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k}\mathbf{P}_m\mathbf{M}_m\right] &= \mathbf{M}_k^{\frac{1}{2}}\mathbf{Q}_k\mathbf{A}_k\mathbf{Q}_k\mathbf{M}_k^{\frac{1}{2}} \\
&\asymp \mathbf{M}_k^{\frac{1}{2}}\left[\bar{\mathbf{Q}}_k\mathbf{A}_k\bar{\mathbf{Q}}_k + \tilde{\rho}_k\tilde{\mu}_k\bar{\mathbf{Q}}_k\mathbf{\Sigma}_k\mathbf{M}_k^{-1}\bar{\mathbf{Q}}_k\right]\mathbf{M}_k^{\frac{1}{2}} \\
&\asymp \left[\prod_{m=k}^{T}\bar{\mathbf{Q}}_m\right]^2 + \sum_{i=k+1}^{T}\left[\prod_{r=k}^{i}\bar{\mathbf{Q}}_r\right]^2\mathbf{\Sigma}_i\tilde{\mu}_i\tilde{\rho}_i + \tilde{\rho}_k\tilde{\mu}_k\mathbf{\Sigma}_k\bar{\mathbf{Q}}_k^2. \qquad (55)
\end{aligned}
$$

Additionally, using Proposition 1, we have,

$$
\begin{aligned}
\tilde{\rho}_k &= \frac{1}{n}\mathrm{Tr}\left[\mathbf{\Sigma}_k\mathbf{M}_k^{-1}\bar{\mathbf{Q}}_k\mathbf{A}_k\bar{\mathbf{Q}}_k\right] = \frac{1}{n}\mathrm{Tr}\left[\mathbf{M}_k^{-1}\bar{\mathbf{Q}}_k^2\mathbf{A}_k\right] \\
&\to \frac{1}{n}\mathrm{Tr}\left[\mathbf{\Sigma}_k\mathbf{M}_k^{-2}\left[\left[\prod_{m=k}^{T}\bar{\mathbf{Q}}_m\right]^2 + \sum_{i=k+1}^{T}\mathbf{\Sigma}_i\left[\prod_{r=k}^{i}\bar{\mathbf{Q}}_r\right]^2\tilde{\mu}_i\tilde{\rho}_i\right]\right].
\end{aligned}
$$

almost surely.

And term 2 in Equation (53) is,

$$
\begin{aligned}
&\frac{1}{n^2}\left[\sum_{k,k'=1}^{T}\mathbf{z}_k^\top\mathbf{W}_k\mathbf{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k'+1}\mathbf{P}_m\mathbf{M}_m\right]\mathbf{P}_{k'}\mathbf{\Sigma}_{k'}^{\frac{1}{2}}\mathbf{W}_{k'}^\top\mathbf{z}_{k'}\right] \\
&\overset{[1]}{\underset{P}{\longrightarrow}}\frac{1}{n^2}\left[\sum_{k=1}^{T}\mathbf{z}_k^\top\mathbf{W}_k\mathbf{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]\mathbf{P}_k\mathbf{\Sigma}_k^{\frac{1}{2}}\mathbf{W}_k^\top\mathbf{z}_k\right] \\
&= \frac{1}{n^2}\left[\sum_{k=1}^{T}\mathrm{Tr}\left[\mathbf{z}_k^\top\mathbf{W}_k\mathbf{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]\mathbf{P}_k\mathbf{\Sigma}_k^{\frac{1}{2}}\mathbf{W}_k^\top\mathbf{z}_k\right]\right] \\
&= \frac{1}{n^2}\left[\sum_{k=1}^{T}\mathrm{Tr}\left[\mathbf{z}_k\mathbf{z}_k^\top\mathbf{W}_k\mathbf{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]\mathbf{P}_k\mathbf{\Sigma}_k^{\frac{1}{2}}\mathbf{W}_k^\top\right]\right] \\
&= \frac{1}{n^2}\sum_{k=1}^{T}\mathrm{Tr}\left[\mathbf{z}_k\mathbf{z}_k^\top\left[\mathbf{W}_k\mathbf{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]\mathbf{P}_k\mathbf{\Sigma}_k^{\frac{1}{2}}\mathbf{W}_k^\top\right]\right]
\end{aligned}
$$

$$
\begin{aligned}
&\xrightarrow[P]{[2]} \frac{v_z}{n^2}\left[\sum_{k=1}^{T}\mathrm{Tr}\left[\mathbf{W}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]\mathbf{P}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{W}_k^\top\right]\right]\\
&=\frac{v_z}{n^2}\left[\sum_{k=1}^{T}\mathrm{Tr}\left[\mathbf{P}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{W}_k^\top\mathbf{W}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]\right]\right]\\
&=\frac{v_z}{n^2}\mathrm{Tr}\left[\sum_{k=1}^{T}\left[\mathbf{P}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{W}_k^\top\mathbf{W}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k\right]\left[\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]\right]\right]
\end{aligned}
$$

where [1] and [2] follows since the noise $\mathbf{z}_k$ is sample independently of the data matrices $\mathbf{X}_1,\ldots,\mathbf{X}_T$ and $\mathbb{E}[\mathbf{z}_k\mathbf{z}_k^\top]=v_z\mathbf{I}$ with the weak law of large numbers.

We already know the equivalent corresponding to $\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]$ from Equation (55). Hence, we now focus on the remaining term.

$$
\begin{aligned}
&\mathbf{P}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{W}_k^\top\mathbf{W}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k\\
&=n^2\left(\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{W}_k^\top\mathbf{W}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}+n\mathbf{M}_k\right)^{-1}\left(\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{W}_k^\top\mathbf{W}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}+n\mathbf{M}_k-n\mathbf{M}_k\right)\left(\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{W}_k^\top\mathbf{W}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}+n\mathbf{M}_k\right)^{-1}\\
&=n\left(\mathbf{P}_k-\mathbf{P}_k\mathbf{M}_k\mathbf{P}_k\right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\overline{\frac{1}{n}\mathbf{P}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{W}_k^\top\mathbf{W}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k}=\mathbf{M}_k^{-\frac{1}{2}}\bar{\mathbf{Q}}_k\mathbf{M}_k^{-\frac{1}{2}}-\overline{\mathbf{M}_k^{-\frac{1}{2}}\mathbf{Q}_k^2\mathbf{M}_k^{-\frac{1}{2}}}\\
&=\mathbf{M}_k^{-\frac{1}{2}}\Bigg(\bar{\mathbf{Q}}_k-\left[\bar{\mathbf{Q}}_k^2+\frac{1}{n}\mathrm{Tr}\left[\boldsymbol{\Sigma}_k\mathbf{M}_k^{-1}\bar{\mathbf{Q}}_k^2\right]\right.\\
&\qquad\left.\left[\left[1+\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}_k\mathbf{M}_k^{-1}\bar{\mathbf{Q}}_k\right)\right]^2-\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}_k\mathbf{M}_k^{-1}\bar{\mathbf{Q}}_k\right)^2\right]^{-1}\boldsymbol{\Sigma}_k\mathbf{M}_k^{-1}\bar{\mathbf{Q}}_k^2\right]\Bigg)\mathbf{M}_k^{-\frac{1}{2}}.
\end{aligned}
$$

Since both $\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]$ and $\mathbf{P}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{Z}_k^\top\mathbf{Z}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k$ have almost surely finite operator norm, then by the properties of deterministic equivalent we have

$$
\begin{aligned}
&\frac{v_z}{n^2}\mathrm{Tr}\left[\sum_{k=1}^{T}\mathbf{P}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{Z}_k^\top\mathbf{Z}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]\right]\\
&\xrightarrow{P}\frac{v_z}{n}\mathrm{Tr}\left[\sum_{k=1}^{T}\overline{\frac{1}{n}\mathbf{P}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{Z}_k^\top\mathbf{Z}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k}\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]\right]\\
&\xrightarrow{P}\frac{v_z}{n}\mathrm{Tr}\left[\sum_{k=1}^{T}\overline{\frac{1}{n}\mathbf{P}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{Z}_k^\top\mathbf{Z}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{P}_k}\overline{\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]^\top\left[\prod_{m=T}^{k+1}\mathbf{P}_m\mathbf{M}_m\right]}\right],
\end{aligned}
$$

where in each step we condition on the event that each one of the matrices has finite operator norm and then use Definition 3. Summing the terms, we get

$$
\begin{aligned}
G_T^{\mathrm{gen}}\xrightarrow{P}&(\mathbf{w}_0-\mathbf{w}^\star)^\top\left[\left[\prod_{m=1}^{T}\bar{\mathbf{Q}}_m\right]^2+\sum_{i=2}^{T}\boldsymbol{\Sigma}_i\left[\prod_{r=1}^{i}\bar{\mathbf{Q}}_r\right]^2\tilde{\mu}_i\tilde{\rho}_i+\tilde{\rho}_1\tilde{\mu}_1\bar{\mathbf{Q}}_1^2\right](\mathbf{w}_0-\mathbf{w}^\star)\\
&+\frac{v_z}{n}\mathrm{Tr}\left[\sum_{k=1}^{T}\bar{\mathbf{Q}}_k\mathbf{M}_k^{-1}\left[\mathbf{I}-\bar{\mathbf{Q}}_k\left[\mathbf{I}+\frac{1}{n}\mathrm{Tr}\left[\boldsymbol{\Sigma}_k^{-1}\mathbf{M}_k^{-1}\bar{\mathbf{Q}}_k^2\right]\tilde{\mu}_k\mathbf{M}_k^{-1}\right]\right]\right.\\
&\qquad\left.\left[\left[\prod_{m=k+1}^{T}\bar{\mathbf{Q}}_m\right]^2+\sum_{i=k+2}^{T}\boldsymbol{\Sigma}_i\left[\prod_{r=k+1}^{i}\bar{\mathbf{Q}}_r\right]^2\tilde{\mu}_i\tilde{\rho}_i+\tilde{\rho}_{k+1}\tilde{\mu}_{k+1}\bar{\mathbf{Q}}_{k+1}^2\right]\right].
\end{aligned}
$$

□

D.4 Proof of Theorem 4: A Failure of Isotropic Ridge Under Sparse Teachers

Recall Theorem 4. Assume that $\alpha := \frac{n}{d} < 1$. Let $\mathbf{w}_1^\star - \mathbf{w}_0$ be k-sparse with support S so that $\mathbf{w}_1^\star \neq \mathbf{w}_0$, and assume that $\mathbf{X}_S$ (the submatrix of $\mathbf{X}$ consisting of the columns indexed by S) has full column rank, then as $n, d \to \infty$ after a single task with no noise, it holds almost surely that,

$$\sup_{\mathbf{M}\in\mathcal{M}} \inf_{\lambda>0} \frac{G_1^{\text{ridge}}}{G_1^{\text{gen}}} = \infty.$$

where $\mathcal{M} = \bigcup_{0<l\le u<\infty} \{\mathbf{M} : l\mathbf{I}_d \preceq \mathbf{M} \preceq u\mathbf{I}_d\}$ and u, l are constants independent of n, d.

Proof. In this proof, we show that ridge has an error floor in the overparmaterized regime, while the generalized ridge can be optimized to sufficiently small error.

Since $\mathbf{y} = \mathbf{X}\mathbf{w}^\star$, the ridge solution is

$$\mathbf{w}_1^{\text{ridge}} = \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d\right)^{-1}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\mathbf{w}_1^\star + \lambda\mathbf{w}_0\right).$$

Equivalently,

$$\begin{aligned}\mathbf{w}_1^{\text{ridge}} - \mathbf{w}_1^\star &= \left(\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d\right)^{-1}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\mathbf{w}_1^\star + \lambda\mathbf{w}_0\right) - \mathbf{I}_d\right)\mathbf{w}_1^\star \\ &= -\lambda\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d\right)^{-1}(\mathbf{w}_1^\star - \mathbf{w}_0).\end{aligned}$$

Hence the squared estimation error is

$$G_1^{\text{ridge}} = \left\|\mathbf{w}_1^{\text{ridge}} - \mathbf{w}^\star\right\|_2^2 = \lambda^2(\mathbf{w}_1^\star - \mathbf{w}_0)^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d\right)^{-2}(\mathbf{w}_1^\star - \mathbf{w}_0).$$

where m_α is the Stieltjes transform of the MP law with aspect ratio α. Therefore, almost surely, as $d \to \infty$ by the Marchenko-Pastur law,

$$G_1^{\text{ridge}} = \|\mathbf{w}_1^{\text{ridge}} - \mathbf{w}^\star\|_2^2 \to \frac{\lambda^2}{v_x^2} m'_\alpha\left(-\frac{\lambda}{v_x}\right)\|\mathbf{w}^\star - \mathbf{w}_0\|^2.$$

$\frac{\lambda^2}{v_x^2} m'_\alpha\left(-\frac{\lambda}{v_x}\right) = \frac{1}{2}\left(1 - \alpha + \frac{\lambda(1+\alpha)+\alpha v_x(1-\alpha)^2}{\sqrt{\lambda^2 + \frac{2\lambda v_x(1+\alpha)}{\alpha} + \frac{v_x^2(1-\alpha)^2}{\alpha^2}}}\right)$ is monotonic increasing as function of λ as established in Karpel et al. (2026, Lemma E.3), therefore $\frac{\lambda^2}{v_x^2} m'_\alpha\left(-\frac{\lambda}{v_x}\right)$ is bounded below by $1 - \alpha > 0$, which implies that $\min_\lambda G_1^{\text{ridge}} > (1-\alpha)\|\mathbf{w_1}^\star - \mathbf{w}_0\|^2$, almost surely, as $d \to \infty$.

Write $\mathbf{P} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ and partition,

$$\mathbf{P} = \begin{pmatrix}\mathbf{P}_{SS} & \mathbf{P}_{SS^c} \\ \mathbf{P}_{S^cS} & \mathbf{P}_{S^cS^c}\end{pmatrix}, \qquad \mathbf{M}_{\tilde\lambda,\psi} = \begin{pmatrix}\tilde\lambda\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \psi\mathbf{I}\end{pmatrix}.$$

Without loss of generality, we assume $S = \{1, \dots, k\}$; otherwise, apply a permutation of coordinates. Since $\mathbf{y} = \mathbf{X}\mathbf{w}_1^\star$ with $\mathbf{w}_1^\star - \mathbf{w}_0 = ((\mathbf{w}_1^\star)_S - (\mathbf{w}_0)_S, \mathbf{0})$, we have

$$\tfrac{1}{n}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{w}_0) = \mathbf{P}(\mathbf{w}_1^\star - \mathbf{w}_0) = \begin{pmatrix}\mathbf{P}_{SS}\left((\mathbf{w}_1^\star)_S - (\mathbf{w}_0)_S\right) \\ \mathbf{P}_{S^cS}\left((\mathbf{w}_1^\star)_S - (\mathbf{w}_0)_S\right)\end{pmatrix}.$$

The estimator $\hat{\mathbf{w}} = (\hat{\mathbf{w}}_S, \hat{\mathbf{w}}_{S^c})$ satisfies

$$(\mathbf{P} + \mathbf{M}_{\tilde\lambda,\psi})(\hat{\mathbf{w}} - \mathbf{w}_0) = \mathbf{P}(\mathbf{w}_1^\star - \mathbf{w}_0),$$

that is,

$$\begin{aligned}(\mathbf{P}_{SS} + \tilde\lambda\mathbf{I})(\hat{\mathbf{w}}_S - (\mathbf{w}_0)_S) + \mathbf{P}_{SS^c}(\hat{\mathbf{w}}_{S^c} - (\mathbf{w}_0)_{S^c}) &= \mathbf{P}_{SS}((\mathbf{w}_1^\star)_S - (\mathbf{w}_0)_S), \\ \mathbf{P}_{S^cS}(\hat{\mathbf{w}}_S - (\mathbf{w}_0)_S) + (\mathbf{P}_{S^cS^c} + \psi\mathbf{I})(\hat{\mathbf{w}}_{S^c} - (\mathbf{w}_0)_{S^c}) &= \mathbf{P}_{S^cS}((\mathbf{w}_1^\star)_S - (\mathbf{w}_0)_S).\end{aligned}$$

Solving the second equation gives

$$\hat{\mathbf{w}}_{S^c} - (\mathbf{w}_0)_{S^c} = (\mathbf{P}_{S^cS^c} + \psi\mathbf{I})^{-1}\mathbf{P}_{S^cS}\left((\mathbf{w}_1^\star)_S - (\mathbf{w}_0)_S - (\hat{\mathbf{w}}_S - (\mathbf{w}_0)_S)\right).$$

Substituting into the first equation yields

$$\begin{aligned}\Big(\mathbf{P}_{SS} + \tilde{\lambda}\mathbf{I} - \mathbf{P}_{SS^c}(\mathbf{P}_{S^cS^c} + \psi\mathbf{I})^{-1}\mathbf{P}_{S^cS}\Big)\left(\hat{\mathbf{w}}_S - (\mathbf{w}_0)_S\right) \\ = \mathbf{P}_{SS}\left((\mathbf{w}_1^\star)_S - (\mathbf{w}_0)_S\right) - \mathbf{P}_{SS^c}(\mathbf{P}_{S^cS^c} + \psi\mathbf{I})^{-1}\mathbf{P}_{S^cS}\left((\mathbf{w}_1^\star)_S - (\mathbf{w}_0)_S\right).\end{aligned}$$

Since $\|(\mathbf{P}_{S^cS^c} + \psi\mathbf{I})^{-1}\|_2 \le \psi^{-1}$, the terms involving $(\mathbf{P}_{S^cS^c} + \psi\mathbf{I})^{-1}$ vanish as $\psi \to \infty$, giving

$$(\mathbf{P}_{SS} + \tilde{\lambda}\mathbf{I})\left(\hat{\mathbf{w}}_S - (\mathbf{w}_0)_S\right) \to \mathbf{P}_{SS}\left((\mathbf{w}_1^\star)_S - (\mathbf{w}_0)_S\right), \qquad \hat{\mathbf{w}}_{S^c} - (\mathbf{w}_0)_{S^c} \to \mathbf{0}.$$

Thus

$$\lim_{\psi\to\infty} \hat{\mathbf{w}}(\tilde{\lambda}, \psi) = \mathbf{w}_0 + \Big((\mathbf{P}_{SS} + \tilde{\lambda}\mathbf{I})^{-1}\mathbf{P}_{SS}\left((\mathbf{w}_1^\star)_S - (\mathbf{w}_0)_S\right), \mathbf{0}\Big).$$

Since $\mathbf{P}_{SS} \succ 0$, then $(\mathbf{P}_{SS} + \tilde{\lambda}\mathbf{I})^{-1}\mathbf{P}_{SS} \to \mathbf{I}$ as $\tilde{\lambda} \to 0$, hence $\mathbf{w}_1 \to \mathbf{w}_1^\star$, which implies that $G_1^{\text{gen}} \to 0$, proving the claim.

In the statement of Theorem 4 the supremum is taken over the set of matrices which are spectrally bounded. This is an immediate consequence of the proof. For each $q \ge 1$ we may define the set $\mathcal{M}_q = \{\frac{1}{q}\mathbf{I} \preceq \mathbf{M} \preceq q\mathbf{I}\} \in \mathcal{M}$ for this choice for all finite q we have that $\mathcal{M}_q$ is spectrally bounded. In the construction of the regularization matrix, we take $\psi \to \infty$ and $\tilde{\lambda} \to 0$, this corresponds to taking $\tilde{\lambda} = 1/q$ and $\psi = q$. The resulting matrix belongs to $\mathcal{M}_q$ for every finite q. The proof shows that, along this sequence, $G_1^{\text{gen}} \to 0$, while the optimal isotropic-ridge loss remains bounded away from zero. Consequently, $\lim_{q\to\infty} \sup_{\mathbf{M}\in\mathcal{M}_q} \inf_{\lambda>0} \frac{G_1^{\text{ridge}}}{G_1^{\text{gen}}} = \infty$. □

D.5 Necessary and Sufficient Symmetry Conditions For a Guaranteed Optimal Isotropic Diagonal Regularizer

Theorem 9 (Necessary and sufficient condition to a guaranteed optimal isotropic regularization)**.** *Assume that the generalization loss admits a unique diagonal minimizer* $\mathbf{M}^\star = \mathrm{diag}(m_1^\star, \ldots, m_d^\star)$*. Then* $\mathbf{M}^\star = \lambda^\star \mathbf{I}_d$ *for some* $\lambda^\star > 0$ *if and only if, for each coordinate, the teacher is equally far from the initialization (i.e.,* $|w_1^\star - (w_0)_1| = |w_2^\star - (w_0)_2| = \ldots = |w_d^\star - (w_0)_d|$*), either deterministically or with probability tending to 1 (as* $n, d \to \infty$*) for random teacher.*

Proof. We write $\boldsymbol{m} = (m_1, \ldots, m_d)$ and $\mathbf{M} = \mathrm{diag}(\boldsymbol{m})$,

Recall Theorem 3, under Assumption 2, the expected generalization loss after T tasks of the generalized ridge algorithm (Algorithm 3), satisfies,

$$G_T^{\mathrm{gen}} \xrightarrow{P} (\mathbf{w}_0 - \mathbf{w}^\star)^\top \left(\bar{\mathbf{Q}}^{2T} + \sum_{k=1}^{T} \bar{\mathbf{Q}}^{2k} \tilde{\mu} \tilde{\rho}_k \right) (\mathbf{w}_0 - \mathbf{w}^\star) \tag{56}$$

$$+ \frac{v_z}{n} \mathrm{Tr} \left[\sum_{k=1}^{T} \left[\mathbf{I} - \bar{\mathbf{Q}} \left[\mathbf{I} + \frac{1}{n} \mathrm{Tr} \left[\mathbf{M}^{-1} \bar{\mathbf{Q}}^2 \right] \tilde{\mu} \mathbf{M}^{-1} \right] \right] \left[\bar{\mathbf{Q}}^{2(T-k)} + \sum_{i=k+1}^{T} \bar{\mathbf{Q}}^{2(i-k)} \tilde{\mu} \tilde{\rho}_i \right] \mathbf{M}^{-1} \bar{\mathbf{Q}} \right] \tag{57}$$

where

$$\bar{\mathbf{Q}} = \left(\mathbf{I} + v_x \tilde{m}_p(-1) \mathbf{M}^{-1} \right)^{-1},$$

with $\tilde{m}_p(-1)$ being the unique positive solution of

$$\frac{1}{\tilde{m}_p(-1)} = 1 + \frac{1}{n} \sum_{i=1}^{d} \frac{v_x M_{i,i}^{-1}}{1 + \tilde{m}_p(-1) v_x M_{i,i}^{-1}}.$$

And,

$$\tilde{\rho}_k = \frac{1}{n} \mathrm{Tr} \Big[\mathbf{M}^{-2} \Big(\bar{\mathbf{Q}}^{2(T-k+1)} + \tilde{\mu} \sum_{i=k+1}^{T} \tilde{\rho}_i \bar{\mathbf{Q}}^{2(i-k+1)} \Big) \Big],$$

$$\tilde{\mu} = \left[\left[1 + \frac{1}{n} \mathrm{Tr} \left(\mathbf{M}^{-1} \bar{\mathbf{Q}} \right) \right]^2 - \frac{1}{n} \mathrm{Tr} \left(\mathbf{M}^{-1} \bar{\mathbf{Q}} \right)^2 \right]^{-1}.$$

With diagonal $\mathbf{M}$, the deterministic equivalent resolvent appearing in Equation (56) reduces to

$$\bar{\mathbf{Q}} = \left(\mathbf{I}_d + v_x \tilde{m}_p(-1) \mathbf{M}^{-1} \right)^{-1} = \left(\mathbf{I}_d + \tilde{m}_p(-1) v_x \mathbf{M}^{-1} \right)^{-1}. \tag{58}$$

Hence $\bar{\mathbf{Q}}$ is diagonal with entries

$$q_i = q_i(m_i) = \frac{1}{1 + \frac{\tilde{m}_p(-1) v_x}{m_i}} = \frac{m_i}{m_i + \tilde{m}_p(-1) v_x}.$$

All matrices built from $\bar{\mathbf{Q}}$, $\mathbf{M}^{\pm 1}$ (e.g. powers of $\bar{\mathbf{Q}}$, $\bar{\mathbf{Q}} \mathbf{M}^{-1} \bar{\mathbf{Q}}$, and trace terms such as $\frac{1}{n} \mathrm{Tr}[(\mathbf{M}^{-1} \bar{\mathbf{Q}})^k]$) are therefore diagonal, and depend on $\mathbf{m}$ only through the coordinate-wise pairs $(m_i, q_i(m_i))$ and through symmetric trace scalars.

Using the closed-form expression in Equation (56) and the diagonal structure from Equation (58), the expected error decomposes as

$$\mathcal{R}_T(\boldsymbol{m}) = \mathcal{R}_T^{\mathrm{teacher}}(\boldsymbol{m}) + \mathcal{R}_T^{\mathrm{noise}}(\boldsymbol{m}).$$

The teacher-dependent term can be written as

$$\mathcal{R}_T^{\mathrm{teacher}}(\boldsymbol{m}) = (\mathbf{w}^\star - \mathbf{w}_0)^\top \mathbf{B}_T(\boldsymbol{m}) (\mathbf{w}^\star - \mathbf{w}_0) = \sum_{i=1}^{d} b_{T,i}(\boldsymbol{m}) (w_i^\star - w_{0i})^2, \tag{59}$$

where $\mathbf{B}_T(\boldsymbol{m})$ is diagonal and $b_{T,i}(\boldsymbol{m})$ denotes its i'th diagonal entry. The noise term $\mathcal{R}_T^{\text{noise}}(\boldsymbol{m})$ does not depend on $\mathbf{w}^\star$ and depends on $\boldsymbol{m}$ only through symmetric trace scalars. Consequently, $\mathcal{R}_T^{\text{noise}}(\boldsymbol{m})$ is invariant under any permutation of the coordinates of $\boldsymbol{m}$.

Fix $\lambda > 0$ and consider the isotropic point $\boldsymbol{m} = \lambda\mathbf{1}$. By permutation symmetry of $\mathcal{R}_T^{\text{noise}}$, there exists a scalar $B_T(\lambda)$ such that

$$\frac{\partial}{\partial m_i}\mathcal{R}_T^{\text{noise}}(\boldsymbol{m})\Big|_{\boldsymbol{m}=\lambda\mathbf{1}} = B_T(\lambda) \qquad \text{for all } i.$$

Likewise, at $\boldsymbol{m} = \lambda\mathbf{1}$ all diagonal entries $b_{t,i}(\boldsymbol{m})$ coincide, and their coordinate derivatives are equal. We define

$$A_T(\lambda) := \frac{\partial}{\partial m_i} b_{T,i}(\boldsymbol{m})\Big|_{\boldsymbol{m}=\lambda\mathbf{1}}, \qquad B_T(\lambda) := \frac{\partial}{\partial m_i}\mathcal{R}_T^{\text{noise}}(\boldsymbol{m})\Big|_{\boldsymbol{m}=\lambda\mathbf{1}}.$$

Differentiating $\mathcal{R}_T(\boldsymbol{m})$ and evaluating at $\boldsymbol{m} = \lambda\mathbf{1}$ yields

$$\frac{\partial}{\partial m_i}\mathcal{R}_T(\boldsymbol{m})\Big|_{\boldsymbol{m}=\lambda\mathbf{1}} = (w_i^\star - w_{0i})^2 A_T(\lambda) + B_T(\lambda), \qquad i = 1, \dots, d. \tag{60}$$

Assume that $\mathbf{M} = \lambda^\star \mathbf{I}_d$ is a diagonal minimizer. Then the first-order optimality conditions imply that Equation (60) vanishes for every i, i.e.

$$(w_i^\star - w_{0i})^2 A_T(\lambda^\star) + B_T(\lambda^\star) = 0 \qquad \text{for all } i.$$

Subtracting the conditions for indices i and j gives

$$\left((w_i^\star - w_{0i})^2 - \left(w_j^\star - w_{0j}\right)^2\right) A_T(\lambda) = 0.$$

In the case of $\mathbf{M} = \lambda\mathbf{I}_d$. We have a compact closed solution given in Karpel et al. (2026, Theorem 3) where[7],

$$\left\|\mathbf{w}_T^{\text{ridge}} - \mathbf{w}^\star\right\|^2 \xrightarrow{P} v_z c\frac{1-a^T}{1-a} + a^T \left\|\mathbf{w}_0 - \mathbf{w}^\star\right\|^2.$$

and

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

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

By Karpel et al. (2026, Lemma D.3) $\frac{\mathrm{d}}{\mathrm{d}\lambda}a = \frac{2\tilde{\lambda}\alpha}{D^3 v_x}$ where $D = \sqrt{\tilde{\lambda}^2 + 2\tilde{\lambda}(1+\alpha) + (1-\alpha)^2}$ and $\tilde{\lambda} = \frac{\lambda\alpha}{v_x}$ which is therefore strictly positive for all $\lambda > 0$, hence $\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathcal{R}_T^{\text{noise}}(\lambda^\star\mathbf{1}) \neq 0$.[8] By permutation symmetry, $\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathcal{R}_T^{\text{noise}}(\lambda\mathbf{1}) = dB_T(\lambda)$, therefore $B_T(\lambda^\star) \neq 0$. Using the isotropic first-order conditions $(w_i^\star - w_{0i})^2 A_T(\lambda^\star) + B_T(\lambda^\star) = 0$ for all i, we conclude that $A_T(\lambda^\star) \neq 0$ (otherwise stationarity would force $B_T(\lambda^\star) = 0$), and therefore $(w_i^\star - w_{0i})^2 = \left(w_j^\star - w_{0j}\right)^2$ for all i, j deterministically, or with probability tending to 1 (as $n, d \to \infty$) for random teacher.

Conversely, assume $(w_1^\star - w_{01})^2 = \ldots = (w_d^\star - w_{0d})^2 = \tau^2$ deterministically, or with probability tending to 1 (as $n, d \to \infty$) for random teacher. Then Equation (59) reduces to $\tau^2 \sum_{i=1}^d b_{T,i}(\boldsymbol{m})$, which is permutation invariant. Together with the invariance of $\mathcal{R}_T^{\text{noise}}(\boldsymbol{m})$, this shows that $\mathcal{R}_T(\boldsymbol{m})$ is a symmetric function of $\boldsymbol{m}$.

Let $\boldsymbol{m}^\star$ be a diagonal minimizer. For any permutation π of $\{1, \dots, d\}$, symmetry implies $\mathcal{R}_T(\boldsymbol{m}^\star) = \mathcal{R}_T(\pi\boldsymbol{m}^\star)$. If the diagonal minimizer is unique, it must be fixed by all permutations, which is only possible if $m_1^\star = \ldots = m_d^\star$. Thus $\mathbf{M}^\star = \lambda^\star\mathbf{I}_d$ for some $\lambda^\star > 0$.

□

[7] While their expression is stated for the expectation, their analysis can be naturally extended to yield the corresponding convergence-in-probability statement under the same asymptotic regime.

[8] In general, if $\|\mathbf{w}^\star - \mathbf{w}_0\| = 0$, then the optimal value of the regularization parameter would force $\lambda^\star \to \infty$. Consequently, an optimal diagonal minimizer does not exist, contrary to the assumption of the Theorem.

E Proofs for Section 6

E.1 Proof of Theorem 5: Closed-Form Expression for Generalization Error Under the Sinusoidal Model

Recall Theorem 5. Under Assumption 3 we have,

$$G_\infty \xrightarrow{P} C_0 + \sqrt{C_1^2 + C_2^2} \sup_{T \geq 1} \cos\left(4\pi f T + \varphi\right)$$

where,

$$\begin{aligned}
\varphi &= 2\pi f - \arctan 2\left(C_2, C_1\right) \\
C_0 &= v_z c' \frac{1}{1-a'} + 2\left\|\mathbf{B}\right\|^2 \sin^2\left(\pi f\right)\left(\frac{1}{1-a'} - 1\right)\left(2c_2 - 1\right), \\
C_1 &= 4\left\|\mathbf{B}\right\|^2 \sin^2\left(\pi f\right)\left(\frac{1}{2}\left(c_1 - 1\right) + \left(c_2 - 1\right)\left(c_1 - 1\right) - c_4 c_3\right), \\
C_2 &= 4\left\|\mathbf{B}\right\|^2 \sin^2\left(\pi f\right)\left(-\frac{1}{2}c_3 + c_4\left(c_1 - 1\right) + c_2 c_3\right), \\
c_1 &= \frac{1 - a'\cos\left(4\pi f\right)}{1 - 2a'\cos\left(4\pi f\right) + a'^2}, \\
c_2 &= \frac{1 - b'\cos\left(2\pi f\right)}{1 - 2b'\cos\left(2\pi f\right) + b'^2}, \\
c_3 &= \frac{a'\sin\left(4\pi f\right)}{1 - 2a'\cos\left(4\pi f\right) + a'^2}, \\
c_4 &= \frac{b'\sin\left(2\pi f\right)}{1 - 2b'\cos\left(2\pi f\right) + b'^2}, \\
a' &= 1 + \sum_{k=0}^{2N}\binom{2N}{k}\left(-\gamma\right)^k \sum_{r=0}^{k-1}\frac{v_x^k}{r+1}\binom{k}{r}\binom{k-1}{r}\frac{1}{\alpha^r}, \\
b' &= 1 + \sum_{k=0}^{N}\binom{N}{k}\left(-\gamma\right)^k \sum_{r=0}^{k-1}\frac{v_x^k}{r+1}\binom{k}{r}\binom{k-1}{r}\frac{1}{\alpha^r}, \\
c' &= \sum_{s=0}^{2N-2} m_s \sum_{l=0}^{s}\binom{s}{l}\left(-\gamma\right)^{l+2}\sum_{r=0}^{l}\frac{v_x^{l+1}}{r+1}\binom{l+1}{r}\binom{l}{r}\frac{1}{\alpha^{r+1}}.
\end{aligned} \tag{61}$$

Proof. We first show that, for all $i \in [T]$,

$$\begin{aligned}
&\limsup_{T\to\infty} \left\|\mathbf{w}_T - \mathbf{w}_i^\star\right\|^2 \\
&\xrightarrow{P} \limsup_{T\to\infty} \Bigg\{ v_z c \frac{1}{1-a} + 2\left\|\mathbf{B}\right\|^2 \sin^2\left(\pi f\right) \\
&\qquad \left(\left(\frac{1}{1-a} - 1\right) + \cos\left(4\pi f T + 2\pi f\right)\left(c_1 - 1\right) + \sin\left(4\pi f T + 2\pi f\right) c_3\right) \\
&\quad + 4\left\|\mathbf{B}\right\|^2 \sin^2\left(\pi f\right)\Bigg\{\left(\frac{1}{1-a} - 1\right)\left(c_2 - 1\right) + \Big[\left(c_2 - 1\right)\cos\left(4\pi f T + 2\pi f\right) \\
&\qquad + \frac{b\sin\left(2\pi f\right)}{1 - 2b\cos\left(2\pi f\right) + b^2}\sin\left(4\pi f T + 2\pi f\right)\Big]\left(c_1 - 1\right)\Bigg\} \\
&\quad + \left[-c_4\cos\left(4\pi f T + 2\pi f\right) + \left(c_2 - 1\right)\sin\left(4\pi f T + 2\pi f\right)\right] c_3
\end{aligned}$$

$$- 8 \|\mathbf{B}\|^2 \cos(\pi f (T+i)) \sin(\pi f (T-i)) \sin(\pi f) (\cos(2\pi f T + \pi f)(c_2 - 1) + \sin(2\pi f T + \pi f) c_4)$$
$$+ 4 \|\mathbf{B}\|^2 \cos^2(\pi f (T+i)) \sin^2(\pi f (T-i)) \Bigg\}$$

Recall Theorem 1,

$$G_T^{\text{esgd}} \xrightarrow{P} v_z c' \frac{1 - a'^T}{1 - a'} + \frac{1}{T} \sum_{i=1}^{T} \|\mathbf{w}_T^\star - \mathbf{w}_i^\star\|^2 + 2\big(\mathbf{w}_T^\star - \bar{\mathbf{w}}^\star\big)^\top \sum_{i=1}^{T} b'^{T-i+1} \big(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\big)$$
$$+ \sum_{i=1}^{T} \sum_{i'=1}^{T} a'^{T - \max(i,i')+1} b'^{|i-i'|} \big(\mathbf{w}_{i-1}^\star - \mathbf{w}_i^\star\big)^\top \big(\mathbf{w}_{i'-1}^\star - \mathbf{w}_{i'}^\star\big),$$

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

$$a' = 1 + \sum_{k=0}^{2N} \binom{2N}{k} (-\gamma)^k \sum_{r=0}^{k-1} \frac{v_x^k}{r+1} \binom{k}{r} \binom{k-1}{r} \frac{1}{\alpha^r},$$
$$b' = 1 + \sum_{k=0}^{N} \binom{N}{k} (-\gamma)^k \sum_{r=0}^{k-1} \frac{v_x^k}{r+1} \binom{k}{r} \binom{k-1}{r} \frac{1}{\alpha^r},$$
$$c' = \sum_{s=0}^{2N-2} m_s \sum_{l=0}^{s} \binom{s}{l} (-\gamma)^{l+2} \sum_{r=0}^{l} \frac{v_x^{l+1}}{r+1} \binom{l+1}{r} \binom{l}{r} \frac{1}{\alpha^{r+1}}.$$

In particular, it is shown in the proof of the Theorem that for all $i \in [T]$ (see Equation (15)),

$$\left\|\mathbf{w}_T^{\text{esgd}} - \mathbf{w}_i^\star\right\|^2 \xrightarrow{P} a'^T \|\mathbf{w}_0 - \mathbf{w}_1^\star\|^2 + v_z c' \frac{1 - a'^T}{1 - a'}$$
$$+ \underbrace{2 \sum_{k=2}^{T} a'^{T-k+1} b'^{k-1} \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top (\mathbf{w}_0 - \mathbf{w}_1^\star)}_{\text{term 1}} + 2 b'^T \left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right)^\top (\mathbf{w}_0 - \mathbf{w}_1^\star)$$
$$+ \underbrace{\sum_{k=2}^{T} \sum_{k'=2}^{T} a'^{T-(k\vee k')+1} b'^{|k-k'|} \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)^\top \left(\mathbf{w}_{k'-1}^\star - \mathbf{w}_{k'}^\star\right)}_{\text{term 2}}$$
$$+ \underbrace{2 \sum_{k=2}^{T} b'^{T-k+1} \left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right)^\top \left(\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star\right)}_{\text{term 3}} + \|\mathbf{w}_T^\star - \mathbf{w}_i^\star\|^2 \tag{62}$$

Let $\mathbf{w}_i^\star = \mathbf{A} + \sin(2\pi f i)\,\mathbf{B}$, for all $k \in [2, T]$ we have

$$\begin{aligned}
\mathbf{w}_{k-1}^\star - \mathbf{w}_k^\star &= \mathbf{A} + \sin(2\pi f (k-1)) \cdot \mathbf{B} - (\mathbf{A} + \sin(2\pi f k) \cdot \mathbf{B}) \\
&= \mathbf{B} (\sin(2\pi f (k-1)) - \sin(2\pi f k)) \\
&= -2\mathbf{B} \sin(\pi f) \cos(\pi f (2k-1))
\end{aligned}$$

and for all $i \in [1, T]$,

$$\begin{aligned}
\mathbf{w}_T^\star - \mathbf{w}_i^\star &= \mathbf{A} + \sin(2\pi f T) \cdot \mathbf{B} - (\mathbf{A} + \sin(2\pi f i) \cdot \mathbf{B}) \\
&= \mathbf{B} (\sin(2\pi f T) - \sin(2\pi f i)) \\
&= 2\mathbf{B} \cos(\pi f (T+i)) \sin(\pi f (T-i))
\end{aligned}$$

We first evaluate the following geometric sums,

$$\begin{aligned}
&\sum_{n=0}^{\infty} q^n \cos(nx) = \frac{1}{2}\sum_{n=0}^{\infty}\left[\left(qe^{ix}\right)^n + \left(qe^{-ix}\right)^n\right] = \frac{1}{2}\left[\frac{1}{1-qe^{ix}} + \frac{1}{1-qe^{-ix}}\right] \\
&= \frac{1}{2}\left[\frac{1-qe^{-ix}}{(1-qe^{ix})(1-qe^{-ix})} + \frac{1-qe^{ix}}{(1-qe^{-ix})(1-qe^{ix})}\right] \\
&= \frac{1}{2}\left[\frac{2-q\left(e^{-ix}+e^{ix}\right)}{1-qe^{ix}-qe^{-ix}+q^2}\right] = \frac{1-q\cos(x)}{1-2q\cos(x)+q^2}
\end{aligned} \tag{63}$$

$$\begin{aligned}
&\sum_{n=0}^{\infty} q^n \sin(nx) = \frac{1}{2i}\sum_{n=0}^{\infty}\left[\left(qe^{ix}\right)^n - \left(qe^{-ix}\right)^n\right] = \frac{1}{2i}\left[\frac{1}{1-qe^{ix}} - \frac{1}{1-qe^{-ix}}\right] \\
&= \frac{1}{2i}\left[\frac{1-qe^{-ix}}{(1-qe^{ix})(1-qe^{-ix})} - \frac{1-qe^{ix}}{(1-qe^{-ix})(1-qe^{ix})}\right] \\
&= \frac{1}{2i}\left[\frac{1-qe^{-ix}}{1-qe^{ix}-qe^{-ix}+q^2} - \frac{1-qe^{ix}}{1-qe^{ix}-qe^{-ix}+q^2}\right] \\
&= \frac{1}{2i}\frac{q\left(e^{ix}-e^{-ix}\right)}{1-2q\cos(x)+q^2} = \frac{q\sin(x)}{1-2q\cos(x)+q^2}
\end{aligned} \tag{64}$$

Then, we proceed with evaluating term 1 in Equation (62)

$$\left|\limsup_{T\to\infty}\sum_{k=2}^{T} a'^{T-k+1}b'^{k-1}\cos(\pi f(2k-1))\right| \le \limsup_{T\to\infty}\sum_{k=2}^{T} a'^{T-k+1}b'^{k-1} = \limsup_{T\to\infty}\sum_{k=2}^{T} a'^{T+1}\left(\frac{b'}{a'}\right)^{k-1}$$

$$= \limsup_{T\to\infty} a'^{T+1}\sum_{j=1}^{T-1}\left(\frac{b'}{a'}\right)^{j} = \limsup_{T\to\infty} a'^{T+1}\frac{\frac{b'}{a'}\left(\left(\frac{b'}{a'}\right)^{T-1}-1\right)}{\frac{b'}{a'}-1} \overset{[1]}{=} \limsup_{T\to\infty}\frac{a'b'^T - a'^T b'}{b'-a'} = 0.$$

Where [1] follows as $a', b' < 1$ for $\gamma \in \left(0, 2v_x^{-1}\left(1+\alpha^{-1/2}\right)^{-2}\right)$. That implies,

$$\limsup_{T\to\infty}\sum_{k=2}^{T} a'^{T-k+1}b'^{k-1}\cos(\pi f(2k-1)) = 0.$$

And term 2 in Equation (62) becomes,

$$\begin{aligned}
&\limsup_{T\to\infty} 4\|\mathbf{B}\|^2\sin^2(\pi f)\sum_{k=2}^{T}\sum_{k'=2}^{T} a'^{T-(k\vee k')+1}b'^{|k-k'|}\cos(\pi f(2k-1))\cos(\pi f(2k'-1)) \\
&= \limsup_{T\to\infty} 4\|\mathbf{B}\|^2\sin^2(\pi f)\sum_{k=0}^{T-2} a'^{T-k-1}\cos^2(\pi f(2k+3)) \\
&\quad + 8\|\mathbf{B}\|^2\sin^2(\pi f)\sum_{p=0}^{T-3} b'^{p+1}\sum_{k=1}^{T-p-2} a'^{k}\cos(\pi f(2(T-k-p)-1))\cos(\pi f(2(T-k)+1)) \\
&= \limsup_{T\to\infty} 4\|\mathbf{B}\|^2\sin^2(\pi f)\sum_{k=1}^{T-1} a'^{k}\left(\frac{1+\cos(2(\pi f(2(T-k)+1)))}{2}\right) \\
&\quad + 8\|\mathbf{B}\|^2\sin^2(\pi f)\sum_{p=0}^{T-3} b'^{p+1}\sum_{r=1}^{T-p-1} a'^{r}\left(\frac{1}{2}\cos(-2\pi f(1+p)) + \frac{1}{2}\cos(4\pi f(T-r) - 2p\pi f)\right) \\
&= \limsup_{T\to\infty} 4\|\mathbf{B}\|^2\sin^2(\pi f)\sum_{k=1}^{T-1} a'^{k}\left(\frac{1+\cos(-4\pi fk + 4\pi fT + 2\pi f)}{2}\right) \\
&\quad + 4\|\mathbf{B}\|^2\sin^2(\pi f)\sum_{p=0}^{T-2} b'^{p+1}\sum_{r=1}^{T-p-1} a'^{r}\left(\cos(-2\pi f(1+p)) + \cos(-4\pi fr + 4\pi fT - 2p\pi f)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \limsup_{T\to\infty} 2\left\|\mathbf{B}\right\|^2 \sin^2(\pi f) \\
&\quad \left[\sum_{k=1}^{T-1} a'^k + \sum_{k=1}^{T-1} a'^k \left(\cos(-4\pi fk)\cos(4\pi fT + 2\pi f) - \sin(-4\pi fk)\sin(4\pi fT + 2\pi f)\right)\right] \\
&\quad + 4\left\|\mathbf{B}\right\|^2 \sin^2(\pi f) \Bigg[\sum_{p=0}^{T-2} b'^{p+1} \sum_{r=1}^{T-p-1} a'^r \cos(-2\pi f(1+p)) \\
&\qquad\qquad + \cos(4\pi fr)\cos(4\pi fT - 2p\pi f) + \sin(4\pi fr)\sin(4\pi fT - 2p\pi f)\Bigg] \\
&= \limsup_{T\to\infty} 2\left\|\mathbf{B}\right\|^2 \sin^2(\pi f) \\
&\quad \left(\left(\frac{1}{1-a'} - 1\right) + \cos(4\pi fT + 2\pi f)(c_1 - 1) + \sin(4\pi fT + 2\pi f)\, c_3\right) \\
&\quad + 4\left\|\mathbf{B}\right\|^2 \sin^2(\pi f) \left[\sum_{p=0}^{T-2} b'^{p+1}\left[\cos(-2\pi f(1+p))\left(\frac{1}{1-a'} - 1\right)\right.\right. \\
&\qquad\qquad \left.\left. + \cos(4\pi fT - 2p\pi f)(c_1 - 1) + \sin(4\pi fT - 2p\pi f)\, c_3\right]\right] \\
&= \limsup_{T\to\infty} 2\left\|\mathbf{B}\right\|^2 \sin^2(\pi f)\left(\left(\frac{1}{1-a'} - 1\right) + \cos(4\pi fT + 2\pi f)(c_1 - 1) + \sin(4\pi fT + 2\pi f)\, c_3\right) \\
&\quad + 4\left\|\mathbf{B}\right\|^2 \sin^2(\pi f) \\
&\quad \left[\sum_{p=0}^{T-2} b'^{p+1}\Big[\left(\frac{1}{1-a'} - 1\right)\cos(-2\pi f(p+1)) + (\cos(-2(p+1)\pi f)\cos(4\pi fT + 2\pi f)\right. \\
&\qquad \left. + \sin(2(p+1)\pi f)\sin(4\pi fT + 2\pi f)(c_1 - 1) + \sin(4\pi fT - 2(p+1)\pi f + 2\pi f)\, c_3\Big]\right] \\
&= \limsup_{T\to\infty} 2\left\|\mathbf{B}\right\|^2 \sin^2(\pi f)\left(\left(\frac{1}{1-a'} - 1\right) + \cos(4\pi fT + 2\pi f)(c_1 - 1) + \sin(4\pi fT + 2\pi f)\, c_3\right) \\
&\quad + 4\left\|\mathbf{B}\right\|^2 \sin^2(\pi f) \\
&\quad \left[\left(\frac{1}{1-a'} - 1\right)(c_2 - 1) + ((c_2 - 1)\cos(4\pi fT + 2\pi f) + c_4 \sin(4\pi fT + 2\pi f))(c_1 - 1)\right. \\
&\quad \left. + \sum_{p=0}^{t-2} b'^{p+1}\left[(-\sin(2(p+1)\pi f)\cos(4\pi fT + 2\pi f) + \cos(2(p+1)\pi f)\sin(4\pi fT + 2\pi f))\, c_3\right]\right] \\
&= \limsup_{t\to\infty} 2\left\|\mathbf{B}\right\|^2 \sin^2(\pi f)\left(\left(\frac{1}{1-a'} - 1\right) + \cos(4\pi fT + 2\pi f)(c_1 - 1) + \sin(4\pi fT + 2\pi f)\, c_3\right) \\
&\quad + 4\left\|\mathbf{B}\right\|^2 \sin^2(\pi f) \\
&\quad \left[\left(\frac{1}{1-a'} - 1\right)(c_2 - 1) + [(c_2 - 1)\cos(4\pi fT + 2\pi f) + c_4 \sin(4\pi fT + 2\pi f)](c_1 - 1)\right. \\
&\qquad \left. + [-c_4 \cos(4\pi fT + 2\pi f) + (c_2 - 1)\sin(4\pi fT + 2\pi f)]\, c_3\right].
\end{aligned}
$$

And term 3 in Equation (62) is,

$$\limsup_{T\to\infty} \sum_{k=2}^{T} b'^{T-k+1} \cos\left(\pi f\left(2k-1\right)\right) = \limsup_{T\to\infty} \sum_{j=1}^{T-1} b'^{j} \cos\left(\pi f\left(2\left(T-j+1\right)-1\right)\right)$$
$$= \limsup_{T\to\infty} \sum_{j=1}^{T-1} b'^{j} \cos\left(-2\pi f j + 2\pi f T + \pi f\right)$$
$$= \limsup_{T\to\infty} \sum_{j=1}^{T-1} b'^{j} \cos\left(-2\pi f j\right)\cos\left(2\pi f T + \pi f\right) - \sum_{j=1}^{T-1} b^{j} \sin\left(-2\pi f j\right)\sin\left(2\pi f T + \pi f\right)$$
$$= \limsup_{T\to\infty} \cos\left(2\pi f T + \pi f\right)\left(\frac{1-b'\cos\left(-2\pi f\right)}{1-2b'\cos\left(-2\pi f\right)+b'^{2}} - 1\right)$$
$$- \sin\left(2\pi f T + \pi f\right)\left(\frac{b'\sin\left(-2\pi f\right)}{1-2b\cos\left(-2\pi f\right)+b'^{2}}\right)$$

Finally, the expression to $\limsup_{T\to\infty} \left\|\mathbf{w}_T - \mathbf{w}_i^\star\right\|^2$ follows by adding the expressions, and noting that,

$$\limsup_{T\to\infty} a'^{T} \left\|\mathbf{w}_0 - \mathbf{w}_1^\star\right\|^2 = 0,$$
$$\limsup_{T\to\infty} v_z c' \frac{-a'^{T}}{1-a'} = 0,$$
$$\limsup_{T\to\infty} 2b'^{T} \left(\mathbf{w}_T^\star - \mathbf{w}_i^\star\right)^\top \left(\mathbf{w}_0 - \mathbf{w}_1^\star\right) = 0,$$

$$\left\|\mathbf{w}_T^\star - \mathbf{w}_i^\star\right\|^2 = 4\left\|\mathbf{B}\right\|^2 \cos^2(\pi f(T+i)) \sin^2(\pi f(T-i)).$$

Note that, $\frac{1}{|A_T|}\sum_{i\in A_T} \left\|\mathbf{w}_T - \mathbf{w}_i^\star\right\|^2 = \left\|\mathbf{w}_T - \mathbf{w}_i^\star\right\|^2$ since $\mathbf{w}_i^\star = \mathbf{w}_j^\star$ for all $i, j \in A_T$.

Therefore, evaluating at the last predictor, we get,

$$\limsup_{T\to\infty} \frac{1}{|A_T|} \sum_{i\in A_T} \left\|\mathbf{w}_T - \mathbf{w}_i^\star\right\|^2 = \limsup_{T\to\infty} C_0 + C_1 \cos\left(4\pi f T + 2\pi f\right) + C_2 \sin\left(4\pi f T + 2\pi f\right)$$
$$= \limsup_{T\to\infty} C_0 + \sqrt{C_1^2 + C_2^2} \cos\left(4\pi f T + 2\pi f - \arctan 2\left(C_2, C_1\right)\right),$$

where

$$C_0 = v_z c' \frac{1}{1-a'} + 2\left\|\mathbf{B}\right\|^2 \sin^2\left(\pi f\right)\left(\frac{1}{1-a'} - 1\right)\left(2c_2 - 1\right),$$
$$C_1 = 4\left\|\mathbf{B}\right\|^2 \sin^2\left(\pi f\right)\left(\frac{1}{2}\left(c_1 - 1\right) + \left(c_2 - 1\right)\left(c_1 - 1\right) - c_4 c_3\right),$$
$$C_2 = 4\left\|\mathbf{B}\right\|^2 \sin^2\left(\pi f\right)\left(-\frac{1}{2}c_3 + c_4\left(c_1 - 1\right) + c_2 c_3\right).$$

□