Learning with Synthetic Data via SGD in High-Dimensional Linear Regression

Jichu Li
School of Statistics
Renmin University of China
lijichu52@gmail.com

Difan Zou
School of Computing and Data Science & Institute of Data Science
The University of Hong Kong
dzou@hku.hk

Abstract

Synthetic data has become a promising way to scale model training beyond limited human-generated data but it may also induce *strong model collapse* [19], where any fixed fraction of synthetic data prevents model performance from improving under data scaling. In this paper, we study how synthetic data affects the generalization of one-pass SGD in high-dimensional linear regression with model shift. We establish finite-sample risk bounds for mixed and two-stage training that disentangle standard SGD error from mismatch-induced effects and quantify how these effects evolve along the optimization trajectory. This highlights a sharp contrast: mixed training induces strong model collapse, while two-stage training avoids the floor by using synthetic data only in the first stage, suggesting that collapse is not inevitable under a simple data curriculum. Under a random sketch model, we further obtain scaling laws for both protocols, showing that larger models may amplify synthetic-induced degradation under mixing, while high-quality synthetic pretraining may reduce bias under two-stage training. Finally, we establish an exact finite-sample condition for two-stage training to strictly outperform real-only training under the same real-data budget and identical real-stage updates. Overall, our results highlight that synthetic data is neither inherently harmful nor beneficial; its effect depends critically on both its quality and the training protocol used to incorporate it.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities, primarily driven by large-scale, high-quality pre-training data [57, 11]. Scaling laws suggest that increasing both model size and data volume consistently leads to improved performance [34, 30, 52]. However, high-quality pre-training data is limited, and its supply is rapidly diminishing, posing a significant challenge for continued model scaling [52, 63]. Consequently, synthetic data, often generated by existing models, has emerged as a promising supplement to human-generated data for training, particularly in scenarios where labeled data is scarce [52, 12, 49, 29]. The training paradigm thus evolves to operate on a mixture of real data from $\mathcal{P}_1$ and synthetic data from $\mathcal{P}_2$, while our ultimate interest remains performance on the real data distribution $\mathcal{P}_1$ alone.

This raises a natural question: Does synthetic data truly benefit model training, and what limitations or adverse effects might arise? A growing body of work has demonstrated the benefits of synthetic data for expanding training sets and improving generalization [25, 43, 47, 50, 49, 1, 56]. At the

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

same time, training on synthetic data can induce *model collapse*, where model performance degrades as generated data is repeatedly incorporated into training [59, 28, 9, 2, 26, 10, 60]. Understanding when synthetic data is beneficial or harmful is therefore increasingly important for large-scale model development.

Theoretical studies typically model learning with synthetic data as regression under distribution mismatch between real and synthetic sources. Existing work has characterized the phenomenon of model collapse [20, 18, 58, 31] and explored ways to mitigate it or leverage synthetic data effectively to improve learning [58, 22, 5, 23, 21]. Of particular relevance to our work, Dohmatob et al. [19] introduce the concept of *strong model collapse*: even a small proportion of synthetic data prevents model performance from improving under data scaling, leaving a non-vanishing excess risk floor, which serves as a key motivation for our analysis. However, most of these works focus on static estimators and asymptotic regimes, without capturing the dynamics of practical training procedures based on iterative optimization algorithms (e.g., SGD), nor providing a detailed finite-sample characterization.

In this paper, we study the effect of synthetic data on model generalization through the lens of one-pass SGD in high-dimensional linear regression. We consider a two-source setting where real and synthetic data share the same covariate distribution but differ in their labeling functions, capturing common synthetic-supervision pipelines where real inputs are labeled by a biased teacher, and analyze SGD under both mixed and two-stage training protocols. We establish finite-sample upper and lower bounds showing that mixed training leads to *strong model collapse* under SGD, while a simple two-stage protocol avoids this issue and achieves vanishing excess risk. Further, we derive scaling laws under a random sketch model, characterizing how model size may affect synthetic-induced degradation and how synthetic data may benefit learning through bias reduction. Finally We give an exact necessary-and-sufficient finite-sample condition for two-stage training to strictly outperform real-only training with the same real-data budget and real-stage updates.
Our contributions can be summarized as follows:

- In Section 3.1, we derive finite-sample upper and lower bounds for last-iterate one-pass SGD under mixed training. The upper bound decomposes the synthetic-data-induced degradation into fluctuation and drift, characterizing how source mismatch affects optimization dynamics. Together, the upper and lower bounds establish a non-vanishing excess risk floor under data scaling, which is exactly *strong model collapse*. Moreover, we show that the same behavior persists for iterate-averaged SGD, indicating that strong model collapse is fundamentally induced by sample mixing in the training protocol, rather than an artifact of a specific optimization scheme.

- In Section 3.2, we show that strong model collapse is not inevitable. We analyze a two-stage training protocol, where the model is first trained on synthetic data and then on real data, and show that this simple data curriculum avoids the non-vanishing excess risk floor. This highlights that the ordering and scheduling of data can be as important as the data itself.

- In Section 3.3, we further establish scaling-law bounds for the two training protocols under a random sketch model, with tight results for mixed training in the optimization-saturated regime, showing that larger models may amplify synthetic-data-induced degradation in mixed training, and providing an explicit characterization of how high-quality synthetic data may reduce bias in two-stage training, thereby benefiting learning.

- In Section 3.4, we establish an exact finite-sample condition under which two-stage training strictly outperforms real-data-only training with a fixed real-data budget and identical real-stage training; we also use a one-dimensional toy model to illustrate the roles of synthetic sample size, source mismatch, and label noise.

Notations. For two positive-valued functions $f(x)$ and $g(x)$, we write $f(x) \lesssim g(x)$ (and $f(x) = \mathcal{O}(g(x))$) or $f(x) \gtrsim g(x)$ (and $f(x) = \Omega(g(x))$) if $f(x) \leq cg(x)$ or $f(x) \geq cg(x)$ holds for some absolute (if not otherwise specified) constant $c > 0$ respectively. We write $f(x) \eqsim g(x)$ (and $f(x) = \Theta(g(x))$) if $f(x) \lesssim g(x) \lesssim f(x)$. For two vectors $\mathbf{u}$ and $\mathbf{v}$ in a Hilbert space, we denote their inner product by $\langle \mathbf{u}, \mathbf{v} \rangle$ or $\mathbf{u}^\top \mathbf{v}$. For two matrices $\mathbf{A}$ and $\mathbf{B}$ of appropriate dimensions, we define their inner product by$\langle \mathbf{A}, \mathbf{B} \rangle := \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$. We use $\|\cdot\|$ to denote the operator norm for matrices and ℓ_2-norm for vectors. For a positive semi-definite (PSD) matrix $\mathbf{A}$ and a vector $\mathbf{v}$ of appropriate dimension, we write $\|\mathbf{v}\|_{\mathbf{A}}^2 := \mathbf{v}^\top \mathbf{A} \mathbf{v}$. Kronecker/tensor product is denoted by $\otimes$.

2 Preliminaries

Problem Setup. We use $\mathbf{x} \in \mathcal{H}$ to denote a feature vector, where $\mathcal{H}$ is a finite d-dimensional or countably infinite dimensional Hilbert space, and $y \in \mathbb{R}$ to denote its label. *Linear Regression* concerns the following objective:

$$\min_{\mathbf{w}} \mathcal{R}(\mathbf{w}), \text{ where } \mathcal{R}(\mathbf{w}) := \frac{1}{2}\mathbb{E}\big(\langle \mathbf{x}, \mathbf{w}\rangle - y\big)^2$$

Here $\mathbf{w} \in \mathcal{H}$ is a weight vector to be learned and the expectation is over $(\mathbf{x}, y) \sim \mathcal{P}$ for some distribution $\mathcal{P}$ on $\mathcal{H} \times \mathbb{R}$.

Following the prior literature [4, 67, 73, 45], we make the following assumptions on the data distribution.

Assumption 1 (Data Covariance) *Let $\mathbf{H} := \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ be the data covariance matrix and assume that $\mathrm{tr}(\mathbf{H})$ and all entries of $\mathbf{H}$ are finite. In addition, we assume that all parameter vectors involved in the analysis have finite $\mathbf{H}$-norm, i.e., $\|\mathbf{w}\|_{\mathbf{H}} < \infty$.*

Assumption 2 (Fourth moment conditions) *(1) There exists constant $\alpha > 0$ such that for every PSD matrix $\mathbf{A}$ it holds that $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top] \preceq \alpha\,\mathrm{tr}(\mathbf{H}\mathbf{A})\mathbf{H}$; (2) There is a constant $\beta > 0$, such that for every PSD matrix $\mathbf{A}$, we have $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top] - \mathbf{H}\mathbf{A}\mathbf{H} \succeq \beta\,\mathrm{tr}(\mathbf{H}\mathbf{A})\mathbf{H}$.*

In this paper, motivated by [19], we consider a two-source setting, where data are drawn from either a real distribution $\mathcal{P}_1$ or a synthetic distribution $\mathcal{P}_2$. We assume that both distributions share the same feature marginal $\mathbf{x} \sim \mathcal{D}$, and differ only in their labeling functions, which allows us to isolate the effect of model mismatch and obtain a clean characterization of the learning dynamics. This formulation naturally captures the prevalent self-training and knowledge distillation paradigms, where a massive pool of unlabeled real-world features is annotated by a potentially biased teacher model to generate supervisory signals.

Assumption 3 (Well-specified models) *For $\ell \in \{1, 2\}$, data from distribution $\mathcal{P}_\ell$ follow a linear model $y = \langle \mathbf{x}, \mathbf{w}_\ell^*\rangle + \xi_\ell$, where $\mathbf{w}_\ell^* \in \mathcal{H}$ is the underlying parameter, and the noise $\xi_\ell \sim N(0, \sigma_\ell^2)$ is independent of $\mathbf{x}$. Define the model shift between the two sources as $\boldsymbol{\delta} := \mathbf{w}_2^* - \mathbf{w}_1^*$.*

While training may involve samples from both $\mathcal{P}_1$ and $\mathcal{P}_2$, our goal is to minimize the population risk on the real distribution $\mathcal{R}(\mathbf{w}) := \frac{1}{2}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_1}(\langle \mathbf{x}, \mathbf{w}\rangle - y)^2$.

Training protocol. We consider two training protocols for combining real and synthetic data:
(1) Mixed training. We are given a dataset consisting of $M + N$ samples, among which M are drawn from the synthetic distribution $\mathcal{P}_2$ and N from the real distribution $\mathcal{P}_1$. Let $p := M/(M + N)$ be the fraction of synthetic data. The training samples are presented in a random order. An equivalent fixed-budget formulation key to the proof is detailed in Appendix C.
(2) Two-stage training. We first train on M synthetic samples drawn from $\mathcal{P}_2$, followed by training on N real samples drawn from $\mathcal{P}_1$.

SGD iteration. We consider training via *stochastic gradient descent* (SGD) with step sizes $\{\gamma_t\}$:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma_t\big(\langle \mathbf{x}_t, \mathbf{w}_t\rangle - y_t\big)\mathbf{x}_t \quad t \geq 1.$$

where $\{(\mathbf{x}_t, y_t)\}_{t=1}^{M+N}$ follow one of the two protocols above, and $\{\gamma_t\}_{t=1}^{M+N}$ are the step sizes. We adopt a well-used geometric decaying stepsize scheduler [67, 66, 45, 46]:

$$\text{for } t = 1, \ldots, T, \quad \gamma_t := \gamma/2^{\ell_t}, \text{ and } \ell_t = \lfloor t/(T/\log(T))\rfloor. \tag{1}$$

Here the initial stepsize γ is a hyperparameter for the algorithm and T denotes the total number of iterations. In the mixed-training setting, we take $T = M + N$, while in the two-stage setting the same schedule is applied separately within each stage.

The output of the algorithm is the last iterate $\mathbf{w}_{M+N}$ and we study the *excess risk* on the real distribution, $\mathcal{E}_1(\mathbf{w}_{M+N}) = \mathcal{R}(\mathbf{w}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*)$.

Additional Notations. let $\mathbf{H} = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ be its eigen-decomposition with non-increasing sorted eigenvalues $\{\lambda_i\}_{i=1}^\infty$ and eigenvectors $\mathbf{v}_i$. Following [73, 66], define $\mathbf{H}_{0:k} := \sum_{i=1}^k \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, $\mathbf{H}_{k:\infty} := \sum_{i>k} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, and $\boldsymbol{\delta} = \sum_i \delta_i \mathbf{v}_i$.

3 Main Results

3.1 Mixed training

In this subsection, we study SGD trained on a mixture of real and synthetic data. We show that, in this setting, the presence of synthetic data induces a non-vanishing excess risk floor, leading to the phenomenon of strong model collapse [19]. Detailed proof is deferred to Appendix C and D.

Theorem 1 (An upper bound for mixed training) *Given M synthetic and N real samples. Consider last iterate SGD with stepsize scheme* (1)*. Suppose Assumptions 1, 2 and 3 hold. Let $\widetilde{N}_{\mathtt{eff}} := (M+N)/\log(M+N)$ and $\bar{\sigma}^2 = (1-p)\sigma_1^2 + p\sigma_2^2$. Suppose $\gamma < 1/(4\alpha\,\mathrm{tr}(\mathbf{H}))$. We have*

$$\mathbb{E}[\mathcal{E}_1(\mathbf{w}_{M+N})] \lesssim \underbrace{\Big\| \prod_{t=1}^{M+N} (\mathbf{I} - \gamma_t \mathbf{H})(\mathbf{w}_0 - \mathbf{w}_1^*)\Big\|_{\mathbf{H}}^2 + \alpha \|\mathbf{w}_0 - \mathbf{w}_1^*\|^2_{\frac{\mathbf{I}_{0:k^*}}{\gamma \widetilde{N}_{\mathtt{eff}}} + \mathbf{H}_{k^*:\infty}} \frac{\widetilde{D}_{\mathtt{eff}}}{\widetilde{N}_{\mathtt{eff}}}}_{\mathsf{BiasError}}$$
$$+ \underbrace{\bar{\sigma}^2 \frac{\widetilde{D}_{\mathtt{eff}}}{\widetilde{N}_{\mathtt{eff}}}}_{\mathsf{VarError}} + \underbrace{\alpha\, p \|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \frac{\widetilde{D}_{\mathtt{eff}}}{\widetilde{N}_{\mathtt{eff}}}}_{\mathsf{FlucError}} + \underbrace{p^2 \Big\| \big(\mathbf{I} - \prod_{t=1}^{M+N} (\mathbf{I} - \gamma_t \mathbf{H})\big)\boldsymbol{\delta}\Big\|_{\mathbf{H}}^2}_{\mathsf{DriftError}},$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{\gamma \widetilde{N}_{\mathtt{eff}}}\}$ and $\widetilde{D}_{\mathtt{eff}} := k^* + \gamma^2 \widetilde{N}_{\mathtt{eff}}^2 \sum_{i>k^*} \lambda_i^2$.

Interpretation. This upper bound decomposes the excess risk into four terms: a bias term, a variance term, and two additional terms induced by the presence of synthetic data, which we refer to as the fluctuation error and the drift error. The bias and variance terms match those in classical analyses of SGD with geometrically decaying step sizes [66], capturing the optimization error and stochastic noise, respectively. The fluctuation error is upper bounded by terms involving $p(1-p)$, p^2 and p, and is therefore summarized at the order level as $O(p\|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \widetilde{D}_{\mathtt{eff}}/\widetilde{N}_{\mathtt{eff}})$. It arises from the stochastic mixing of the two data sources. Particularly, when $\|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \eqsim \bar{\sigma}^2$, this term is of the same order as the standard variance term and behaves similarly to it. In contrast, the drift error $p^2\|(\mathbf{I} - \prod_{t=1}^{M+N}(\mathbf{I} - \gamma_t \mathbf{H}))\boldsymbol{\delta}\|_{\mathbf{H}}^2$ captures a systematic bias in the optimization trajectory toward the synthetic model, which reflects the accumulation of updates aligned with $\mathbf{w}_2^*$, leading the iterate to drift away from the target $\mathbf{w}_1^*$. Together, these two terms reveal a fundamental trade-off: while incorporating synthetic data can increase sample size to reduce bias and variance, it also introduces a bias proportional to the model mismatch $\boldsymbol{\delta}$, which may dominate and degrade performance.

Theorem 2 (A lower bound for mixed training) *Given M synthetic and N real samples. Consider last iterate SGD with stepsize scheme* (1)*. Suppose Assumptions 1, 2 and 3 hold. Let $\widetilde{N}_{\mathtt{eff}} := (M+N)/\log(M+N)$ and $\bar{\sigma}^2 = (1-p)\sigma_1^2 + p\sigma_2^2$. Suppose $M + N \geq 500$ and $\gamma < 1/\lambda_1$, then*

$$\mathbb{E}[\mathcal{E}_1(\mathbf{w}_{M+N})] \gtrsim \underbrace{\Big\| \prod_{t=1}^{M+N} (\mathbf{I} - \gamma_t \mathbf{H})(\mathbf{w}_0 - \mathbf{w}_1^*) + p\big(\mathbf{I} - \prod_{t=1}^{M+N} (\mathbf{I} - \gamma_t \mathbf{H})\big)\boldsymbol{\delta}\Big\|_{\mathbf{H}}^2}_{\mathsf{Non-vanishing\ term}}$$
$$+ \big(\beta \|\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}\|^2_{\mathbf{H}_{k^*:\infty}} + \bar{\sigma}^2\big) \frac{\widetilde{D}_{\mathtt{eff}}}{\widetilde{N}_{\mathtt{eff}}} + p(1-p)\gamma^2 \widetilde{N}_{\mathtt{eff}} \|\boldsymbol{\delta}\|^2_{\mathbf{H}^3_{k^*:\infty}}$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{\gamma \widetilde{N}_{\mathtt{eff}}}\}$ and $\widetilde{D}_{\mathtt{eff}} := k^* + \gamma^2 \widetilde{N}_{\mathtt{eff}}^2 \sum_{i>k^*} \lambda_i^2$.

Interpretation. The lower bound complements Theorem 1 by showing that the excess risk cannot vanish in the large-sample regime. In particular, its first term captures the mean optimization error, jointly determined by the residual initialization error and the drift toward the synthetic model. At finite sample sizes, the spectral filter determines which directions remain dominated by the initialization and which have accumulated synthetic-induced bias. The remaining terms further quantify label-noise variance and mismatch-induced fluctuations, providing a more complete finite-sample characterization. Comparing with the upper bound, we observe that both bounds contain

a term governed by the quantity $p\big(\mathbf{I} - \prod_{t=1}^{M+N}(\mathbf{I} - \gamma_t \mathbf{H})\big)\boldsymbol{\delta}$, which characterizes the drift of the optimization trajectory toward the synthetic model. While the upper bound isolates this effect as the drift error, the lower bound shows that such a term is unavoidable as the initialization is progressively forgotten. Crucially, when the total sample size scales to infinity with a fixed synthetic proportion p, and under the condition $\widetilde{D}_{\texttt{eff}} = o(\widetilde{N}_{\texttt{eff}})$, all remaining terms vanish, and both bounds converge to the same non-zero limit of order $p^2\|\boldsymbol{\delta}\|_{\mathbf{H}}^2$. This directly leads to the following corollary, which formalizes the phenomenon of *strong model collapse*[19].

Corollary 3 (Strong model collapse) *Consider last iterate SGD with stepsize scheme* (1)*. Suppose Assumptions 1, 2 and 3 hold. Suppose $\gamma < 1/4\alpha\,\mathrm{tr}(\mathbf{H})$ and $\widetilde{D}_{\texttt{eff}} = o(\widetilde{N}_{\texttt{eff}})$. As the total sample size $M + N$ scales to infinity with a fixed synthetic proportion $p = M/(M+N)$, we have*

$$\lim_{M+N\to\infty} \mathbb{E}[\mathcal{E}_1(\mathbf{w}_{M+N})] \eqsim p^2\|\boldsymbol{\delta}\|_{\mathbf{H}}^2.$$

Corollary 3 shows that the strong model collapse phenomenon persists under SGD with decaying stepsizes. In particular, even when the total sample size $M + N$ tends to infinity, as long as the proportion of synthetic data p remains non-vanishing, the excess risk does not converge to zero, but instead admits a strictly positive limit of order $p^2\|\boldsymbol{\delta}\|_{\mathbf{H}}^2$. This demonstrates that strong model collapse is not merely a property of batch estimators or scaling-law limits, but also arises intrinsically from the optimization dynamics of SGD. Here the condition $\widetilde{D}_{\texttt{eff}} = o(\widetilde{N}_{\texttt{eff}})$ requires the effective spectral dimension to grow sublinearly with the effective sample size, and holds for many standard eigendecay regimes; see [73, 67].

Our result should be compared with Dohmatob et al. [19], who established strong model collapse for regression models trained on mixtures of real and synthetic data. In their classical linear estimator setting, the error floor is of order $p^2\mathbb{E}[\|\mathbf{w}_2^* - \mathbf{w}_1^*\|_{\mathbf{H}}^2]$. Hence, their expression can be viewed as taking an additional expectation over the model mismatch, while ours characterizes the floor for a fixed $\boldsymbol{\delta}$. In this sense, our result is consistent with the estimator-level picture of Dohmatob et al. [19], but recast in stochastic optimization dynamics. However, a key distinction is that their analysis focuses on classical batch estimators (ridge/OLS) under a proportional high-dimensional asymptotic regime, and the resulting error floor is an estimator-level asymptotic characterization. In contrast, we analyze SGD under finite-sample training and establish non-asymptotic upper and lower bounds.

Moreover, we extend our analysis to constant-stepsize SGD with iterate averaging, establishing corresponding upper and lower risk bounds that again imply a non-vanishing error floor of order $p^2\|\boldsymbol{\delta}\|_{\mathbf{H}}^2$ under mixed training; see Appendix H. This naturally raises an important question:

***Q:** Is strong model collapse inevitable with fixed synthetic data proportion?*

In the next subsection, we answer this question by showing that it can in fact be avoided. Specifically, we demonstrate that a two-stage training protocol eliminates the non-vanishing error floor, leading to vanishing excess risk, highlighting the importance of how synthetic and real data are scheduled.

3.2 Two-stage Training

In this subsection, we study a two-stage training protocol, where the model is first trained on synthetic data and then further trained on real data. We show that, in contrast to the mixed-training setting, this training scheme avoids strong model collapse and leads to vanishing excess risk under mild conditions. The detailed proof is deferred to Appendix E.

Theorem 4 (Two-stage training) *Given M synthetic samples and N real samples. Consider last iterate SGD with stepsize scheme* (1)*. Suppose Assumptions 1, 2 and 3 hold. Let $N_{\texttt{eff}} := N/\log N$ and $\mathcal{E}_2(\mathbf{w}) := \frac{1}{2}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_2}[(\langle\mathbf{x},\mathbf{w}\rangle - y)^2] - \frac{1}{2}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_2}[(\langle\mathbf{x},\mathbf{w}_2^*\rangle - y)^2]$ denote the excess risk on the synthetic distribution. Suppose $\gamma < 1/(4\alpha\,\mathrm{tr}(\mathbf{H}))$. Then we have*

$$\mathbb{E}[\mathcal{E}_1(\mathbf{w}_{M+N})] \lesssim \underbrace{\mathbb{E}[\mathcal{E}_2(\mathbf{w}_M)] + \Big\|\prod_{t=1}^{N}(\mathbf{I} - \gamma_t\mathbf{H})\boldsymbol{\delta}\Big\|_{\mathbf{H}}^2}_{\mathsf{EffectiveBias}} + \underbrace{\Big(\alpha\,\mathbb{E}[\mathcal{E}_2(\mathbf{w}_M)] + \alpha\|\boldsymbol{\delta}\|_{\mathbf{H}}^2 + \sigma_1^2\Big)\frac{D_{\texttt{eff}}}{N_{\texttt{eff}}}}_{\mathsf{EffectiveVariance}}.$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{\gamma N_{\texttt{eff}}}\}$ and $D_{\texttt{eff}} := k^* + \gamma^2 N_{\texttt{eff}}^2 \sum_{i>k^*}\lambda_i^2$.

As a direct consequence of Theorem 4, by considering the asymptotic regime where the sample size goes to infinity, we can establish that strong model collapse does not occur:

Corollary 5 (No strong model collapse) *Under the assumptions of Theorem 4. As the total sample size $M+N$ scales to infinity with a fixed synthetic proportion $p = M/(M+N)$, suppose that $D_{\texttt{eff}} = o(N_{\texttt{eff}})$, then $\mathbb{E}[\mathcal{E}_2(\mathbf{w}_M)] \xrightarrow{M\to\infty} 0$, and we have*

$$\lim_{M+N\to\infty} \mathbb{E}[\mathcal{E}_1(\mathbf{w}_{M+N})] = 0.$$

Corollary 5 shows that, in sharp contrast to the mixed-training setting, strong model collapse can be avoided under a two-stage training protocol. The key distinction is that synthetic data no longer appears throughout the entire optimization trajectory. Instead, the synthetic stage only determines the initialization of the second stage, after which all subsequent updates are performed on real data. As a result, the synthetic model does not induce a persistent drift in the trajectory. Broadly, this suggests that successfully leveraging synthetic data may hinge crucially on the training curriculum, as the specific ordering and scheduling of data may be just as important as the data itself.

Further, Theorem 4 sheds light on the mechanism through which synthetic data can be beneficial. The main benefit does *not* come from reducing variance: the $\mathsf{EffectiveVariance}$ term still scales as $D_{\texttt{eff}}/N_{\texttt{eff}}$, which is the same order as in the standard single-source SGD bound trained on real data only [66]. Instead, the main benefit of synthetic data lies in reducing the *bias*. First, the synthetic training phase error $\mathbb{E}[\mathcal{E}_2(\mathbf{w}_M)]$ can be driven to zero as M increases. Second, the residual mismatch term $\| \prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{H})\boldsymbol{\delta}\|_{\mathbf{H}}^2$ replaces the real-only initialization term $\| \prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{H})(\mathbf{w}_0-\mathbf{w}_1^*)\|_{\mathbf{H}}^2$. Thus, two-stage training effectively replaces random initialization error by the source mismatch $\boldsymbol{\delta} = \mathbf{w}_2^* - \mathbf{w}_1^*$. When the synthetic data is of high quality so that $\|\boldsymbol{\delta}\|_{\mathbf{H}}^2$ is small, first-stage places the model substantially closer to the target $\mathbf{w}_1^*$, yielding a smaller bias after second-stage training.

Consequently, synthetic data primarily acts as a *bias-reduction mechanism*: it improves the initialization for the real-data optimization stage, thereby accelerating convergence and reducing the optimization error when real data is limited. This highlights that the effectiveness of synthetic data critically depends on its quality (i.e., the magnitude of $\boldsymbol{\delta}$), as well as on the training protocol used to incorporate it. In the next section, we make this intuition more precise by analyzing a random sketch model under a power-law eigendecay assumption on the data covariance, which allows us to characterize more explicitly when and how synthetic data improves performance.

Learning-rate restart. Throughout the main text, including the scaling-law analysis below, we restart the learning-rate schedule to treat the real-data phase as a separate fine-tuning stage, with a larger cumulative stepsize that accelerates forgetting of the synthetic initialization. Without restart, the mismatch filter becomes $\prod_{t=M+1}^{M+N}(\mathbf{I}-\gamma_t\mathbf{H})$, which generally forgets $\boldsymbol{\delta}$ more slowly. Nevertheless, restart does not affect the qualitative behavior: for fixed $p \in (0,1)$, the cumulative real-stage stepsize diverges and the excess risk still has no non-vanishing floor; see detailed discussion in Appendix E.1.

3.3 Random Sketch Model & Scaling laws

The results in the previous subsections establish a sharp qualitative distinction between mixed training and two-stage training. However, these results do not yet reveal how the effect of synthetic data interacts with model size. To study this question and establish scaling laws for the two training protocols, we adopt the random sketch model with Gaussian sketch matrix [45, 46, 70]. Detailed proof of this subsection is deferred to Appendix F.

Concretely, following Lin et al. [45], we introduce a Gaussian random sketch operator $\mathbf{S} : \mathcal{H} \to \mathbb{R}^D$, where entries of $\mathbf{S}$ are independently sampled from $N(0, 1/D)$, and replace the original feature vector $\mathbf{x}$ in our problem setup by its D-dimensional sketch $\mathbf{S}\mathbf{x}$. We then train a linear predictor with D trainable parameters in the sketched space,

$$f_{\mathbf{v}}(\mathbf{x}) = \langle \mathbf{v}, \mathbf{S}\mathbf{x} \rangle, \qquad \mathbf{v} \in \mathbb{R}^D,$$

starting from zero initialization $\mathbf{v}_0 = \mathbf{0}$ and using the same two-source data model and training protocols introduced in Section 2. Accordingly, the population risk on the real distribution becomes

$$\mathcal{R}_D(\mathbf{v}) := \frac{1}{2}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_1}\big(\langle \mathbf{v}, \mathbf{S}\mathbf{x} \rangle - y\big)^2.$$

where the expectation is conditioned on $\mathbf{S}$. The corresponding optimal parameters are given by $\mathbf{v}_\ell^* := (\mathbf{SHS}^\top)^{-1}\mathbf{SHw}_\ell^*$, $\ell \in \{1, 2\}$. Thus, the sketch dimension D plays the role of model size, allowing us to study how model scaling interacts with source mismatch and synthetic data. As in Lin et al. [45], we can decompose the risk into irreducible, approximation, and excess terms:

$$\mathcal{R}_D(\mathbf{v}_{M+N}) = \underbrace{\min \mathcal{R}(\cdot)}_{\text{Irreducible}} + \underbrace{\min \mathcal{R}_D(\cdot) - \min \mathcal{R}(\cdot)}_{\text{Approx}} + \underbrace{\mathcal{R}_D(\mathbf{v}_{M+N}) - \min \mathcal{R}_D(\cdot)}_{\text{Excess}}.$$

Following the prior literature [45, 46, 39, 70], we further make the following additional assumptions to establish scaling law upper bounds for the two training protocols and to explicitly characterize the effects of synthetic data and model capacity.

Assumption 4 (Distributional conditions) *We assume $\mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$, and the random vectors $\mathbf{w}_1^*$ and $\boldsymbol{\delta} := \mathbf{w}_2^* - \mathbf{w}_1^*$ satisfy $\mathbb{E}[\mathbf{w}_1^*(\mathbf{w}_1^*)^\top] = \mathbf{I}$, $\mathbb{E}[\boldsymbol{\delta}\boldsymbol{\delta}^\top] = \boldsymbol{\Sigma}$, and $\mathbb{E}[\mathbf{w}_1^*\boldsymbol{\delta}^\top] = \mathbf{0}$.*

Assumption 5 (Power-law spectrum and source condition) *Let $(\lambda_i, \mathbf{v}_i)_{i\geq 1}$ be the eigenvalue-eigenvector pairs of $\mathbf{H}$ with $(\lambda_i)_{i\geq 1}$ in non-increasing order. We assume $\lambda_i \eqsim i^{-a}$ for some $a > 1$, and for the mismatch $\boldsymbol{\delta}$ we assume $\mathbb{E}[\langle \mathbf{v}_i, \boldsymbol{\delta}\rangle\langle \mathbf{v}_j, \boldsymbol{\delta}\rangle] = 0$ for $i \neq j$ and $\mathbb{E}[\lambda_i\langle \mathbf{v}_i, \boldsymbol{\delta}\rangle^2] \eqsim i^{-b}$ for some $b > 1$.*

The exponent b in Assumption 5 characterizes how hard the mismatch $\boldsymbol{\delta}$ is relative to the data spectrum $\mathbf{H}$. A larger b implies a simpler task, as the mismatch energy decays more rapidly along $\mathbf{H}$'s eigen-directions.

Theorem 6 (Scaling law for mixed training) *Given M synthetic and N real samples. Consider last iterate SGD with stepsize scheme* (1)*. Suppose that Assumptions 1, 2, 3, 4 and 5 hold. Let $\widetilde{N}_{\mathtt{eff}} = (M + N)/\log(M + N)$. Suppose $\gamma < 1/4\alpha\,\mathrm{tr}(\mathbf{SHS}^\top)$, $\sigma_1^2, \sigma_2^2 \eqsim 1$ and choose $\gamma \eqsim 1$, then with probability at least $1 - e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$, we have*

$$\mathbb{E}\mathcal{R}_D(\mathbf{v}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*) \lesssim \underbrace{\frac{1}{D^{a-1}} + \frac{1}{(\widetilde{N}_{\mathtt{eff}})^{1-1/a}}}_{\text{Approx+Bias}} + \underbrace{\Big(\frac{M}{M+N}\Big)^2\Big(1 - \frac{1}{D^{b-1}}\Big)}_{\text{Drift Floor}}$$

Moreover, if the sketched problem is optimization-saturated that $\gamma\widetilde{N}_{\mathtt{eff}} \gtrsim D^a$, then the preceding upper bound is sharp up to constants:

$$\mathbb{E}\mathcal{R}_D(\mathbf{v}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*) \eqsim D^{1-a} + \widetilde{N}_{\mathtt{eff}}^{\,1/a-1} + p^2(1 - D^{1-b}).$$

Interpretation. The first part of Theorem 6 provides a clean upper-bound scaling picture for mixed training under the random sketch model. When $\gamma\widetilde{N}_{\mathtt{eff}} \gtrsim D^a$, the cumulative stepsize is large enough to learn even the smallest sketched eigendirection, whose eigenvalue is of order D^{-a}, and a matching lower bound holds in this optimization-saturated regime. The fluctuation and variance terms are absorbed into higher-order contributions and can be bounded by the same order as the Approx + Bias term. Hence, they do not affect the leading-order behavior, which is dominated by Approx + Bias and the drift floor. Second, synthetic data improves the reducible part of the risk at the pre-asymptotic level. Compared with the real-only scaling $1/N_{\mathtt{eff}}^{1-1/a}$ in Lin et al. [45], the term $1/\widetilde{N}_{\mathtt{eff}}^{1-1/a}$ shows that synthetic and real samples contribute jointly, effectively increasing the sample size and reducing Approx + Bias. However, this benefit comes with a fundamental trade-off. While increasing the sketch dimension D decreases the approximation term D^{1-a} and therefore improves Approx + Bias, it simultaneously increases the drift-floor term $p^2(1 - \frac{1}{D^{b-1}})$. Indeed, as D grows, the factor D^{1-b} vanishes and the drift contribution approaches its limiting level of order p^2. Therefore, larger models fit the real signal better, but they may amplify synthetic-induced degradation by exposing the mismatch between the real and synthetic sources more fully, revealing a fundamental trade-off in mixed training.

Theorem 7 (Scaling law upper bound for two-stage training) *Given M synthetic and N real samples. Consider last iterate SGD with stepsize scheme* (1)*. Suppose that Assumptions 1, 2, 3, 4 and 5 hold. Let $N_{\mathtt{eff}} = N/\log N$ and $M_{\mathtt{eff}} = M/\log M$. Suppose $\gamma < 1/4\alpha\,\mathrm{tr}(\mathbf{SHS}^\top)$ and*

$\sigma_1^2, \sigma_2^2 \eqsim 1$. *Choose* $\gamma \eqsim 1$*, and assume* $b < a + 1$ *for simplicity and clarity, then with probability at least* $1 - e^{-\Omega(D)}$ *over the randomness of the sketch matrix* $\mathbf{S}$*, we have*

$$\mathbb{E}\mathcal{R}_D(\mathbf{v}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*) \lesssim \underbrace{\frac{1}{D^{a-1}}}_{\text{Approx}} + \underbrace{\frac{\min\{D, (N_{\mathtt{eff}})^{1/a}\}}{N_{\mathtt{eff}}}}_{\text{Variance}} + \underbrace{e^{-\Omega(\frac{N_{\mathtt{eff}}}{D^a})} \cdot \max\Big\{\frac{1}{D^{a\wedge b-1}}, \frac{1}{(M_{\mathtt{eff}})^{\frac{a\wedge b-1}{a}}}\Big\} + \max\Big\{\frac{1}{D^{b-1}}, \frac{1}{(N_{\mathtt{eff}})^{\frac{b-1}{a}}}\Big\}}_{\text{Bias}}$$

Interpretation. The variance term is of the same order as in the real-only scaling law of Lin et al. [45], namely $\min\{D, N_{\mathtt{eff}}^{1/a}\}/N_{\mathtt{eff}}$, and therefore synthetic data does not improve the variance, which is consistent with our earlier analysis. The bias term is split into two parts, reflecting two distinct effects of synthetic training phase. The first term, $e^{-\Omega(N_{\mathtt{eff}}/D^a)} \max\{D^{1-a\wedge b}, M_{\mathtt{eff}}^{-(a\wedge b-1)/a}\}$, captures the residual error from the first stage, and is exponentially forgotten during training on real data. The second term, $\max\{D^{1-b}, N_{\mathtt{eff}}^{-(b-1)/a}\}$, captures the remaining source mismatch after the second stage real-data training. Compared with the real-only bias upper bound $\max\{D^{1-a}, N_{\mathtt{eff}}^{-(a-1)/a}\}$[45], this term is strictly smaller when $b > a$, corresponding to high-quality synthetic data. In this regime, the mismatch is easier than the original task, so synthetic training phase places the model closer to the target and reduces the bias more effectively than random initialization. So both bias contributions can be improved at the upper-bound level than in the real-only case. Therefore, unlike mixed training, two-stage training can genuinely benefit from synthetic data: high-quality synthetic data can reduce the bias without introducing any non-vanishing drift floor.

This subsection establishes only an upper bound for two-stage training, since a uniform termwise matching lower bound is generally impossible, because upper-bound analysis separately bounds the propagated first-stage error and the source-mismatch contribution, although their cross term may be negative and substantially cancel in the actual final iterate. Accordingly, Theorem 7 should be interpreted as an upper-bound scaling characterization of how high-quality synthetic pretraining may reduce bias rather than claiming that it uniformly outperforms real-only training.

3.4 Finite-Sample Conditions for Synthetic Benefit

We now ask when the potential synthetic bias-reduction translates into a strict improvement in the actual risk. In this subsection, under a fixed real-data budget and identical real-stage training, we establish an exact finite-sample necessary-and-sufficient condition for two-stage training to outperform real-only training, and then use a one-dimensional toy model to characterize the roles of synthetic sample size, label noise, and source mismatch. Detailed proof is deferred to Appendix G.

Theorem 8 (Exact condition for strict synthetic benefit) *Let* $\mathbf{w}_N^{\mathrm{real}}$ *and* $\mathbf{w}_{M+N}$ *denote the outputs of real-only and two-stage SGD, respectively. Both procedures start from* $\mathbf{w}_0$*; the two-stage procedure additionally uses* M *synthetic samples before the two procedures undergo the same* N *real updates, with identical samples, order, and stepsizes. Define* $\Phi_N := \prod_{t=N}^{1}(\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)$ *and* $\mathbf{K}_N := \mathbb{E}_{\mathbf{x}_1,\ldots,\mathbf{x}_N}[\Phi_N^\top \mathbf{H} \Phi_N]$*. Under Assumptions 1 and 3,*

$$2\left(\mathbb{E}[\mathcal{E}_1(\mathbf{w}_N^{\mathrm{real}})] - \mathbb{E}[\mathcal{E}_1(\mathbf{w}_{M+N})]\right) = \|\mathbf{w}_0 - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2 - \|\mathbb{E}[\mathbf{w}_M] - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2 - \operatorname{tr}\left(\mathbf{K}_N \operatorname{Cov}(\mathbf{w}_M)\right).$$

Consequently, two-stage training strictly outperforms real-only training if and only if

$$\|\mathbb{E}[\mathbf{w}_M] - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2 + \operatorname{tr}\left(\mathbf{K}_N \operatorname{Cov}(\mathbf{w}_M)\right) < \|\mathbf{w}_0 - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2.$$

This condition compares three quantities. $\|\mathbf{w}_0 - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2$ is the real-only initialization error that remains after N real updates. $\|\mathbb{E}[\mathbf{w}_M] - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2$ is the remaining bias of the synthetic initialization; it includes both finite-M optimization bias and source mismatch. $\operatorname{tr}\left(\mathbf{K}_N \operatorname{Cov}(\mathbf{w}_M)\right)$ is the remaining uncertainty generated by finite synthetic training, including synthetic sampling and label noise. The role of $\mathbf{K}_N$ follows from $\|\mathbf{u}\|_{\mathbf{K}_N}^2 = \mathbb{E}\|\Phi_N \mathbf{u}\|_{\mathbf{H}}^2$, which measures the expected error remaining after applying the same N real-data updates to an initial error $\mathbf{u}$. Therefore, synthetic pretraining is beneficial exactly when $\mathbf{w}_M$ provides a better starting point than $\mathbf{w}_0$ after both are followed by the

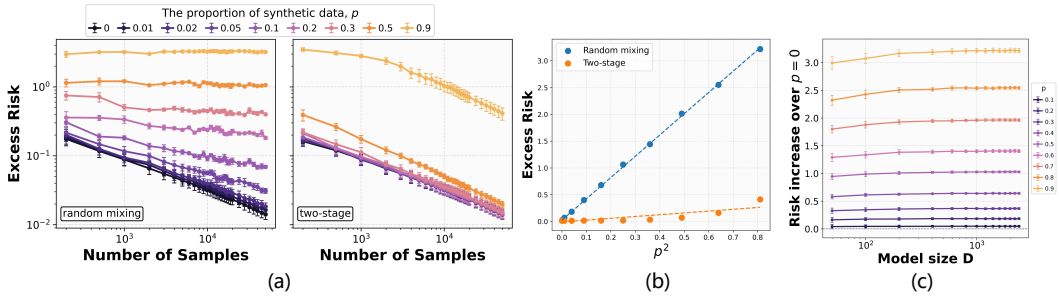


Figure 1: **Simulation results (i).** (a) Excess risk of two protocols in full high-dimensional linear model: mixed training exhibits a clear p-dependent error floor as T scales, while two-stage training continues to improve without saturation. (b) Excess risk at fixed $T = 50000$ versus p^2: under mixing, the excess risk scales approximately linearly with p^2, consistent with the theoretical prediction $p^2\|\boldsymbol{\delta}\|_{\mathbf{H}}^2$, whereas two-stage training shows a much weaker dependence, matching our theory. (c) Risk increase over $p = 0$ under mixed training in the random sketch model at fixed $T = 50000$: the degradation grows with model size D and saturates, consistent with the theoretical scaling $p^2(1 - D^{1-b})$, indicating that larger models amplify the contamination from synthetic mixed training.

same real-data training and it must improve the initialization in directions that cannot be efficiently corrected by the available real data.

Theorem 8 gives an exact criterion once the mean and covariance of $\mathbf{w}_M$ are characterized. In general, however, these quantities depend jointly on the synthetic sample size, label noise, source mismatch, data spectrum, and SGD dynamics, making their individual effects difficult to separate. To disentangle these factors and obtain an explicit sample-size threshold, we next specialize the condition to a one-dimensional toy model.

One-dimensional illustration. Let x be Rademacher, independent of the label noises, and consider

$$y^{\text{real}} = \theta x + \xi_1, \qquad y^{\text{syn}} = (\theta + \delta)x + \xi_2, \qquad \mathbb{E}[\xi_\ell] = 0, \quad \text{Var}(\xi_\ell) = \sigma_\ell^2.$$

Starting from w_0, the synthetic stage performs M SGD-updates with stepsize $1/s$ at iteration s. Afterward, both procedures use the same N real samples and stepsizes $0 < \gamma_t < 1$.

Corollary 9 (One-dimensional strict-benefit condition) *For every $M, N \geq 1$,*

$$\mathbb{E}[\mathcal{E}_1(w_{M+N})] - \mathbb{E}[\mathcal{E}_1(w_N^{\text{real}})] = \frac{1}{2}\left(\prod_{t=1}^{N}(1-\gamma_t)\right)^2 \left[\delta^2 + \frac{\sigma_2^2}{M} - (w_0 - \theta)^2\right].$$

Hence, synthetic pretraining strictly improves the expected risk if and only if $\delta^2 + \sigma_2^2/M < (w_0 - \theta)^2$. Such an improvement is achievable for some M if and only if $\delta^2 < (w_0 - \theta)^2$; under this condition, the exact requirement on the synthetic sample size is $M > \sigma_2^2/\big((w_0 - \theta)^2 - \delta^2\big)$.

This condition displays the bias-variance trade-off directly. The mismatch δ^2 is the error that remains even with infinitely many synthetic samples, whereas σ_2^2/M is the uncertainty caused by finite and noisy synthetic data. Thus, increasing M can eliminate the synthetic variance but cannot compensate for a mismatch larger than the original initialization error. The required synthetic sample size increases linearly with σ_2^2 and diverges as δ^2 approaches $(w_0 - \theta)^2$.

4 Experiments

Simulation. We first validate our theory in a Gaussian high-dimensional linear regression setting with dimension $d = 5000$. Features are drawn from a Gaussian distribution with covariance spectrum $\lambda_i \propto i^{-1.5}$, and the synthetic teacher satisfies $\boldsymbol{w}_2^* = \boldsymbol{w}_1^* + \boldsymbol{\delta}$ with source condition $\lambda_i \mathbb{E}[\delta_i^2] \propto i^{-b}$. We train last-iterate SGD under the mixed and two-stage protocols in Section 2, and evaluate excess risk on the real distribution. This experiments contain two parts. The first part varies the synthetic proportion $p \in (0, 1)$, sample size $T \in [50, 50000]$, and sketch width D to validate the predicted

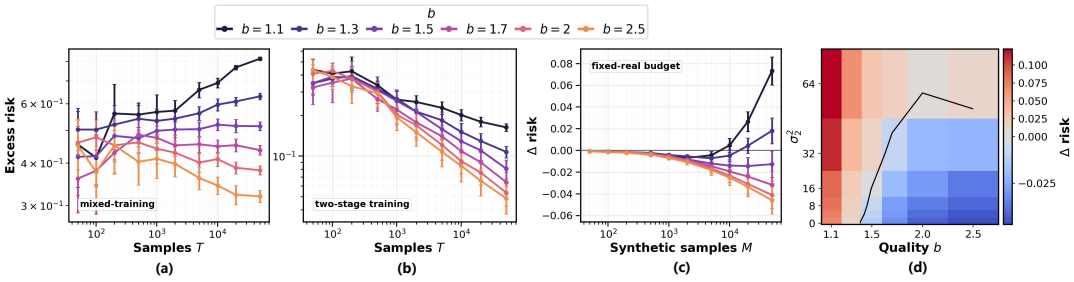


Figure 2: **Simulation results (ii).** (a) Excess risk of mixed training under different synthetic qualities: low-quality synthetic data induces a larger error floor, while increasing b substantially reduces the degradation. (b) Excess risk of two-stage training under different synthetic qualities: two-stage training continues to improve with T for all b, and higher-quality synthetic data leads to lower final risk. (c) Risk change from adding synthetic samples under a fixed real-data budget: synthetic data helps when b is large, but low-quality synthetic data can eventually hurt as M increases. (d) Benefit region over synthetic quality and label noise: additional synthetic data improves real risk only when the teacher mismatch is sufficiently benign and σ_2^2 is not too large.

synthetic-induced error floor. The second part keeps the same high-dimensional setup but varies the synthetic quality b, synthetic sample size M, and synthetic label noise σ_2^2 to study when synthetic data helps. All experiments are repeated with 5 random seeds, , and the results of the two parts are reported in Figure 1 and Figure 2, respectively.

Real-world data. We further extend the experiments to validate our theory on `CIFAR10` [37] using a 5-layer CNN and on `WikiText` $-$ `103` via language model training. Synthetic data is generated by a pretrained teacher model. We observe the same qualitative behavior as predicted by our theory. Full experimental details are provided in Appendix I.

5 Conclusion

We studied how synthetic data affects the generalization of one-pass SGD in high-dimensional linear regression with model shift. We establish finite-sample risk bounds for mixed and two-stage training that characterize how source mismatch enters the SGD dynamics under different training protocols. As a consequence, mixed training exhibits *strong model collapse* while two-stage training avoids this by using synthetic data only in the first stage, showing that collapse is not inevitable through a simple data curriculum. Our random sketch model analysis further establishes scaling-law bounds for both protocols, showing that larger models may amplify synthetic-induced degradation and giving an explicit characterization of how high-quality synthetic pre-training may reduce bias. We also establish an exact finite-sample condition under which two-stage training strictly outperforms real-data-only training with a fixed real-data budget and identical real-stage training. Overall, our results highlight that synthetic data is not inherently harmful or beneficial; its effect depends critically on both its quality and how it is incorporated during training.

Limitations. Our analysis focuses on one-pass SGD with shared covariates. Extensions to joint covariate and model shift, as well as multi-epoch SGD, are natural next steps beyond our current setting. In addition, the mixed-training scaling law is tight only in the optimization-saturated regime, while the two-stage result remains an upper bound; this disparity between the two protocols leaves tight characterizations for data-limited and two-stage settings open.

Acknowledgments

We sincerely thank the reviewers and the area chair for their careful reading, constructive feedback, and thoughtful suggestions, which helped us improve the clarity and presentation of this work. This work is supported in part by NSFC 62306252, Hong Kong GRF award 17312526, Guangdong NSF 2024A1515012444, and the central fund from HKU.

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[2] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go mad. In *The Twelfth International Conference on Learning Representations*, 2023.

[3] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). *Advances in neural information processing systems*, 26, 2013.

[4] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[5] Daniel Barzilai and Ohad Shamir. When models don't collapse: On the consistency of iterative mle. *arXiv preprint arXiv:2505.19046*, 2025.

[6] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1): 151–175, 2010.

[7] Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *Advances in Neural Information Processing Systems*, 33:2576–2586, 2020.

[8] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20, 2007.

[9] Matyas Bohacek and Hany Farid. Nepotistically trained generative-ai models collapse. *arXiv e-prints*, pages arXiv–2311, 2023.

[10] Martin Briesch, Dominik Sobania, and Franz Rothlauf. Large language models suffer from their own output: An analysis of the self-consuming training loop, 2024. URL `https://arxiv.org/abs/2311.16822`.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[12] Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I Abdin. On the diversity of synthetic data and its impact on training large language models. *arXiv preprint arXiv:2410.15226*, 2024.

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

[14] Yuyang Deng and Samory Kpotufe. Mixed-sample sgd: an end-to-end analysis of supervised transfer learning. *arXiv preprint arXiv:2507.04194*, 2025.

[15] Aymeric Dieuleveut and Francis Bach. Non-parametric stochastic approximation with large step sizes, 2016. URL `https://arxiv.org/abs/1408.0361`.

[16] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(101): 1–51, 2017.

[17] Meng Ding, Kaiyi Ji, Di Wang, and Jinhui Xu. Understanding forgetting in continual learning with linear regression. *arXiv preprint arXiv:2405.17583*, 2024.

[18] Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model collapse demystified: The case of regression. *Advances in Neural Information Processing Systems*, 37:46979–47013, 2024.

[19] Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. *arXiv preprint arXiv:2410.04840*, 2024.

[20] Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024.

[21] Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. Self-consuming generative models with curated data provably optimize human preferences. *arXiv preprint arXiv:2407.09499*, 2024.

[22] Aymane El Firdoussi, Mohamed El Amine Seddik, Soufiane Hayou, Reda Alami, Ahmed Alzubaidi, and Hakim Hacid. Maximizing the potential of synthetic data: Insights from random matrix theory. *arXiv preprint arXiv:2410.08942*, 2024.

[23] Anvit Garg, Sohom Bhattacharya, and Pragya Sur. Preventing model collapse under overparametrization: Optimal mixing ratios for interpolation learning and ridge regression. *arXiv preprint arXiv:2509.22341*, 2025.

[24] Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Advances in neural information processing systems*, 32, 2019.

[25] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

[26] Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, 2024.

[27] Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[28] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20555–20565, 2023.

[29] Alex Havrilla, Andrew Dai, Laura O'Mahony, Koen Oostermeijer, Vera Zisler, Alon Albalak, Fabrizio Milo, Sharath Chandra Raparthy, Kanishk Gandhi, Baber Abbasi, et al. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. *arXiv preprint arXiv:2412.02980*, 2024.

[30] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10, 2022.

[31] Ayush Jain, Andrea Montanari, and Eren Sasoglu. Scaling laws for learning with real and surrogate data. *Advances in Neural Information Processing Systems*, 37:110246–110289, 2024.

[32] Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). *arXiv preprint arXiv:1710.09430*, 2017.

[33] Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of machine learning research*, 18(223):1–42, 2018.

[34] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[35] Yeichan Kim, Ilmun Kim, and Seyoung Park. Transfer learning for benign overfitting in high-dimensional linear regression. *arXiv preprint arXiv:2510.15337*, 2025.

[36] Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference On Learning Theory*, pages 1882–1886. PMLR, 2018.

[37] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

[38] Yunwen Lei, Ting Hu, and Ke Tang. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *Journal of Machine Learning Research*, 22(25): 1–41, 2021.

[39] Binghui Li, Fengling Chen, Zixun Huang, Lean Wang, and Lei Wu. Functional scaling laws in kernel regression: Loss dynamics and learning rate schedules. *arXiv preprint arXiv:2509.19189*, 2025.

[40] Binghui Li, Zilin Wang, Fengling Chen, Shiyang Zhao, Ruiheng Zheng, and Lei Wu. Optimal learning-rate schedules under functional scaling laws: Power decay and warmup-stable-decay. *arXiv preprint arXiv:2602.06797*, 2026.

[41] Haoran Li, Jingfeng Wu, and Vladimir Braverman. Fixed design analysis of regularization-based continual learning. In *Conference on lifelong learning agents*, pages 513–533. PMLR, 2023.

[42] Haoran Li, Jingfeng Wu, and Vladimir Braverman. Memory-statistics tradeoff in continual learning with structural regularization. *arXiv preprint arXiv:2504.04039*, 2025.

[43] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.

[44] Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

[45] Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *Advances in Neural Information Processing Systems*, 37:60556–60606, 2024.

[46] Licong Lin, Jingfeng Wu, and Peter L Bartlett. Improved scaling laws in linear regression via data reuse. *arXiv preprint arXiv:2506.08415*, 2025.

[47] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data. *arXiv preprint arXiv:2404.07503*, 2024.

[48] Yuanshi Liu, Haihan Zhang, Qian Chen, and Cong Fang. Optimal algorithms in linear regression under covariate shift: On the importance of precondition. *arXiv preprint arXiv:2502.09047*, 2025.

[49] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On llms-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, 2024.

[50] Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14044–14072, 2024.

[51] Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating sgd saturation with tail-averaging and minibatching. *Advances in Neural Information Processing Systems*, 32, 2019.

[52] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

[53] Reese Pathak, Cong Ma, and Martin Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, pages 17517–17530. PMLR, 2022.

[54] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018.

[55] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[56] Zeyu Qin, Qingxiu Dong, Xingxing Zhang, Li Dong, Xiaolong Huang, Ziyi Yang, Mahmoud Khademi, Dongdong Zhang, Hany Hassan Awadalla, Yi R Fung, et al. Scaling laws of synthetic data for language models. *arXiv preprint arXiv:2503.19551*, 2025.

[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[58] Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. How bad is training on synthetic data? a statistical analysis of language model collapse. *arXiv preprint arXiv:2404.05090*, 2024.

[59] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

[60] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759, 2024.

[61] Javan Tahir, Surya Ganguli, and Grant M Rotskoff. Features are fate: a theory of transfer learning in high-dimensional regression. *arXiv preprint arXiv:2410.08194*, 2024.

[62] Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. *Advances in Neural Information Processing Systems*, 34:21581–21591, 2021.

[63] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*, 2024.

[64] Jinbo Wang, Binghui Li, Zhanpeng Zhou, Mingze Wang, Yuxuan Sun, Jiaqi Zhang, Xunliang Cai, and Lei Wu. Fast catch-up, late switching: Optimal batch size scheduling via functional scaling laws. *arXiv preprint arXiv:2602.14208*, 2026.

[65] Xuezhi Wang and Jeff G Schneider. Generalization bounds for transfer learning under model shift. In *UAI*, pages 922–931, 2015.

[66] Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. The power and limitation of pretraining-finetuning for linear regression under covariate shift. *Advances in Neural Information Processing Systems*, 35:33041–33053, 2022.

[67] Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. In *International conference on machine learning*, pages 24280–24314. PMLR, 2022.

[68] Tingkai Yan, Haodong Wen, Binghui Li, Kairong Luo, Wenguang Chen, and Kaifeng Lyu. Larger datasets can be repeated more: A theoretical analysis of multi-epoch scaling in linear regression. *arXiv preprint arXiv:2511.13421*, 2025.

[69] Dechen Zhang, Junwei Su, and Difan Zou. Learning under quantization for high-dimensional linear regression. *arXiv preprint arXiv:2510.18259*, 2025.

[70] Dechen Zhang, Xuan Tang, Yingyu Liang, and Difan Zou. Scaling laws for precision in high-dimensional linear regression. *arXiv preprint arXiv:2602.19241*, 2026.

[71] Haihan Zhang, Yuanshi Liu, Qianwen Chen, and Cong Fang. The optimality of (accelerated) sgd for high-dimensional quadratic optimization. *arXiv preprint arXiv:2409.09745*, 2024.

[72] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Risk bounds of multi-pass sgd for least squares in the interpolation regime. *Advances in Neural Information Processing Systems*, 35:12909–12920, 2022.

[73] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Benign overfitting of constant-stepsize sgd for linear regression. *Journal of Machine Learning Research*, 24(326):1–58, 2023.

Contents

1 Introduction **1**

2 Preliminaries **3**

3 Main Results **4**

3.1 Mixed training 4

3.2 Two-stage Training 5

3.3 Random Sketch Model & Scaling laws 6

3.4 Finite-Sample Conditions for Synthetic Benefit 8

4 Experiments **9**

5 Conclusion **10**

A Related Work **17**

B Technical Operator Preliminaries **18**

C Mixed Training Upper Bound Analysis **18**

C.1 Excess Risk Decomposition 19

C.2 Bias upper bound 20

C.3 Variance upper bound 20

C.4 Drift upper bound 20

C.5 Fluctuation upper bound 23

C.6 Proof of Theorem 1 28

D Mixed Training Lower Bound Analysis **29**

D.1 The mean recursion 29

D.2 The centered fluctuation term 30

D.3 Proof of Theorem 2 35

D.4 Proof of Corollary 3 36

E Two-stage Training Analysis **37**

E.1 Two-stage training without learning-rate restart 39

F Random Sketch Model **40**

F.1 Mixed Training Upper Bound 40

F.2 Mixed Training Lower Bound 42

F.3 Two-stage Training Analysis 43

G Exact Synthetic-Benefit Condition **46**

H constant stepsize SGD with iterate averaging **48**

H.1 Upper bound analysis 48
H.2 Lower bound analysis 56
H.3 Strong model collapse behavior 58

I Additional Experiments **60**
I.1 CIFAR10 CNN Experiments 60
I.2 Language Model Experiments 61

A Related Work

Synthetic data and model collapse. Training with synthetic data has been shown to induce *model collapse* [59], a phenomenon in which model performance degrades significantly when consistently trained with poor-quality synthetic data, as has been empirically verified by a number of subsequent works [28, 9, 2, 26, 10, 60]. Theoretical studies of model collapse typically frame it as a regression problem under distribution mismatch between real and synthetic data [20, 18, 58], showing how synthetic data degrade performance, while others explore how to avoid collapse or leverage synthetic data to benefit learning [58, 22, 5, 23]. In particular, Dohmatob et al. [19] introduce the concept of *strong model collapse*, where even a small proportion of synthetic data prevents model performance from improving under data scaling, leaving a non-vanishing excess risk floor. Moreover, such collapse cannot be mitigated by simple strategies such as data reweighting ([31, 21]) unless the proportion of synthetic data vanishes asymptotically. However, most of them do not capture the dynamics of practical training, where the order and interplay of data sources matter. In this paper, we study finite-sample one-pass SGD and show that while mixed training induces strong model collapse, it can be avoided by a simple two-stage training protocol.

Learning with multiple data sources. Learning from multiple data sources has been widely studied in transfer learning, domain adaptation, and continual learning under covariate or model shift [8, 6, 65, 38, 36, 27, 53, 41, 61, 42, 35]. However, most of them focus on statistical estimators rather than optimization dynamics. More recently, several works have analyzed SGD under covariate shift. Wu et al. [66] study pretraining-finetuning under covariate shift via SGD and establish excess risk bounds. Ding et al. [17] extend the analysis to continual learning settings and show how covariate shifts across tasks affect SGD dynamics and forgetting. Liu et al. [48] characterize the minimax optimality of SGD-type methods for high-dimensional linear regression under covariate shift. Beyond standard covariate shift, Deng and Kpotufe [14] propose an adaptive mixed-sample SGD that further accounts for model shift while achieving transfer-optimal target risk. In this paper, we consider different data sources with shared covariates but mismatched labeling functions, as commonly arises with synthetic data, and characterize how synthetic data affects finite-sample SGD dynamics.

SGD in high-dimensional linear regression. The generalization of stochastic gradient descent (SGD) in linear regression has been extensively studied. In the classical underparameterized regime, a large number of works have studied its behavior [55, 3, 13, 16, 32, 33, 54]. One-pass SGD under different iterate schemes (e.g., last iterate, averaging) has also been rigorously studied in the overparameterized setting [15, 24, 7, 62, 67, 73, 71, 69], providing framework for analyzing the generalization of SGD in high-dimensional regime. In parallel, Several works have extended the analysis to multipass SGD [44, 54, 51, 38, 72]. More recently, a line of research has focused on the theoretical scaling laws of SGD in high-dimensional linear regression. Lin et al. [45] analyzed the last-iterate test error of one-pass SGD in a random sketch model, providing the first systematic derivation of a finite-sample joint scaling law. Lin et al. [46] further extended their analysis to multi-epoch SGD, followed by an even finer-grained characterization from Yan et al. [68]. Li et al. [39] established functional scaling laws and analyzed optimal learning rate and batch size schedules [40, 64]. Overall, this paper builds on high-dimensional SGD and scaling-law analyses, and extend them from single-source learning to a two-source setting where source mismatch induces new drift and fluctuation effects.

B Technical Operator Preliminaries

Operators. We first summarize the linear operators (on symmetric matrices) to be used in the proof:

$$\mathcal{I} = \mathbf{I} \otimes \mathbf{I}, \quad \mathcal{M} = \mathbb{E}[\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x}], \quad \tilde{\mathcal{M}} = \mathbf{H} \otimes \mathbf{H},$$
$$\mathcal{T}_t = \mathbf{H} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H} - \gamma_t \mathcal{M}, \quad \tilde{\mathcal{T}}_t = \mathbf{H} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H} - \gamma_t \mathbf{H} \otimes \mathbf{H}.$$

We use the notation $\mathcal{O} \circ \mathbf{A}$ to denotes the operator $\mathcal{O}$ acting on a symmetric matrix $\mathbf{A}$. For example, with these definitions, we have that for a symmetric matrix $\mathbf{A}$, [73]:

$$\mathcal{I} \circ \mathbf{A} = \mathbf{A}, \quad \mathcal{M} \circ \mathbf{A} = \mathbb{E}[(\mathbf{x}^\top \mathbf{A} \mathbf{x}) \mathbf{x}\mathbf{x}^\top], \quad \tilde{\mathcal{M}} \circ \mathbf{A} = \mathbf{H}\mathbf{A}\mathbf{H},$$
$$(\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \mathbf{A} = \mathbb{E}[(\mathbf{I} - \gamma_t \mathbf{x}\mathbf{x}^\top)\mathbf{A}(\mathbf{I} - \gamma_t \mathbf{x}\mathbf{x}^\top)], \quad (\mathcal{I} - \gamma_t \tilde{\mathcal{T}}_t) \circ \mathbf{A} = (\mathbf{I} - \gamma_t \mathbf{H})\mathbf{A}(\mathbf{I} - \gamma_t \mathbf{H}). \tag{2}$$

For the linear operators we have the following technical lemma from Zou et al. [73].

Lemma 10 (Lemma B.1, Zou et al. [73]) *An operator $\mathcal{O}$ defined on symmetric matrices is called PSD mapping, if $\mathbf{A} \succeq 0$ implies $\mathcal{O} \circ \mathbf{A} \succeq 0$. Then we have*

1. *$\mathcal{M}$ and $\widetilde{\mathcal{M}}$ are both PSD mappings.*
2. *$\mathcal{I} - \gamma \mathcal{T}_t$ and $\mathcal{I} - \gamma \widetilde{\mathcal{T}}_t$ are both PSD mappings.*
3. *$\mathcal{M} - \widetilde{\mathcal{M}}$ and $\widetilde{\mathcal{T}}_t - \mathcal{T}_t$ are both PSD mappings.*
4. *If $0 < \gamma_t < 1/\lambda_1$, then $\widetilde{\mathcal{T}}_t^{-1}$ exists, and is a PSD mapping.*
5. *If $0 < \gamma_t < 1/(\alpha \operatorname{tr}(\mathbf{H}))$, then $\mathcal{T}_t^{-1} \circ \mathbf{A}$ exists for PSD matrix $\mathbf{A}$, and $\mathcal{T}_t^{-1}$ is a PSD mapping.*

Proof. See proof of Lemma B.1 in Zou et al. [73]. ∎

C Mixed Training Upper Bound Analysis

Local notation. For notational simplicity, throughout this section we depart slightly from the notation in the main text and use N to denote the total sample size. Thus, the numbers of synthetic and real samples are pN and $(1-p)N$, respectively, where $p \in [0,1]$ is the synthetic-data proportion. We assume that pN and $(1-p)N$ are integers; replacing them by the corresponding floor or ceiling does not affect any of our results.

Fixed-budget sampling model. We formalize the random-order mixed-training protocol used in our analysis. Let

$$\mathcal{Z}_{N,p} := \left\{ \mathbf{z} = (z_1, \ldots, z_N) \in \{0,1\}^N : \sum_{t=1}^{N} z_t = pN \right\},$$

where $z_t = 1$ indicates that the sample at iteration t is synthetic and $z_t = 0$ indicates that it is real. We first draw the source schedule $\mathbf{z}$ uniformly from $\mathcal{Z}_{N,p}$. Conditional on the complete source schedule $\mathbf{z}$, the training samples are generated according to

$$\mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \mathcal{D}, \qquad y_t = \langle \mathbf{x}_t, \mathbf{w}_1^* + z_t \boldsymbol{\delta} \rangle + \xi_t, \qquad t = 1, \ldots, N,$$

where $\boldsymbol{\delta} := \mathbf{w}_2^* - \mathbf{w}_1^*$ and

$$\xi_t \mid z_t \sim \begin{cases} N(0, \sigma_1^2), & z_t = 0, \\ N(0, \sigma_2^2), & z_t = 1. \end{cases}$$

Conditional on $\mathbf{z}$, the noises $\{\xi_t\}_{t=1}^N$ are independent across iterations and independent of all covariates.

This construction is equivalent in distribution to independently drawing pN complete samples from $\mathcal{P}_2$ and $(1-p)N$ complete samples from $\mathcal{P}_1$, and then uniformly shuffling their union. Indeed, a uniform shuffle induces a source schedule that is uniform over $\mathcal{Z}_{N,p}$. Conditional on any such schedule, the observations at positions with $z_t = 1$ are independent samples from $\mathcal{P}_2$, while those at positions with $z_t = 0$ are independent samples from $\mathcal{P}_1$. Since samples within each source are exchangeable, the two constructions induce the same joint distribution over the ordered training sequence.

Our analysis does not assume independence of the source indicators. To identify the conditional independence used in the SGD recursions, define the filtration

$$\mathcal{F}_{t-1} := \sigma(\mathbf{z}, \mathbf{x}_1, \xi_1, \ldots, \mathbf{x}_{t-1}, \xi_{t-1}) .$$

Since the real and synthetic distributions share the same covariate marginal $\mathcal{D}$, the current covariate $\mathbf{x}_t$ is independent of $\mathcal{F}_{t-1}$ and satisfies

$$\mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top \mid \mathcal{F}_{t-1}] = \mathbf{H}.$$

Moreover, the sample-wise noise construction gives

$$\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}, \mathbf{x}_t] = 0.$$

Consequently, the noise cross terms appearing in the bias–variance recursions vanish. Averaging over the source schedule further yields

$$\mathbb{E}[\xi_t^2 \mathbf{x}_t \mathbf{x}_t^\top] = \bar{\sigma}^2 \mathbf{H}, \qquad \bar{\sigma}^2 := (1-p)\sigma_1^2 + p\sigma_2^2.$$

Therefore, the covariance and noise recursions used below remain exact under the fixed-budget sampling protocol. They rely on fresh covariates and sample-wise conditionally independent noises, rather than on independence of the source indicators.

C.1 Excess Risk Decomposition

Starting from $\mathbf{w}_0 \in \mathbb{R}^{\mathcal{H}}$, SGD on the squared loss performs

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \gamma_t (y_t - \mathbf{x}_t^\top \mathbf{w}_{t-1}) \mathbf{x}_t, \qquad t \geq 1.$$

Substituting the label model yields

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \gamma_t \big(\mathbf{x}_t^\top (\mathbf{w}_1^* + z_t \boldsymbol{\delta} - \mathbf{w}_{t-1}) + \xi_t \big) \mathbf{x}_t.$$

Let $\boldsymbol{\eta}_t := \mathbf{w}_t - \mathbf{w}_1^*$ denote the parameter error relative to the real-data optimum. Then equation (3) serves as the starting point of the analysis:

$$\boldsymbol{\eta}_t = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top) \boldsymbol{\eta}_{t-1} + \gamma_t z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} + \gamma_t \xi_t \mathbf{x}_t. \tag{3}$$

With as light abuse of probability spaces, one can view the centered SGD iterates as the sum of four random processes:

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}_t^{\text{bias}} + \boldsymbol{\eta}_t^{\text{var}} + \boldsymbol{\eta}_t^{\text{drift}} + \boldsymbol{\eta}_t^{\text{fluct}}, \qquad t = 1, 2, \ldots N.$$

where

$$\begin{cases} \boldsymbol{\eta}_t^{\text{bias}} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top) \boldsymbol{\eta}_{t-1}^{\text{bias}}; \\ \boldsymbol{\eta}_0^{\text{bias}} = \mathbf{w}_0 - \mathbf{w}_1^*, \end{cases} \qquad \begin{cases} \boldsymbol{\eta}_t^{\text{var}} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top) \boldsymbol{\eta}_{t-1}^{\text{var}} + \gamma_t \xi_t \mathbf{x}_t; \\ \boldsymbol{\eta}_0^{\text{var}} = \mathbf{0}, \end{cases}$$

$$\begin{cases} \boldsymbol{\eta}_t^{\text{drift}} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top) \boldsymbol{\eta}_{t-1}^{\text{drift}} + \gamma_t p \mathbf{H} \boldsymbol{\delta}; \\ \boldsymbol{\eta}_0^{\text{drift}} = \mathbf{0}, \end{cases} \qquad \begin{cases} \boldsymbol{\eta}_t^{\text{fluct}} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top) \boldsymbol{\eta}_{t-1}^{\text{fluct}} + \gamma_t z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} - \gamma_t p \mathbf{H} \boldsymbol{\delta}; \\ \boldsymbol{\eta}_0^{\text{fluct}} = \mathbf{0}, \end{cases}$$

Moreover, we denote

$$\mathbf{B}_t = \mathbb{E}[\boldsymbol{\eta}_t^{\text{bias}} \otimes \boldsymbol{\eta}_t^{\text{bias}}], \quad \mathbf{V}_t = \mathbb{E}[\boldsymbol{\eta}_t^{\text{var}} \otimes \boldsymbol{\eta}_t^{\text{var}}],$$
$$\mathbf{D}_t = \mathbb{E}[\boldsymbol{\eta}_t^{\text{drift}} \otimes \boldsymbol{\eta}_t^{\text{drift}}], \quad \mathbf{F}_t = \mathbb{E}[\boldsymbol{\eta}_t^{\text{fluct}} \otimes \boldsymbol{\eta}_t^{\text{fluct}}].$$

Using the definition of $\mathbf{B}_t, \mathbf{V}_t, \mathbf{D}_t, \mathbf{F}_t$, we can then decompose Excess Risk using Cauchy–Schwarz Inequality:

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(\mathbf{w}_N) - \mathcal{R}(\mathbf{w}_1^*)] &= \frac{1}{2}\big\langle \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_N \otimes \boldsymbol{\eta}_N]\big\rangle \\
&= \frac{1}{2}\big\langle \mathbf{H}, \mathbb{E}[(\boldsymbol{\eta}_N^{\mathrm{bias}} + \boldsymbol{\eta}_N^{\mathrm{var}} + \boldsymbol{\eta}_N^{\mathrm{drift}} + \boldsymbol{\eta}_N^{\mathrm{fluct}}) \otimes (\boldsymbol{\eta}_N^{\mathrm{bias}} + \boldsymbol{\eta}_N^{\mathrm{var}} + \boldsymbol{\eta}_N^{\mathrm{drift}} + \boldsymbol{\eta}_N^{\mathrm{fluct}})]\big\rangle \\
&\le 2\big\langle \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_N^{\mathrm{bias}} \otimes \boldsymbol{\eta}_N^{\mathrm{bias}}]\big\rangle + 2\big\langle \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_N^{\mathrm{var}} \otimes \boldsymbol{\eta}_N^{\mathrm{var}}]\big\rangle \\
&\quad + 2\big\langle \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_N^{\mathrm{drift}} \otimes \boldsymbol{\eta}_N^{\mathrm{drift}}]\big\rangle + 2\big\langle \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_N^{\mathrm{fluct}} \otimes \boldsymbol{\eta}_N^{\mathrm{fluct}}]\big\rangle \\
&= 2\big\langle \mathbf{H}, \mathbf{B}_N\big\rangle + 2\big\langle \mathbf{H}, \mathbf{V}_N\big\rangle + 2\big\langle \mathbf{H}, \mathbf{D}_N\big\rangle + 2\big\langle \mathbf{H}, \mathbf{F}_N\big\rangle.
\end{aligned}$$

C.2 Bias upper bound

Using the defined operators, the update rule of the iterates imply the following recursive form of $\mathbf{B}_t$:

$$\mathbf{B}_t = (\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \mathbf{B}_{t-1}, \quad \mathbf{B}_0 = \boldsymbol{\eta}_0 \otimes \boldsymbol{\eta}_0, \tag{4}$$

Note that this bias term $\big\langle \mathbf{H}, \mathbf{B}_N \big\rangle$ is the same as that in [67], so we can apply Lemma in [67] to bound it directly:

Lemma 11 (A bias upper bound) *Suppose Assumptions 1 and 2 hold. Let $N_{\mathtt{eff}} = N/\log N$. Consider* (4)*. Suppose $\gamma < 1/(4\alpha \operatorname{tr}(\mathbf{H}))$. We have*

$$\big\langle \mathbf{H}, \mathbf{B}_N \big\rangle \lesssim \Big\| \prod_{t=1}^{N} (\mathbf{I} - \gamma_t \mathbf{H})(\mathbf{w}_0 - \mathbf{w}^*) \Big\|_{\mathbf{H}}^2 + \alpha \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}/\gamma N_{\mathtt{eff}} + \mathbf{H}_{k^*:\infty}}} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}$$

where $k^* = \max\{k : \lambda_k \ge \frac{1}{\gamma N_{\mathtt{eff}}}\}$ and $D_{\mathtt{eff}} := k^* + \gamma^2 N_{\mathtt{eff}}^2 \sum_{i>k^*} \lambda_i^2$.

Proof. See proof of Theorem D.1 in [66]. ∎

C.3 Variance upper bound

Using the defined operators, the update rule of the iterates imply the following recursive form of $\mathbf{V}_t$:

$$\mathbf{V}_t = (\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \mathbf{V}_{t-1} + \gamma_t^2 \boldsymbol{\Sigma}_t, \qquad \mathbf{V}_0 = \mathbf{0}. \tag{5}$$

where $\boldsymbol{\Sigma}_t = \mathbb{E}[\xi_t^2 \mathbf{x}_t \mathbf{x}_t^\top]$. Note that $\{\mathbf{x}_t\}$ are independent of $\{z_t\}$. So we have:

$$\boldsymbol{\Sigma}_t = \mathbb{E}[\mathbb{E}[\xi_t^2 \mathbf{x}_t \mathbf{x}_t^\top] | z_t] = (1-p)\boldsymbol{\Sigma}_1 + p\boldsymbol{\Sigma}_2 \preceq ((1-p)\sigma_1^2 + p\sigma_2^2)\mathbf{H}.$$

Then we can apply Lemma C.2 in [67] to bound variance term $\big\langle \mathbf{H}, \mathbf{V}_N \big\rangle$ by simply replace σ^2 with $(1-p)\sigma_1^2 + p\sigma_2^2$.

Lemma 12 (A variance bound) *Suppose Assumptions 1, 2 and 3 hold. Consider* (5)*. Let $N_{\mathtt{eff}} = N/\log N$. Suppose $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$. We have*

$$\big\langle \mathbf{H}, \mathbf{V}_N \big\rangle \le \frac{8\sigma^2}{1 - \gamma\alpha \operatorname{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}},$$

where $k^* = \max\{k : \lambda_k \ge \frac{1}{\gamma N_{\mathtt{eff}}}\}$ and $D_{\mathtt{eff}} := k^* + \gamma^2 N_{\mathtt{eff}}^2 \sum_{i>k^*} \lambda_i^2$.

Proof. See proof of Theorem C.2 in [67]. ∎

C.4 Drift upper bound

We now analyze the drift term

$$\big\langle \mathbf{H}, \mathbf{D}_N \big\rangle = \big\langle \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_N^{\mathrm{drift}} \otimes \boldsymbol{\eta}_N^{\mathrm{drift}}] \big\rangle.$$

Lemma 13 (A drift bound) *Suppose Assumptions 1, 2 and 3 hold. Consider* (5). *Let* $N_{\texttt{eff}} = N/\log N$. *Suppose* $\gamma < 1/(\alpha\,\mathrm{tr}(\mathbf{H}))$. *We have*

$$\big\langle \mathbf{H}, \mathbf{D}_N \big\rangle \le p^2 \Big\| \Big(\mathbf{I} - \prod_{t=1}^{N} (\mathbf{I} - \gamma_t \mathbf{H})\Big)\boldsymbol{\delta} \Big\|_{\mathbf{H}}^2 + \frac{8\alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1 - \gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \frac{D_{\texttt{eff}}}{N_{\texttt{eff}}},$$

where $k^* = \max\{k : \lambda_k \ge \frac{1}{\gamma N_{\texttt{eff}}}\}$ *and* $D_{\texttt{eff}} := k^* + \gamma^2 N_{\texttt{eff}}^2 \sum_{i>k^*} \lambda_i^2$.

Proof.

Recall that $\boldsymbol{\eta}_t^{\mathrm{drift}}$ satisfies

$$\boldsymbol{\eta}_t^{\mathrm{drift}} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{drift}} + \gamma_t p \mathbf{H}\boldsymbol{\delta}, \qquad \boldsymbol{\eta}_0^{\mathrm{drift}} = \mathbf{0}.$$

We first center $\boldsymbol{\eta}_t^{\mathrm{drift}}$ by separating its mean and fluctuation parts. Define $\overline{\boldsymbol{\eta}}_t^{\mathrm{drift}} := \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{drift}}]$. Since $\mathbf{x}_t$ is independent of $\boldsymbol{\eta}_{t-1}^{\mathrm{drift}}$ and $\mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top] = \mathbf{H}$, taking expectation on both sides yields

$$\overline{\boldsymbol{\eta}}_t^{\mathrm{drift}} = (\mathbf{I} - \gamma_t \mathbf{H})\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}} + \gamma_t p \mathbf{H}\boldsymbol{\delta}, \qquad \overline{\boldsymbol{\eta}}_0^{\mathrm{drift}} = \mathbf{0}. \tag{6}$$

Define the centered drift fluctuation $\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}} := \boldsymbol{\eta}_t^{\mathrm{drift}} - \overline{\boldsymbol{\eta}}_t^{\mathrm{drift}}$ and subtracting (6) from the recursion of $\boldsymbol{\eta}_t^{\mathrm{drift}}$, we obtain

$$\begin{aligned}
\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}} &= (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{drift}} - (\mathbf{I} - \gamma_t \mathbf{H})\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}} \\
&= (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\widetilde{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}} + \gamma_t (\mathbf{H} - \mathbf{x}_t \mathbf{x}_t^\top)\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}}, \qquad \widetilde{\boldsymbol{\eta}}_0^{\mathrm{drift}} = \mathbf{0}.
\end{aligned} \tag{7}$$

By definition,

$$\mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}}] = \mathbf{0}, \qquad \forall t \ge 1.$$

Using $\boldsymbol{\eta}_t^{\mathrm{drift}} = \overline{\boldsymbol{\eta}}_t^{\mathrm{drift}} + \widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}}$, we can expand $\mathbf{D}_t$ as

$$\begin{aligned}
\mathbf{D}_t &= \mathbb{E}\big[(\overline{\boldsymbol{\eta}}_t^{\mathrm{drift}} + \widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}}) \otimes (\overline{\boldsymbol{\eta}}_t^{\mathrm{drift}} + \widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}})\big] \\
&= \overline{\boldsymbol{\eta}}_t^{\mathrm{drift}} \otimes \overline{\boldsymbol{\eta}}_t^{\mathrm{drift}} + \mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}} \otimes \widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}}] + \overline{\boldsymbol{\eta}}_t^{\mathrm{drift}} \otimes \mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}}] + \mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}}] \otimes \overline{\boldsymbol{\eta}}_t^{\mathrm{drift}}.
\end{aligned}$$

Since $\mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}}] = \mathbf{0}$, the two cross terms vanish. Hence

$$\mathbf{D}_t = \overline{\boldsymbol{\eta}}_t^{\mathrm{drift}} \otimes \overline{\boldsymbol{\eta}}_t^{\mathrm{drift}} + \widetilde{\mathbf{D}}_t, \qquad \widetilde{\mathbf{D}}_t := \mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}} \otimes \widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}}].$$

Therefore the drift contribution to the excess risk can be decomposed into two parts:

$$\big\langle \mathbf{H}, \mathbf{D}_N \big\rangle = \big\| \overline{\boldsymbol{\eta}}_N^{\mathrm{drift}} \big\|_{\mathbf{H}}^2 + \big\langle \mathbf{H}, \widetilde{\mathbf{D}}_N \big\rangle. \tag{8}$$

In the following, we will bound the two terms in (8) separately.

Mean part of the drift term. We now derive the explicit form of $\big\|\overline{\boldsymbol{\eta}}_N^{\mathrm{drift}}\big\|_{\mathbf{H}}^2$. Recall from (6) that

$$\overline{\boldsymbol{\eta}}_t^{\mathrm{drift}} = (\mathbf{I} - \gamma_t \mathbf{H})\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}} + \gamma_t p \mathbf{H}\boldsymbol{\delta}, \qquad \overline{\boldsymbol{\eta}}_0^{\mathrm{drift}} = \mathbf{0}.$$

Unrolling the recursion gives

$$\overline{\boldsymbol{\eta}}_N^{\mathrm{drift}} = p \sum_{s=1}^{N} \gamma_s \left(\prod_{t=s+1}^{N} (\mathbf{I} - \gamma_t \mathbf{H}) \right) \mathbf{H}\boldsymbol{\delta}. \tag{9}$$

Since each factor $(\mathbf{I} - \gamma_t \mathbf{H})$ is a polynomial in $\mathbf{H}$, it commutes with $\mathbf{H}$. Hence

$$\left(\prod_{t=s+1}^{N} (\mathbf{I} - \gamma_t \mathbf{H}) \right) \mathbf{H} = \mathbf{H} \left(\prod_{t=s+1}^{N} (\mathbf{I} - \gamma_t \mathbf{H}) \right),$$

and therefore

$$\sum_{s=1}^{N} \gamma_s \left(\prod_{t=s+1}^{N} (\mathbf{I} - \gamma_t \mathbf{H}) \right) \mathbf{H} = \mathbf{I} - \prod_{t=1}^{N} (\mathbf{I} - \gamma_t \mathbf{H}).$$

Substituting this identity into (9), we obtain

$$\overline{\boldsymbol{\eta}}_N^{\mathrm{drift}} = p\Big(\mathbf{I} - \prod_{t=1}^{N}(\mathbf{I} - \gamma_t \mathbf{H})\Big)\boldsymbol{\delta}.$$

As a consequence,

$$\big\|\overline{\boldsymbol{\eta}}_N^{\mathrm{drift}}\big\|_{\mathbf{H}}^2 = p^2\Big\|\Big(\mathbf{I} - \prod_{t=1}^{N}(\mathbf{I} - \gamma_t \mathbf{H})\Big)\boldsymbol{\delta}\Big\|_{\mathbf{H}}^2. \tag{10}$$

This is exactly the first drift term appearing in the Lemma.

Fluctuation part of the drift term. We now bound the second term in (8), namely

$$\big\langle \mathbf{H}, \widetilde{\mathbf{D}}_N \big\rangle, \qquad \widetilde{\mathbf{D}}_t := \mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}} \otimes \widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}}].$$

Recall from (7) that

$$\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}} = (\mathbf{I} - \gamma_t \mathbf{x}_t\mathbf{x}_t^\top)\widetilde{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}} + \gamma_t(\mathbf{H} - \mathbf{x}_t\mathbf{x}_t^\top)\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}}, \qquad \widetilde{\boldsymbol{\eta}}_0^{\mathrm{drift}} = \mathbf{0}.$$

Define

$$\widetilde{\boldsymbol{\Sigma}}_t^{\mathrm{drift}} := \mathbb{E}\big[\big((\mathbf{H} - \mathbf{x}_t\mathbf{x}_t^\top)\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}}\big) \otimes \big((\mathbf{H} - \mathbf{x}_t\mathbf{x}_t^\top)\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}}\big)\big].$$

Since $\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}}$ is deterministic, the recursion for $\widetilde{\mathbf{D}}_t$ takes the form

$$\widetilde{\mathbf{D}}_t = (\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \widetilde{\mathbf{D}}_{t-1} + \gamma_t^2 \widetilde{\boldsymbol{\Sigma}}_t^{\mathrm{drift}}, \qquad \widetilde{\mathbf{D}}_0 = \mathbf{0}. \tag{11}$$

Indeed, the cross term vanishes because

$$\mathbb{E}\big[(\mathbf{H} - \mathbf{x}_t\mathbf{x}_t^\top)\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}}\big] = (\mathbf{H} - \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top])\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}} = \mathbf{0}.$$

Let $\mathbf{A}_t := \overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}} \otimes \overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}}$, then

$$\begin{aligned}
\widetilde{\boldsymbol{\Sigma}}_t^{\mathrm{drift}} &= \mathbb{E}\big[(\mathbf{H} - \mathbf{x}_t\mathbf{x}_t^\top)\mathbf{A}_t(\mathbf{H} - \mathbf{x}_t\mathbf{x}_t^\top)\big] \\
&= \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top \mathbf{A}_t \mathbf{x}_t\mathbf{x}_t^\top] - \mathbf{H}\mathbf{A}_t\mathbf{H} \\
&= (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{A}_t.
\end{aligned}$$

Since $\mathcal{M} - \widetilde{\mathcal{M}}$ is a PSD mapping by Lemma 10, we have $\widetilde{\boldsymbol{\Sigma}}_t^{\mathrm{drift}} \succeq 0$. On the other hand, by Assumption 2,

$$\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top \mathbf{A}_t \mathbf{x}_t\mathbf{x}_t^\top] = \mathcal{M} \circ \mathbf{A}_t \preceq \alpha\ \mathrm{tr}(\mathbf{H}\mathbf{A}_t)\,\mathbf{H} = \alpha\,\|\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}}\|_{\mathbf{H}}^2\,\mathbf{H}.$$

Therefore,

$$\widetilde{\boldsymbol{\Sigma}}_t^{\mathrm{drift}} \preceq \alpha\,\|\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}}\|_{\mathbf{H}}^2\,\mathbf{H}. \tag{12}$$

Next, using (C.4), $\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}} = p\Big(\mathbf{I} - \prod_{s=1}^{t-1}(\mathbf{I} - \gamma_s\mathbf{H})\Big)\boldsymbol{\delta}$, hence

$$\begin{aligned}
\|\overline{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}}\|_{\mathbf{H}}^2 &= p^2 \sum_i \lambda_i \left(1 - \prod_{s=1}^{t-1}(1 - \gamma_s\lambda_i)\right)^2 \delta_i^2 \\
&\le p^2 \sum_i \lambda_i \delta_i^2 = p^2\|\boldsymbol{\delta}\|_{\mathbf{H}}^2.
\end{aligned} \tag{13}$$

Combining (12) and (13), we obtain

$$\widetilde{\boldsymbol{\Sigma}}_t^{\mathrm{drift}} \preceq \alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \mathbf{H}. \tag{14}$$

The recursion (11) has exactly the same form as the variance recursion (5). Therefore, applying Lemma 12 with

$$\sigma^2 \ \leftarrow\ \alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2,$$

we immediately get

$$\big\langle \mathbf{H}, \widetilde{\mathbf{D}}_N \big\rangle \le \frac{8\alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}, \tag{15}$$

where

$$k^* = \max\left\{k : \lambda_k \ge \frac{1}{\gamma N_{\mathtt{eff}}}\right\}, \qquad D_{\mathtt{eff}} := k^* + \gamma^2 N_{\mathtt{eff}}^2 \sum_{i>k^*} \lambda_i^2.$$

Combining (10) and (15), we conclude that

$$\big\langle \mathbf{H}, \mathbf{D}_N \big\rangle = \big\|\overline{\boldsymbol{\eta}}_N^{\mathrm{drift}}\big\|_{\mathbf{H}}^2 + \big\langle \mathbf{H}, \widetilde{\mathbf{D}}_N \big\rangle \le p^2 \Big\| \Big(\mathbf{I} - \prod_{t=1}^{N} (\mathbf{I} - \gamma_t \mathbf{H})\Big)\boldsymbol{\delta}\Big\|_{\mathbf{H}}^2 + \frac{8\alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}.$$

■

C.5 Fluctuation upper bound

We now analyze the drift term

$$\big\langle \mathbf{H}, \mathbf{F}_N \big\rangle = \big\langle \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_N^{\mathrm{fluct}} \otimes \boldsymbol{\eta}_N^{\mathrm{fluct}}] \big\rangle.$$

Lemma 14 (A fluctuation bound) *Suppose Assumptions 1, 2 and 3 hold. Consider* (5)*. Let* $N_{\mathtt{eff}} = N/\log N$*. Suppose* $\gamma < 1/(\alpha\,\mathrm{tr}(\mathbf{H}))$*. We have*

$$\big\langle \mathbf{H}, \mathbf{F}_N \big\rangle \le \frac{8Np(1-p)\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{(N-1)\big(1-\alpha\gamma\,\mathrm{tr}(\mathbf{H})\big)} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}} + \frac{16\alpha p \|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}},$$

where $k^* = \max\{k : \lambda_k \ge \frac{1}{\gamma N_{\mathtt{eff}}}\}$ *and* $D_{\mathtt{eff}} := k^* + \gamma^2 N_{\mathtt{eff}}^2 \sum_{i>k^*} \lambda_i^2$*.*

Proof.

Recall that $\boldsymbol{\eta}_t^{\mathrm{fluct}}$ satisfies

$$\boldsymbol{\eta}_t^{\mathrm{fluct}} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{fluct}} + \gamma_t z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} - \gamma_t p \mathbf{H}\boldsymbol{\delta}, \qquad \boldsymbol{\eta}_0^{\mathrm{fluct}} = \mathbf{0}.$$

We decompose the driving term as

$$z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} - p\mathbf{H}\boldsymbol{\delta} = z_t(\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta} + (z_t - p)\mathbf{H}\boldsymbol{\delta}.$$

Accordingly, define two auxiliary processes

$$\boldsymbol{\eta}_t^{\mathrm{fluct}} = \boldsymbol{\eta}_t^{\mathrm{fluct},1} + \boldsymbol{\eta}_t^{\mathrm{fluct},2}, \qquad t \ge 0,$$

where

$$\begin{cases} \boldsymbol{\eta}_t^{\mathrm{fluct},1} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{fluct},1} + \gamma_t z_t (\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}, \\ \boldsymbol{\eta}_0^{\mathrm{fluct},1} = \mathbf{0}, \end{cases} \qquad \begin{cases} \boldsymbol{\eta}_t^{\mathrm{fluct},2} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{fluct},2} + \gamma_t (z_t - p)\mathbf{H}\boldsymbol{\delta}, \\ \boldsymbol{\eta}_0^{\mathrm{fluct},2} = \mathbf{0}. \end{cases}$$

Then by Cauchy–Schwarz,

$$\begin{aligned} \big\langle \mathbf{H}, \mathbf{F}_N \big\rangle &= \big\langle \mathbf{H}, \mathbb{E}\big[(\boldsymbol{\eta}_N^{\mathrm{fluct},1} + \boldsymbol{\eta}_N^{\mathrm{fluct},2}) \otimes (\boldsymbol{\eta}_N^{\mathrm{fluct},1} + \boldsymbol{\eta}_N^{\mathrm{fluct},2})\big] \big\rangle \\ &\le 2\big\langle \mathbf{H}, \mathbf{F}_N^{(1)} \big\rangle + 2\big\langle \mathbf{H}, \mathbf{F}_N^{(2)} \big\rangle, \end{aligned} \tag{16}$$

where

$$\mathbf{F}_t^{(1)} := \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{fluct},1} \otimes \boldsymbol{\eta}_t^{\mathrm{fluct},1}], \qquad \mathbf{F}_t^{(2)} := \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{fluct},2} \otimes \boldsymbol{\eta}_t^{\mathrm{fluct},2}].$$

We first analyze the term $\mathbf{F}_t^{(1)}$. Expanding the outer product gives

$$\begin{aligned} \mathbf{F}_t^{(1)} &= \mathbb{E}\Big[\Big((\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{fluct},1} + \gamma_t z_t (\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\Big) \\ &\qquad \otimes \Big((\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{fluct},1} + \gamma_t z_t (\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\Big)\Big] \\ &= \mathbb{E}\Big[(\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{fluct},1} \otimes (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{fluct},1}\Big] \\ &\quad + \gamma_t^2 \mathbb{E}\Big[z_t^2 \big((\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big) \otimes \big((\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big)\Big] \\ &\quad + \gamma_t \mathbb{E}\Big[(\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{fluct},1} \otimes z_t (\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\Big] \\ &\quad + \gamma_t \mathbb{E}\Big[z_t (\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta} \otimes (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{fluct},1}\Big]. \end{aligned}$$

Define

$$\mathbf{\Sigma}_t^{\text{fluct},1} := \mathbb{E}\Big[z_t^2\big((\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big) \otimes \big((\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big)\Big]. \tag{17}$$

Then

$$\mathbf{F}_t^{(1)} = (\mathcal{I} - \gamma_t\mathcal{T}_t) \circ \mathbf{F}_{t-1}^{(1)} + \gamma_t^2\mathbf{\Sigma}_t^{\text{fluct},1} + \mathbf{C}_t^{(1)} + (\mathbf{C}_t^{(1)})^\top, \tag{18}$$

where

$$\mathbf{C}_t^{(1)} := \gamma_t\mathbb{E}\Big[(\mathbf{I} - \gamma_t\mathbf{x}_t\mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{fluct},1} \otimes z_t(\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\Big].$$

We now prove that $\mathbf{C}_t^{(1)} = \mathbf{0}$. Let

$$\mathcal{Z} := \sigma(z_1, \dots, z_N)$$

be the sigma-field generated by the entire source schedule. We first claim that

$$\mathbb{E}\big[\boldsymbol{\eta}_t^{\text{fluct},1} \mid \mathcal{Z}\big] = \mathbf{0}, \qquad \forall t \geq 0. \tag{19}$$

Indeed, the claim is trivial for $t = 0$. Suppose it holds at time $t - 1$. Using the recursion

$$\boldsymbol{\eta}_t^{\text{fluct},1} = (\mathbf{I} - \gamma_t\mathbf{x}_t\mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{fluct},1} + \gamma_t z_t(\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta},$$

and using the independence of $\mathbf{x}_t$ from $\mathcal{Z} \vee \sigma(\mathbf{x}_1, \dots, \mathbf{x}_{t-1})$, we have

$$\begin{aligned}
\mathbb{E}\big[\boldsymbol{\eta}_t^{\text{fluct},1} \mid \mathcal{Z}\big] &= \mathbb{E}\Big[\mathbb{E}\big[(\mathbf{I} - \gamma_t\mathbf{x}_t\mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{fluct},1} + \gamma_t z_t(\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta} \mid \mathcal{Z}, \mathbf{x}_1, \dots, \mathbf{x}_{t-1}\big] \mid \mathcal{Z}\Big] \\
&= \mathbb{E}\Big[(\mathbf{I} - \gamma_t\mathbf{H})\boldsymbol{\eta}_{t-1}^{\text{fluct},1} + \gamma_t z_t\mathbb{E}\big[(\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big] \mid \mathcal{Z}\Big] \\
&= (\mathbf{I} - \gamma_t\mathbf{H})\mathbb{E}\big[\boldsymbol{\eta}_{t-1}^{\text{fluct},1} \mid \mathcal{Z}\big] \\
&= \mathbf{0}.
\end{aligned}$$

This proves (19) by induction.

We now show that the cross term vanishes. Define, for any deterministic vector $\mathbf{u}$, the linear map

$$\mathcal{L}_t(\mathbf{u}) := \mathbb{E}_{\mathbf{x}_t}\Big[(\mathbf{I} - \gamma_t\mathbf{x}_t\mathbf{x}_t^\top)\mathbf{u} \otimes (\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\Big].$$

The map $\mathcal{L}_t(\cdot)$ is linear in $\mathbf{u}$. Since z_t is $\mathcal{Z}$-measurable, and since $\mathbf{x}_t$ is independent of $\mathcal{Z} \vee \sigma(\mathbf{x}_1, \dots, \mathbf{x}_{t-1})$, we obtain

$$\begin{aligned}
\mathbf{C}_t^{(1)} &= \gamma_t\mathbb{E}\Big[(\mathbf{I} - \gamma_t\mathbf{x}_t\mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{fluct},1} \otimes z_t(\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\Big] \\
&= \gamma_t\mathbb{E}\Big[\mathbb{E}\Big[(\mathbf{I} - \gamma_t\mathbf{x}_t\mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{fluct},1} \otimes z_t(\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta} \mid \mathcal{Z}, \mathbf{x}_1, \dots, \mathbf{x}_{t-1}\Big]\Big] \\
&= \gamma_t\mathbb{E}\Big[z_t\,\mathcal{L}_t\big(\boldsymbol{\eta}_{t-1}^{\text{fluct},1}\big)\Big] \\
&= \gamma_t\mathbb{E}\Big[\mathbb{E}\Big[z_t\,\mathcal{L}_t\big(\boldsymbol{\eta}_{t-1}^{\text{fluct},1}\big) \mid \mathcal{Z}\Big]\Big] \\
&= \gamma_t\mathbb{E}\Big[z_t\,\mathcal{L}_t\Big(\mathbb{E}[\boldsymbol{\eta}_{t-1}^{\text{fluct},1} \mid \mathcal{Z}]\Big)\Big] \\
&= \mathbf{0},
\end{aligned}$$

where the penultimate equality uses the linearity of $\mathcal{L}_t$, and the last equality follows from (19). Hence

$$\mathbf{C}_t^{(1)} = \mathbf{0}.$$

Taking transpose also gives

$$(\mathbf{C}_t^{(1)})^\top = \mathbf{0}.$$

Therefore (18) reduces to

$$\mathbf{F}_t^{(1)} = (\mathcal{I} - \gamma_t\mathcal{T}_t) \circ \mathbf{F}_{t-1}^{(1)} + \gamma_t^2\mathbf{\Sigma}_t^{\text{fluct},1}, \qquad \mathbf{F}_0^{(1)} = \mathbf{0}. \tag{20}$$

Since $z_t \in \{0,1\}$, we have $\mathbb{E}[z_t^2] = \mathbb{E}[z_t] = p$. Moreover,

$$\begin{aligned}\boldsymbol{\Sigma}_t^{\mathrm{fluct},1} &= \mathbb{E}\Big[z_t^2\big((\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big) \otimes \big((\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big)\Big] \\ &\preceq p\,\mathbb{E}\Big[\big((\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big) \otimes \big((\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big)\Big].\end{aligned}$$

Write $\mathbf{A} := \boldsymbol{\delta} \otimes \boldsymbol{\delta}$. Then

$$\begin{aligned}&\mathbb{E}\Big[\big((\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big) \otimes \big((\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big)\Big] \\ &\quad= \mathbb{E}\Big[(\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\mathbf{A}(\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\Big] \\ &\quad= \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top\mathbf{A}\mathbf{x}_t\mathbf{x}_t^\top] - \mathbf{H}\mathbf{A}\mathbf{H}.\end{aligned}$$

Since $\mathbf{H}\mathbf{A}\mathbf{H} \succeq 0$, it follows that

$$\boldsymbol{\Sigma}_t^{\mathrm{fluct},1} \preceq p\,\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top\mathbf{A}\mathbf{x}_t\mathbf{x}_t^\top]. \tag{21}$$

Now apply Assumption 2 with the PSD matrix $\mathbf{A} = \boldsymbol{\delta} \otimes \boldsymbol{\delta}$:

$$\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top\mathbf{A}\mathbf{x}_t\mathbf{x}_t^\top] \preceq \alpha\ \mathrm{tr}(\mathbf{H}\mathbf{A})\,\mathbf{H} = \alpha\,\|\boldsymbol{\delta}\|_{\mathbf{H}}^2\,\mathbf{H}. \tag{22}$$

Combining (21) and (22), we obtain

$$\boldsymbol{\Sigma}_t^{\mathrm{fluct},1} \preceq p\,\alpha\|\boldsymbol{\delta}\|_{\mathbf{H}}^2\,\mathbf{H}. \tag{23}$$

Comparing (20) with the variance recursion (5), we see that $\mathbf{F}_t^{(1)}$ satisfies exactly the same recursion form as the variance term, with noise covariance $\boldsymbol{\Sigma}_t^{\mathrm{fluct},1}$ in place of $\boldsymbol{\Sigma}_t$. By (23), we may apply Lemma 12 with

$$\sigma^2 \leftarrow p\alpha\|\boldsymbol{\delta}\|_{\mathbf{H}}^2.$$

Therefore,

$$\big\langle \mathbf{H}, \mathbf{F}_N^{(1)} \big\rangle \le \frac{8p\alpha\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1 - \gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}, \tag{24}$$

where

$$k^* = \max\left\{k : \lambda_k \ge \frac{1}{\gamma N_{\mathtt{eff}}}\right\}, \qquad D_{\mathtt{eff}} := k^* + \gamma^2 N_{\mathtt{eff}}^2 \sum_{i>k^*} \lambda_i^2.$$

Consequently, by (16),

$$\big\langle \mathbf{H}, \mathbf{F}_N \big\rangle \le 2\big\langle \mathbf{H}, \mathbf{F}_N^{(1)} \big\rangle + 2\big\langle \mathbf{H}, \mathbf{F}_N^{(2)} \big\rangle,$$

and the first term is controlled by (24). The second term $\big\langle \mathbf{H}, \mathbf{F}_N^{(2)} \big\rangle$ is be bounded by Lemma 16.

■

Lemma 15 (Fixed-budget variance identity) *Let $m \in \{0, 1, \ldots, N\}$, and set*

$$p := \frac{m}{N}.$$

Suppose that $(z_1, \ldots, z_N)$ is uniformly distributed over all binary vectors in $\{0,1\}^N$ with exactly m ones. Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N \in \mathbb{R}^{\mathcal{H}}$ be deterministic vectors, and define

$$\bar{\boldsymbol{v}} := \frac{1}{N}\sum_{s=1}^N \boldsymbol{v}_s.$$

Then

$$\mathbb{E}\left\|\sum_{s=1}^N (z_s - p)\boldsymbol{v}_s\right\|_{\mathbf{H}}^2 = \frac{Np(1-p)}{N-1}\sum_{s=1}^N \|\boldsymbol{v}_s - \bar{\boldsymbol{v}}\|_{\mathbf{H}}^2. \tag{25}$$

In particular,

$$\mathbb{E}\left\|\sum_{s=1}^N (z_s - p)\boldsymbol{v}_s\right\|_{\mathbf{H}}^2 \le \frac{Np(1-p)}{N-1}\sum_{s=1}^N \|\boldsymbol{v}_s\|_{\mathbf{H}}^2. \tag{26}$$

Proof. Since the law of $(z_1, \ldots, z_N)$ is exchangeable and exactly m coordinates are equal to 1, we have

$$\mathbb{E}[z_s] = \frac{m}{N} = p, \qquad s = 1, \ldots, N.$$

Moreover, because $z_s \in \{0, 1\}$,

$$\mathbb{E}[(z_s - p)^2] = \mathrm{Var}(z_s) = p(1-p).$$

For $s \neq r$, using the fixed-budget constraint $\sum_{t=1}^{N} z_t = m$, we also have

$$\mathbb{E}[z_s z_r] = \frac{m(m-1)}{N(N-1)}.$$

Therefore

$$\begin{aligned}
\mathbb{E}[(z_s - p)(z_r - p)] &= \mathbb{E}[z_s z_r] - p^2 \\
&= \frac{m(m-1)}{N(N-1)} - \frac{m^2}{N^2} \\
&= -\frac{m(N-m)}{N^2(N-1)} \\
&= -\frac{p(1-p)}{N-1}.
\end{aligned}$$

Now expand the quadratic form:

$$\begin{aligned}
\mathbb{E}\left\| \sum_{s=1}^{N} (z_s - p)\boldsymbol{v}_s \right\|_{\mathbf{H}}^2 &= \mathbb{E}\left[\sum_{s=1}^{N} \sum_{r=1}^{N} (z_s - p)(z_r - p) \langle \boldsymbol{v}_s, \boldsymbol{v}_r \rangle_{\mathbf{H}} \right] \\
&= \sum_{s=1}^{N} \mathbb{E}[(z_s - p)^2] \|\boldsymbol{v}_s\|_{\mathbf{H}}^2 + \sum_{s \neq r} \mathbb{E}[(z_s - p)(z_r - p)] \langle \boldsymbol{v}_s, \boldsymbol{v}_r \rangle_{\mathbf{H}} \\
&= p(1-p) \sum_{s=1}^{N} \|\boldsymbol{v}_s\|_{\mathbf{H}}^2 - \frac{p(1-p)}{N-1} \sum_{s \neq r} \langle \boldsymbol{v}_s, \boldsymbol{v}_r \rangle_{\mathbf{H}}.
\end{aligned}$$

Thus

$$\mathbb{E}\left\| \sum_{s=1}^{N} (z_s - p)\boldsymbol{v}_s \right\|_{\mathbf{H}}^2 = \frac{p(1-p)}{N-1} \left[(N-1) \sum_{s=1}^{N} \|\boldsymbol{v}_s\|_{\mathbf{H}}^2 - \sum_{s \neq r} \langle \boldsymbol{v}_s, \boldsymbol{v}_r \rangle_{\mathbf{H}} \right]. \tag{27}$$

On the other hand,

$$\begin{aligned}
\sum_{s=1}^{N} \|\boldsymbol{v}_s - \bar{\boldsymbol{v}}\|_{\mathbf{H}}^2 &= \sum_{s=1}^{N} \|\boldsymbol{v}_s\|_{\mathbf{H}}^2 - 2 \sum_{s=1}^{N} \langle \boldsymbol{v}_s, \bar{\boldsymbol{v}} \rangle_{\mathbf{H}} + \sum_{s=1}^{N} \|\bar{\boldsymbol{v}}\|_{\mathbf{H}}^2 \\
&= \sum_{s=1}^{N} \|\boldsymbol{v}_s\|_{\mathbf{H}}^2 - N \|\bar{\boldsymbol{v}}\|_{\mathbf{H}}^2.
\end{aligned}$$

Since

$$N^2 \|\bar{\boldsymbol{v}}\|_{\mathbf{H}}^2 = \left\| \sum_{s=1}^{N} \boldsymbol{v}_s \right\|_{\mathbf{H}}^2 = \sum_{s=1}^{N} \|\boldsymbol{v}_s\|_{\mathbf{H}}^2 + \sum_{s \neq r} \langle \boldsymbol{v}_s, \boldsymbol{v}_r \rangle_{\mathbf{H}},$$

it follows that

$$\begin{aligned}
N \sum_{s=1}^{N} \|\boldsymbol{v}_s - \bar{\boldsymbol{v}}\|_{\mathbf{H}}^2 &= N \sum_{s=1}^{N} \|\boldsymbol{v}_s\|_{\mathbf{H}}^2 - \left(\sum_{s=1}^{N} \|\boldsymbol{v}_s\|_{\mathbf{H}}^2 + \sum_{s \neq r} \langle \boldsymbol{v}_s, \boldsymbol{v}_r \rangle_{\mathbf{H}} \right) \\
&= (N-1) \sum_{s=1}^{N} \|\boldsymbol{v}_s\|_{\mathbf{H}}^2 - \sum_{s \neq r} \langle \boldsymbol{v}_s, \boldsymbol{v}_r \rangle_{\mathbf{H}}.
\end{aligned}$$

Substituting this identity into (27) yields

$$\mathbb{E}\left\|\sum_{s=1}^{N}(z_s-p)\boldsymbol{v}_s\right\|_{\mathbf{H}}^2=\frac{Np(1-p)}{N-1}\sum_{s=1}^{N}\|\boldsymbol{v}_s-\bar{\boldsymbol{v}}\|_{\mathbf{H}}^2,$$

which proves (25).

Finally, since

$$\sum_{s=1}^{N}\|\boldsymbol{v}_s-\bar{\boldsymbol{v}}\|_{\mathbf{H}}^2=\sum_{s=1}^{N}\|\boldsymbol{v}_s\|_{\mathbf{H}}^2-N\|\bar{\boldsymbol{v}}\|_{\mathbf{H}}^2\le\sum_{s=1}^{N}\|\boldsymbol{v}_s\|_{\mathbf{H}}^2,$$

we obtain (26). ∎

Lemma 16 (An upper bound for the second fluctuation term) *Recall that*

$$\boldsymbol{\eta}_t^{\mathrm{fluct},2}=(\mathbf{I}-\gamma_t\mathbf{x}_t\mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{fluct},2}+\gamma_t(z_t-p)\mathbf{H}\boldsymbol{\delta},\qquad\boldsymbol{\eta}_0^{\mathrm{fluct},2}=\mathbf{0},$$

and define

$$\mathbf{F}_t^{(2)}:=\mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{fluct},2}\otimes\boldsymbol{\eta}_t^{\mathrm{fluct},2}].$$

Suppose Assumption 2 holds and $\gamma<\frac{1}{\alpha\,\mathrm{tr}(\mathbf{H})}$. *Let* $N_{\mathtt{eff}}:=N/\log N$, *Then*

$$\left\langle\mathbf{H},\mathbf{F}_N^{(2)}\right\rangle\le\frac{8Np(1-p)}{(N-1)\big(1-\alpha\gamma\,\mathrm{tr}(\mathbf{H})\big)}\|\boldsymbol{\delta}\|_{\mathbf{H}}^2\cdot\frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}},$$

where $k^*=\max\{k:\lambda_k\ge\frac{1}{\gamma N_{\mathtt{eff}}}\}$ *and* $D_{\mathtt{eff}}:=k^*+\gamma^2N_{\mathtt{eff}}^2\sum_{i>k^*}\lambda_i^2$.

Proof. For each $t=1,\dots,N$, define

$$\mathbf{B}_t:=\mathbf{I}-\gamma_t\mathbf{x}_t\mathbf{x}_t^\top.$$

Iterating the recursion for $\boldsymbol{\eta}_t^{\mathrm{fluct},2}$ yields

$$\boldsymbol{\eta}_N^{\mathrm{fluct},2}=\sum_{s=1}^{N}\mathbf{B}_N\mathbf{B}_{N-1}\cdots\mathbf{B}_{s+1}\,\gamma_s(z_s-p)\mathbf{H}\boldsymbol{\delta},$$

where the empty product is interpreted as $\mathbf{I}$. Therefore

$$\boldsymbol{\eta}_N^{\mathrm{fluct},2}=\sum_{s=1}^{N}(z_s-p)\boldsymbol{v}_s,\qquad\boldsymbol{v}_s:=\gamma_s\mathbf{B}_N\mathbf{B}_{N-1}\cdots\mathbf{B}_{s+1}\mathbf{H}\boldsymbol{\delta}.$$

Conditional on the feature sequence $(\mathbf{x}_1,\dots,\mathbf{x}_N)$, the vectors $\boldsymbol{v}_1,\dots,\boldsymbol{v}_N$ are deterministic. Since the source indicators are independent of the features and follow the fixed-budget law, Lemma 15 gives

$$\mathbb{E}\Big[\|\boldsymbol{\eta}_N^{\mathrm{fluct},2}\|_{\mathbf{H}}^2\,\Big|\,\mathbf{x}_1,\dots,\mathbf{x}_N\Big]\le\frac{Np(1-p)}{N-1}\sum_{s=1}^{N}\|\boldsymbol{v}_s\|_{\mathbf{H}}^2.$$

Taking expectation over the features gives

$$\left\langle\mathbf{H},\mathbf{F}_N^{(2)}\right\rangle=\mathbb{E}\Big[\|\boldsymbol{\eta}_N^{\mathrm{fluct},2}\|_{\mathbf{H}}^2\Big]\le\frac{Np(1-p)}{N-1}\sum_{s=1}^{N}\mathbb{E}\big[\|\boldsymbol{v}_s\|_{\mathbf{H}}^2\big]\,.\tag{28}$$

For each $t=0,1,\dots,N$ and each $s\in\{1,\dots,N\}$, define

$$\boldsymbol{v}_{s,t}:=\begin{cases}\gamma_s\mathbf{B}_t\mathbf{B}_{t-1}\cdots\mathbf{B}_{s+1}\mathbf{H}\boldsymbol{\delta}, & 1\le s\le t,\\ \mathbf{0}, & s>t.\end{cases}$$

In particular, $\boldsymbol{v}_{s,N}=\boldsymbol{v}_s$. Now define

$$\mathbf{S}_t:=\sum_{s=1}^{t}\boldsymbol{v}_{s,t}\boldsymbol{v}_{s,t}^\top,\qquad t=0,1,\dots,N,$$

with the convention $\mathbf{S}_0 := \mathbf{0}$. Then

$$\langle \mathbf{H}, \mathbf{S}_N \rangle = \sum_{s=1}^{N} \langle \mathbf{H}, \boldsymbol{v}_s \boldsymbol{v}_s^\top \rangle = \sum_{s=1}^{N} \|\boldsymbol{v}_s\|_{\mathbf{H}}^2.$$

Moreover, for $s < t$,

$$\boldsymbol{v}_{s,t} = \mathbf{B}_t \boldsymbol{v}_{s,t-1}, \qquad \text{while} \qquad \boldsymbol{v}_{t,t} = \gamma_t \mathbf{H}\boldsymbol{\delta}.$$

Hence

$$\mathbf{S}_t = \mathbf{B}_t \mathbf{S}_{t-1} \mathbf{B}_t^\top + \gamma_t^2 \mathbf{O}, \qquad \mathbf{O} := \mathbf{H}\boldsymbol{\delta}\boldsymbol{\delta}^\top \mathbf{H}.$$

Define $\bar{\mathbf{S}}_t := \mathbb{E}[\mathbf{S}_t]$. Since $\mathbf{S}_{t-1}$ is measurable with respect to $\sigma(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1})$ and $\mathbf{x}_t$ is independent of the past, taking conditional expectation yields

$$\bar{\mathbf{S}}_t = (\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \bar{\mathbf{S}}_{t-1} + \gamma_t^2 \mathbf{O}, \qquad \bar{\mathbf{S}}_0 = \mathbf{0}.$$

We first bound the forcing matrix $\mathbf{O}$ by a multiple of $\mathbf{H}$. For any vector $\mathbf{u}$,

$$\mathbf{u}^\top \mathbf{O} \mathbf{u} = (\boldsymbol{\delta}^\top \mathbf{H} \mathbf{u})^2 \le (\boldsymbol{\delta}^\top \mathbf{H} \boldsymbol{\delta})(\mathbf{u}^\top \mathbf{H} \mathbf{u}) = \|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \, \mathbf{u}^\top \mathbf{H} \mathbf{u},$$

where we used Cauchy–Schwarz in the inner product induced by $\mathbf{H}$. Therefore

$$\mathbf{O} \preceq \|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \, \mathbf{H}. \tag{29}$$

Now define an auxiliary matrix sequence $(\mathbf{C}_t^\delta)_{t=0}^N$ by

$$\mathbf{C}_0^\delta := \mathbf{0}, \qquad \mathbf{C}_t^\delta = (\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \mathbf{C}_{t-1}^\delta + \gamma_t^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \, \mathbf{H}, \qquad t = 1, \ldots, N.$$

By Lemma 10, the operator $(\mathcal{I} - \gamma_t \mathcal{T}_t)$ is a PSD mapping. Since $\bar{\mathbf{S}}_0 = \mathbf{C}_0^\delta = \mathbf{0}$ and (29) holds, an induction on t shows that

$$\bar{\mathbf{S}}_t \preceq \mathbf{C}_t^\delta, \qquad t = 0, 1, \ldots, N.$$

Consequently,

$$\sum_{s=1}^{N} \mathbb{E}\big[\|\boldsymbol{v}_s\|_{\mathbf{H}}^2\big] = \langle \mathbf{H}, \bar{\mathbf{S}}_N \rangle \le \langle \mathbf{H}, \mathbf{C}_N^\delta \rangle. \tag{30}$$

The recursion defining $\mathbf{C}_t^\delta$ has exactly the same form as the variance recursion (5), with noise covariance $\boldsymbol{\Sigma}^\delta := \|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \, \mathbf{H}$. Applying Lemma 12 with

$$\sigma^2 \leftarrow \|\boldsymbol{\delta}\|_{\mathbf{H}}^2$$

yields

$$\langle \mathbf{H}, \mathbf{C}_N^\delta \rangle \le \frac{8\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1 - \alpha\gamma \operatorname{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}.$$

Combining this with (28) and (30), we obtain

$$\big\langle \mathbf{H}, \mathbf{F}_N^{(2)} \big\rangle \le \frac{Np(1-p)}{N-1} \cdot \frac{8\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1 - \alpha\gamma \operatorname{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}},$$

■

C.6 Proof of Theorem 1

Proof. Combining Bias upper bound, Variance upper bound, Drift upper bound and Fluctuation upper bound yields the desired result. ■

D Mixed Training Lower Bound Analysis

In this section, we derive a lower bound for the mixed-training excess risk. The starting point is again the error recursion

$$\boldsymbol{\eta}_t = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1} + \gamma_t z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} + \gamma_t \xi_t \mathbf{x}_t, \qquad \boldsymbol{\eta}_0 = \mathbf{w}_0 - \mathbf{w}_1^*. \tag{31}$$

Recall that

$$\boldsymbol{\eta}_t = \mathbf{w}_t - \mathbf{w}_1^*.$$

By definition of excess risk,

$$\mathbb{E}[\mathcal{R}(\mathbf{w}_N) - \mathcal{R}(\mathbf{w}_1^*)] = \frac{1}{2}\left\langle \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_N \otimes \boldsymbol{\eta}_N] \right\rangle.$$

To derive a lower bound, we directly center the total error process $\boldsymbol{\eta}_t$ around its mean. Define

$$\bar{\boldsymbol{\eta}}_t := \mathbb{E}[\boldsymbol{\eta}_t], \qquad \widetilde{\boldsymbol{\eta}}_t := \boldsymbol{\eta}_t - \bar{\boldsymbol{\eta}}_t.$$

Then

$$\boldsymbol{\eta}_t = \bar{\boldsymbol{\eta}}_t + \widetilde{\boldsymbol{\eta}}_t, \qquad \mathbb{E}[\widetilde{\boldsymbol{\eta}}_t] = \mathbf{0}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\eta}_N \otimes \boldsymbol{\eta}_N] &= \mathbb{E}\big[(\bar{\boldsymbol{\eta}}_N + \widetilde{\boldsymbol{\eta}}_N) \otimes (\bar{\boldsymbol{\eta}}_N + \widetilde{\boldsymbol{\eta}}_N)\big] \\
&= \bar{\boldsymbol{\eta}}_N \otimes \bar{\boldsymbol{\eta}}_N + \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N \otimes \widetilde{\boldsymbol{\eta}}_N] + \bar{\boldsymbol{\eta}}_N \otimes \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N] + \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N] \otimes \bar{\boldsymbol{\eta}}_N \\
&= \bar{\boldsymbol{\eta}}_N \otimes \bar{\boldsymbol{\eta}}_N + \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N \otimes \widetilde{\boldsymbol{\eta}}_N].
\end{aligned}$$

Hence

$$\mathbb{E}[\mathcal{R}(\mathbf{w}_N) - \mathcal{R}(\mathbf{w}_1^*)] = \frac{1}{2}\|\bar{\boldsymbol{\eta}}_N\|_{\mathbf{H}}^2 + \frac{1}{2}\big\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N \otimes \widetilde{\boldsymbol{\eta}}_N] \big\rangle. \tag{32}$$

D.1 The mean recursion

Taking expectation on both sides of (31), and using that

$$\mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top] = \mathbf{H}, \qquad \mathbb{E}[z_t] = p, \qquad \mathbb{E}[\xi_t \mathbf{x}_t] = \mathbf{0},$$

we obtain

$$\bar{\boldsymbol{\eta}}_t = (\mathbf{I} - \gamma_t \mathbf{H})\bar{\boldsymbol{\eta}}_{t-1} + \gamma_t p \mathbf{H}\boldsymbol{\delta}, \qquad \bar{\boldsymbol{\eta}}_0 = \mathbf{w}_0 - \mathbf{w}_1^*.$$

Unrolling the recursion gives

$$\bar{\boldsymbol{\eta}}_N = \prod_{t=1}^{N}(\mathbf{I} - \gamma_t \mathbf{H})(\mathbf{w}_0 - \mathbf{w}_1^*) + p\sum_{s=1}^{N}\gamma_s \left(\prod_{t=s+1}^{N}(\mathbf{I} - \gamma_t \mathbf{H})\right)\mathbf{H}\boldsymbol{\delta}. \tag{33}$$

Since each factor $(\mathbf{I} - \gamma_t \mathbf{H})$ is a polynomial in $\mathbf{H}$, it commutes with $\mathbf{H}$. Therefore,

$$\sum_{s=1}^{N}\gamma_s \left(\prod_{t=s+1}^{N}(\mathbf{I} - \gamma_t \mathbf{H})\right)\mathbf{H} = \mathbf{I} - \prod_{t=1}^{N}(\mathbf{I} - \gamma_t \mathbf{H}).$$

Substituting this identity into (33), we get the closed form

$$\bar{\boldsymbol{\eta}}_N = \prod_{t=1}^{N}(\mathbf{I} - \gamma_t \mathbf{H})(\mathbf{w}_0 - \mathbf{w}_1^*) + p\Big(\mathbf{I} - \prod_{t=1}^{N}(\mathbf{I} - \gamma_t \mathbf{H})\Big)\boldsymbol{\delta}.$$

Consequently,

$$\|\bar{\boldsymbol{\eta}}_N\|_{\mathbf{H}}^2 = \Big\| \prod_{t=1}^{N}(\mathbf{I} - \gamma_t \mathbf{H})(\mathbf{w}_0 - \mathbf{w}_1^*) + p\big(\mathbf{I} - \prod_{t=1}^{N}(\mathbf{I} - \gamma_t \mathbf{H})\big)\boldsymbol{\delta} \Big\|_{\mathbf{H}}^2. \tag{34}$$

The remaining task is to lower bound the centered fluctuation term

$$\frac{1}{2}\big\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N \otimes \widetilde{\boldsymbol{\eta}}_N] \big\rangle,$$

which will be handled next.

D.2 The centered fluctuation term

First, we introduce the lower bound Theorem of $\langle \mathbf{H}, \mathbf{V}_N \rangle$ in [67], which will be used in our analysis.

Lemma 17 (A variance lower bound) *Suppose Assumptions 1 and 3 hold. Let $N_{\texttt{eff}} = N/\log N$. Suppose $N_{\texttt{eff}} \geq 10$ and $\gamma < 1/\lambda_1$. We have*

$$\langle \mathbf{H}, \mathbf{V}_N \rangle \geq \frac{\sigma_1^2}{400} \cdot \frac{D_{\texttt{eff}}}{N_{\texttt{eff}}},$$

where $k^* = \max\left\{k : \lambda_k \geq \frac{1}{\gamma N_{\texttt{eff}}}\right\}$, $\quad D_{\texttt{eff}} := k^* + \gamma^2 N_{\texttt{eff}}^2 \sum_{i>k^*} \lambda_i^2$.

Proof. See proof of Theorem C.2 in [67]. ∎

We now continue the analysis of the centered fluctuation term

$$\frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N \otimes \widetilde{\boldsymbol{\eta}}_N] \rangle.$$

Starting from the total recursion

$$\boldsymbol{\eta}_t = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1} + \gamma_t z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} + \gamma_t \xi_t \mathbf{x}_t, \qquad \boldsymbol{\eta}_0 = \mathbf{w}_0 - \mathbf{w}_1^*,$$

and subtracting the mean recursion

$$\bar{\boldsymbol{\eta}}_t = (\mathbf{I} - \gamma_t \mathbf{H})\bar{\boldsymbol{\eta}}_{t-1} + \gamma_t p \mathbf{H}\boldsymbol{\delta}, \qquad \bar{\boldsymbol{\eta}}_0 = \mathbf{w}_0 - \mathbf{w}_1^*,$$

we obtain

$$\widetilde{\boldsymbol{\eta}}_t = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\widetilde{\boldsymbol{\eta}}_{t-1} + \gamma_t \xi_t \mathbf{x}_t + \gamma_t\big(z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} - p\mathbf{H}\boldsymbol{\delta}\big) + \gamma_t(\mathbf{H} - \mathbf{x}_t \mathbf{x}_t^\top)\bar{\boldsymbol{\eta}}_{t-1}. \tag{35}$$

We further decompose

$$\widetilde{\boldsymbol{\eta}}_t = \widetilde{\boldsymbol{\eta}}_t^{\text{noise}} + \widetilde{\boldsymbol{\eta}}_t^{\text{res}}, \qquad t = 0, 1, \ldots, N,$$

where

$$\begin{cases} \widetilde{\boldsymbol{\eta}}_t^{\text{noise}} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\widetilde{\boldsymbol{\eta}}_{t-1}^{\text{noise}} + \gamma_t \xi_t \mathbf{x}_t, \\ \widetilde{\boldsymbol{\eta}}_0^{\text{noise}} = \mathbf{0}, \end{cases} \quad \begin{cases} \widetilde{\boldsymbol{\eta}}_t^{\text{res}} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\widetilde{\boldsymbol{\eta}}_{t-1}^{\text{res}} + \gamma_t\big(z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} - p\mathbf{H}\boldsymbol{\delta}\big) + \gamma_t(\mathbf{H} - \mathbf{x}_t \mathbf{x}_t^\top)\bar{\boldsymbol{\eta}}_{t-1}, \\ \widetilde{\boldsymbol{\eta}}_0^{\text{res}} = \mathbf{0}. \end{cases}$$

Thus

$$\widetilde{\boldsymbol{\eta}}_N = \widetilde{\boldsymbol{\eta}}_N^{\text{noise}} + \widetilde{\boldsymbol{\eta}}_N^{\text{res}}.$$

Expanding the second moment gives

$$\begin{aligned} \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N \otimes \widetilde{\boldsymbol{\eta}}_N] = \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{noise}}] + \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{res}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{res}}] \\ + \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{res}}] + \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{res}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{noise}}]. \end{aligned} \tag{36}$$

We now prove that

$$\mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_t^{\text{res}}] = \mathbf{0}, \qquad \forall t \geq 0. \tag{37}$$

Let

$$\mathbf{B}_t := \mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top, \qquad \mathcal{G} := \sigma\big((\mathbf{x}_s, z_s) : 1 \leq s \leq N\big)$$

be the sigma-field generated by the entire feature sequence and source schedule.

We first observe that

$$\widetilde{\boldsymbol{\eta}}_t^{\text{res}} \quad \text{is } \mathcal{G}\text{-measurable for every } t.$$

Indeed, $\widetilde{\boldsymbol{\eta}}_0^{\text{res}} = \mathbf{0}$, and the recursion

$$\widetilde{\boldsymbol{\eta}}_t^{\text{res}} = \mathbf{B}_t \widetilde{\boldsymbol{\eta}}_{t-1}^{\text{res}} + \gamma_t\big(z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} - p\mathbf{H}\boldsymbol{\delta}\big) + \gamma_t(\mathbf{H} - \mathbf{x}_t \mathbf{x}_t^\top)\bar{\boldsymbol{\eta}}_{t-1}$$

only involves $(\mathbf{x}_s, z_s)_{s \leq t}$ and the deterministic vector $\bar{\boldsymbol{\eta}}_{t-1}$. Hence, by induction, $\widetilde{\boldsymbol{\eta}}_t^{\text{res}}$ is $\mathcal{G}$-measurable.

Next, we show that

$$\mathbb{E}\big[\widetilde{\boldsymbol{\eta}}_t^{\text{noise}} \mid \mathcal{G}\big] = \mathbf{0}, \qquad \forall t \geq 0. \tag{38}$$

To see this, unroll the noise recursion

$$\widetilde{\boldsymbol{\eta}}_t^{\text{noise}} = \mathbf{B}_t \widetilde{\boldsymbol{\eta}}_{t-1}^{\text{noise}} + \gamma_t \xi_t \mathbf{x}_t, \qquad \widetilde{\boldsymbol{\eta}}_0^{\text{noise}} = \mathbf{0}.$$

This gives

$$\widetilde{\boldsymbol{\eta}}_t^{\text{noise}} = \sum_{s=1}^{t} \Big(\mathbf{B}_t \mathbf{B}_{t-1} \cdots \mathbf{B}_{s+1}\Big) \gamma_s \xi_s \mathbf{x}_s,$$

where the empty product is understood as the identity matrix. Conditional on $\mathcal{G}$, all matrices $\mathbf{B}_j$ and vectors $\mathbf{x}_s$ are deterministic. Moreover, by the zero-mean noise assumption,

$$\mathbb{E}[\xi_s \mid \mathcal{G}] = 0, \qquad s = 1, \ldots, N.$$

Therefore,

$$\mathbb{E}\big[\widetilde{\boldsymbol{\eta}}_t^{\text{noise}} \mid \mathcal{G}\big] = \sum_{s=1}^{t} \Big(\mathbf{B}_t \mathbf{B}_{t-1} \cdots \mathbf{B}_{s+1}\Big) \gamma_s \mathbf{x}_s \mathbb{E}[\xi_s \mid \mathcal{G}] = \mathbf{0},$$

which proves (38).

Now, since $\widetilde{\boldsymbol{\eta}}_t^{\text{res}}$ is $\mathcal{G}$-measurable, we have

$$\begin{aligned}
\mathbb{E}\big[\widetilde{\boldsymbol{\eta}}_t^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_t^{\text{res}}\big] &= \mathbb{E}\Big[\mathbb{E}\big[\widetilde{\boldsymbol{\eta}}_t^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_t^{\text{res}} \mid \mathcal{G}\big]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[\widetilde{\boldsymbol{\eta}}_t^{\text{noise}} \mid \mathcal{G}\big] \otimes \widetilde{\boldsymbol{\eta}}_t^{\text{res}}\Big] \\
&= \mathbf{0}.
\end{aligned}$$

This proves (37). Taking transpose also gives

$$\mathbb{E}\big[\widetilde{\boldsymbol{\eta}}_t^{\text{res}} \otimes \widetilde{\boldsymbol{\eta}}_t^{\text{noise}}\big] = \mathbf{0}.$$

Consequently,

$$\mathbb{E}[\widetilde{\boldsymbol{\eta}}_N \otimes \widetilde{\boldsymbol{\eta}}_N] = \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{noise}}] + \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{res}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{res}}].$$

Hence

$$\big\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N \otimes \widetilde{\boldsymbol{\eta}}_N] \big\rangle = \big\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{noise}}] \big\rangle + \big\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{res}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{res}}] \big\rangle. \tag{39}$$

We now identify the first term with a variance-type recursion. Define

$$\mathbf{V}_t^{\text{mix}} := \mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_t^{\text{noise}}].$$

Then $\mathbf{V}_t^{\text{mix}}$ satisfies

$$\mathbf{V}_t^{\text{mix}} = (\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \mathbf{V}_{t-1}^{\text{mix}} + \gamma_t^2 \bar{\boldsymbol{\Sigma}}, \qquad \mathbf{V}_0^{\text{mix}} = \mathbf{0},$$

where

$$\bar{\boldsymbol{\Sigma}} := \mathbb{E}[\xi_t^2 \mathbf{x}_t \mathbf{x}_t^\top] = (1-p)\boldsymbol{\Sigma}_1 + p\boldsymbol{\Sigma}_2.$$

Under Assumption 3, we have

$$\boldsymbol{\Sigma}_1 = \sigma_1^2 \mathbf{H}, \qquad \boldsymbol{\Sigma}_2 = \sigma_2^2 \mathbf{H},$$

and therefore

$$\bar{\boldsymbol{\Sigma}} = \big((1-p)\sigma_1^2 + p\sigma_2^2\big)\mathbf{H}.$$

Denote

$$\bar{\sigma}^2 := (1-p)\sigma_1^2 + p\sigma_2^2.$$

Then $\mathbf{V}_t^{\text{mix}}$ is exactly the same variance recursion as in Lemma 17, with σ_1^2 replaced by $\bar{\sigma}^2$. Hence

$$\big\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{noise}}] \big\rangle = \big\langle \mathbf{H}, \mathbf{V}_N^{\text{mix}} \big\rangle \geq \frac{\bar{\sigma}^2}{400} \cdot \frac{D_{\texttt{eff}}}{N_{\texttt{eff}}}. \tag{40}$$

The remaining term

$$\big\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{res}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{res}}] \big\rangle$$

will be lower bounded separately by the following Lemma.

Lemma 18 (A lower bound for the residual fluctuation term) *Recall that*

$$\widetilde{\boldsymbol{\eta}}_t^{\mathrm{res}} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\widetilde{\boldsymbol{\eta}}_{t-1}^{\mathrm{res}} + \gamma_t\big(z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} - p\mathbf{H}\boldsymbol{\delta}\big) + \gamma_t(\mathbf{H} - \mathbf{x}_t \mathbf{x}_t^\top)\bar{\boldsymbol{\eta}}_{t-1}, \qquad \widetilde{\boldsymbol{\eta}}_0^{\mathrm{res}} = \mathbf{0},$$

Suppose Assumptions 1, 2 and 3 hold. Let $N_{\mathtt{eff}} := N/\log N$, *Assume* $N \geq 500$ *and* $\gamma < 1/\lambda_1$. *Then*

$$\big\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}}] \big\rangle \geq \frac{\beta e^{-8}}{400} \, \|\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}\|_{\mathbf{H}_{k^*:\infty}}^2 \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}} + \frac{(e^{-4} - 2^{-7})^2}{40} \, p(1-p)\gamma^2 N_{\mathtt{eff}} \sum_{i > k^*} \lambda_i^3 \delta_i^2,$$

where $k^* = \max\left\{k : \lambda_k \geq \frac{1}{\gamma N_{\mathtt{eff}}}\right\}$, $\qquad D_{\mathtt{eff}} := k^* + \gamma^2 N_{\mathtt{eff}}^2 \sum_{i > k^*} \lambda_i^2$.

Proof. Let$\mathbf{c}_{t-1} = p\boldsymbol{\delta} - \bar{\boldsymbol{\eta}}_{t-1}$, we can rewrite the driving term in the recursion of $\widetilde{\boldsymbol{\eta}}_t^{\mathrm{res}}$ as

$$z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} - p\mathbf{H}\boldsymbol{\delta} + (\mathbf{H} - \mathbf{x}_t \mathbf{x}_t^\top)\bar{\boldsymbol{\eta}}_{t-1} = (z_t - p)\mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} + (\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\mathbf{c}_{t-1}.$$

Hence

$$\widetilde{\boldsymbol{\eta}}_t^{\mathrm{res}} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\widetilde{\boldsymbol{\eta}}_{t-1}^{\mathrm{res}} + \gamma_t (z_t - p)\mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} + \gamma_t (\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\mathbf{c}_{t-1}, \qquad \widetilde{\boldsymbol{\eta}}_0^{\mathrm{res}} = \mathbf{0}. \tag{41}$$

On the other hand, from the closed form of the mean process,

$$\bar{\boldsymbol{\eta}}_t = (\mathbf{I} - \gamma_t \mathbf{H})\bar{\boldsymbol{\eta}}_{t-1} + \gamma_t p\mathbf{H}\boldsymbol{\delta},$$

we obtain

$$\mathbf{c}_t = p\boldsymbol{\delta} - \bar{\boldsymbol{\eta}}_t = (\mathbf{I} - \gamma_t \mathbf{H})(p\boldsymbol{\delta} - \bar{\boldsymbol{\eta}}_{t-1}) = (\mathbf{I} - \gamma_t \mathbf{H})\mathbf{c}_{t-1},$$

In particular,

$$\mathbf{c}_0 = p\boldsymbol{\delta} - (\mathbf{w}_0 - \mathbf{w}_1^*) = -(\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}).$$

Let

$$\mathcal{Z} := \sigma(z_1, \ldots, z_N), \qquad \mathcal{X} := \sigma(\mathbf{x}_1, \ldots, \mathbf{x}_N),$$

and define

$$\mathbf{S}_Z := \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}} \mid \mathcal{Z}], \qquad \mathbf{S}_X := \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}} \mid \mathcal{X}], \qquad \mathbf{R}^\circ := \widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}} - \mathbf{S}_Z - \mathbf{S}_X. \tag{42}$$

We first show that

$$\mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}}] = \mathbf{0}. \tag{43}$$

Indeed, taking expectation in (41) and using

$$\mathbb{E}[z_t - p] = 0, \qquad \mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H}] = \mathbf{0}, \qquad \mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top] = \mathbf{H},$$

gives

$$\mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\mathrm{res}}] = (\mathbf{I} - \gamma_t \mathbf{H})\mathbb{E}[\widetilde{\boldsymbol{\eta}}_{t-1}^{\mathrm{res}}],$$

and since $\widetilde{\boldsymbol{\eta}}_0^{\mathrm{res}} = \mathbf{0}$, (43) follows.

Next, $\mathcal{Z}$ and $\mathcal{X}$ are independent, and $\mathbb{E}[\mathbf{S}_X] = \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}}] = 0$, hence

$$\mathbb{E}[\mathbf{S}_X \mid \mathcal{Z}] = \mathbb{E}[\mathbf{S}_X] = 0.$$

Therefore,

$$\mathbb{E}\langle \mathbf{S}_Z, \mathbf{H}\mathbf{S}_X \rangle = \mathbb{E}\Big[\langle \mathbf{S}_Z, \mathbf{H}\,\mathbb{E}[\mathbf{S}_X \mid \mathcal{Z}] \rangle\Big] = 0.$$

Moreover, by (42),

$$\mathbb{E}[\mathbf{R}^\circ \mid \mathcal{Z}] = \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}} \mid \mathcal{Z}] - \mathbf{S}_Z - \mathbb{E}[\mathbf{S}_X \mid \mathcal{Z}] = 0,$$

and similarly $\mathbb{E}[\mathbf{R}^\circ \mid \mathcal{X}] = 0$. Thus

$$\mathbb{E}\langle \mathbf{S}_Z, \mathbf{H}\mathbf{R}^\circ \rangle = 0, \qquad \mathbb{E}\langle \mathbf{S}_X, \mathbf{H}\mathbf{R}^\circ \rangle = 0.$$

Consequently,

$$\mathbb{E}\|\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}}\|_{\mathbf{H}}^2 = \mathbb{E}\|\mathbf{S}_Z\|_{\mathbf{H}}^2 + \mathbb{E}\|\mathbf{S}_X\|_{\mathbf{H}}^2 + \mathbb{E}\|\mathbf{R}^\circ\|_{\mathbf{H}}^2 \geq \mathbb{E}\|\mathbf{S}_Z\|_{\mathbf{H}}^2 + \mathbb{E}\|\mathbf{S}_X\|_{\mathbf{H}}^2. \tag{44}$$

Since $\big\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}}] \big\rangle = \mathbb{E}\|\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}}\|_{\mathbf{H}}^2$, it remains to lower bound the two terms on the right-hand side of (44).

Define $\mathbf{f}_t := \mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\mathrm{res}} \mid \mathcal{X}]$.Taking conditional expectation in (41) and using $\mathbb{E}[z_t - p \mid \mathcal{X}] = 0$, we obtain

$$\mathbf{f}_t = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{f}_{t-1} + \gamma_t(\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\mathbf{c}_{t-1}, \qquad \mathbf{f}_0 = \mathbf{0}. \tag{45}$$

Thus $\mathbf{S}_X = \mathbf{f}_N$. Further define

$$\boldsymbol{\zeta}_u := (\mathbf{x}_u\mathbf{x}_u^\top - \mathbf{H})\mathbf{c}_{u-1}, \qquad \mathbf{K}_{u,N} := \gamma_u \prod_{j=u+1}^{N} (\mathbf{I} - \gamma_j \mathbf{x}_j\mathbf{x}_j^\top).$$

Unrolling (45) gives

$$\mathbf{S}_X = \sum_{u=1}^{N} \mathbf{U}_u, \qquad \mathbf{U}_u := \mathbf{K}_{u,N}\boldsymbol{\zeta}_u. \tag{46}$$

We first prove the cross terms vanish. Let $u < v$. Since $\mathbf{K}_{u,N}$ and $\mathbf{U}_v$ are measurable with respect to $\sigma(\mathbf{x}_{u+1}, \ldots, \mathbf{x}_N)$, while $\boldsymbol{\zeta}_u$ depends only on $\mathbf{x}_u$, we have

$$\mathbb{E}[\boldsymbol{\zeta}_u] = (\mathbb{E}[\mathbf{x}_u\mathbf{x}_u^\top] - \mathbf{H})\mathbf{c}_{u-1} = 0.$$

Hence

$$\mathbb{E}\big[\langle \mathbf{U}_u, \mathbf{H}\mathbf{U}_v\rangle \mid \mathbf{x}_{u+1}, \ldots, \mathbf{x}_N\big] = \mathbb{E}\big[\langle \boldsymbol{\zeta}_u, \mathbf{K}_{u,N}^\top \mathbf{H}\mathbf{U}_v\rangle \mid \mathbf{x}_{u+1}, \ldots, \mathbf{x}_N\big] = 0,$$

and therefore

$$\mathbb{E}\langle \mathbf{U}_u, \mathbf{H}\mathbf{U}_v\rangle = 0, \qquad u \neq v.$$

It follows that

$$\mathbb{E}\|\mathbf{S}_X\|_{\mathbf{H}}^2 = \sum_{u=1}^{N} \mathbb{E}\|\mathbf{U}_u\|_{\mathbf{H}}^2. \tag{47}$$

Now let

$$\mathbf{A}_u := \mathbf{c}_{u-1}\mathbf{c}_{u-1}^\top \succeq 0.$$

Using the independence of $\mathbf{K}_{u,N}$ and $\boldsymbol{\zeta}_u$, we have

$$\mathbb{E}\|\mathbf{U}_u\|_{\mathbf{H}}^2 = \mathbb{E}\big[\boldsymbol{\zeta}_u^\top \mathbf{K}_{u,N}^\top \mathbf{H}\mathbf{K}_{u,N}\boldsymbol{\zeta}_u\big] = \big\langle \mathbb{E}[\mathbf{K}_{u,N}^\top \mathbf{H}\mathbf{K}_{u,N}],\, \mathbf{G}_u \big\rangle, \tag{48}$$

where

$$\mathbf{G}_u := \mathbb{E}[\boldsymbol{\zeta}_u\boldsymbol{\zeta}_u^\top] = \mathbb{E}[(\mathbf{x}_u\mathbf{x}_u^\top - \mathbf{H})\mathbf{A}_u(\mathbf{x}_u\mathbf{x}_u^\top - \mathbf{H})].$$

By expansion,

$$\mathbf{G}_u = \mathbb{E}[\mathbf{x}_u\mathbf{x}_u^\top \mathbf{A}_u \mathbf{x}_u\mathbf{x}_u^\top] - \mathbf{H}\mathbf{A}_u\mathbf{H}.$$

Applying Assumption 2 with $A = \mathbf{A}_u$, we obtain

$$\mathbf{G}_u \succeq \beta\, \mathrm{tr}(\mathbf{H}\mathbf{A}_u)\, \mathbf{H} = \beta\, \|\mathbf{c}_{u-1}\|_{\mathbf{H}}^2\, \mathbf{H}.$$

Substituting this into (48) yields

$$\mathbb{E}\|\mathbf{U}_u\|_{\mathbf{H}}^2 \geq \beta\, \|\mathbf{c}_{u-1}\|_{\mathbf{H}}^2\, \mathbb{E}\big[\,\mathrm{tr}(\mathbf{H}\mathbf{K}_{u,N}\mathbf{H}\mathbf{K}_{u,N}^\top)\big]. \tag{49}$$

Set

$$\mathbf{M}_u := \mathbf{H}^{1/2}\mathbf{K}_{u,N}\mathbf{H}^{1/2}.$$

Then

$$\mathrm{tr}(\mathbf{H}\mathbf{K}_{u,N}\mathbf{H}\mathbf{K}_{u,N}^\top) = \|\mathbf{M}_u\|_F^2.$$

By Jensen's inequality,

$$\mathbb{E}\|\mathbf{M}_u\|_F^2 \geq \|\mathbb{E}[\mathbf{M}_u]\|_F^2.$$

Moreover,

$$\mathbb{E}[\mathbf{K}_{u,N}] = \gamma_u \prod_{j=u+1}^{N} (\mathbf{I} - \gamma_j\mathbf{H}).$$

Hence, in the eigenbasis of $\mathbf{H}$,

$$\|\mathbb{E}[\mathbf{M}_u]\|_F^2 = \sum_{i\geq 1} \gamma_u^2\lambda_i^2 \prod_{j=u+1}^{N} (1 - \gamma_j\lambda_i)^2.$$

Define

$$\kappa_u := \sum_{i\geq 1} \gamma_u^2 \lambda_i^2 \prod_{j=u+1}^{N} (1-\gamma_j\lambda_i)^2. \tag{50}$$

Then (49) gives

$$\mathbb{E}\|\mathbf{U}_u\|_{\mathbf{H}}^2 \geq \beta \, \|\mathbf{c}_{u-1}\|_{\mathbf{H}}^2 \, \kappa_u.$$

Combining this with (47),

$$\mathbb{E}\|\mathbf{S}_X\|_{\mathbf{H}}^2 \geq \beta \sum_{u=1}^{N} \|\mathbf{c}_{u-1}\|_{\mathbf{H}}^2 \, \kappa_u. \tag{51}$$

We now lower bound $\|\mathbf{c}_{u-1}\|_{\mathbf{H}}^2$ uniformly on the tail space. In the eigenbasis of $\mathbf{H}$,

$$(\mathbf{c}_{u-1})_i = -\prod_{j=1}^{u-1}(1-\gamma_j\lambda_i)\,(\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta})_i.$$

Hence

$$\|\mathbf{c}_{u-1}\|_{\mathbf{H}_{k^*:\infty}}^2 = \sum_{i>k^*} \lambda_i \prod_{j=1}^{u-1}(1-\gamma_j\lambda_i)^2 (\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta})_i^2. \tag{52}$$

Since $i > k^*$, we have $\gamma\lambda_i < 1/N_{\texttt{eff}} \leq 1/10$ for $N_{\texttt{eff}} \geq 10$. Also

$$\sum_{t=1}^{N} \gamma_t \leq 2\gamma N_{\texttt{eff}}.$$

Using $1 - a \geq e^{-2a}$ for $0 \leq a \leq 1/2$, we obtain

$$\prod_{j=1}^{u-1}(1-\gamma_j\lambda_i)^2 \geq \exp\left(-4\lambda_i \sum_{j=1}^{u-1}\gamma_j\right) \geq \exp(-8\gamma N_{\texttt{eff}}\lambda_i) \geq e^{-8}.$$

Substituting this into (52) gives

$$\|\mathbf{c}_{u-1}\|_{\mathbf{H}}^2 \geq \|\mathbf{c}_{u-1}\|_{\mathbf{H}_{k^*:\infty}}^2 \geq e^{-8}\|\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}\|_{\mathbf{H}_{k^*:\infty}}^2, \qquad u = 1, \ldots, N. \tag{53}$$

Therefore (51) yields

$$\mathbb{E}\|\mathbf{S}_X\|_{\mathbf{H}}^2 \geq \beta e^{-8}\|\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}\|_{\mathbf{H}_{k^*:\infty}}^2 \sum_{u=1}^{N} \kappa_u. \tag{54}$$

Finally, by applying Lemma 17,

$$\sum_{u=1}^{N} \kappa_u = \left\langle \mathbf{H},\, \sum_{u=1}^{N} \gamma_u^2 \prod_{j=u+1}^{N} (\mathbf{I}-\gamma_j\mathbf{H})^2\mathbf{H} \right\rangle \geq \frac{1}{400}\left(\frac{k^*}{N_{\texttt{eff}}} + \gamma^2 N_{\texttt{eff}} \sum_{i>k^*} \lambda_i^2\right).$$

Substituting this into (54), we conclude that

$$\mathbb{E}\|\mathbf{S}_X\|_{\mathbf{H}}^2 \geq \frac{\beta e^{-8}}{400} \, \|\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}\|_{\mathbf{H}_{k^*:\infty}}^2 \left(\frac{k^*}{N_{\texttt{eff}}} + \gamma^2 N_{\texttt{eff}} \sum_{i>k^*} \lambda_i^2\right). \tag{55}$$

Then we lower bound schedule component $\mathbf{S}_Z$. For each t, define $\mathbf{a}_t := \mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\text{res}} \mid \mathcal{Z}]$. Taking conditional expectation in (41) and using independence of $\mathbf{x}_t$ from $(\mathcal{Z}, \mathbf{x}_1, \ldots, \mathbf{x}_{t-1})$, we obtain

$$\mathbf{a}_t = (\mathbf{I} - \gamma_t\mathbf{H})\mathbf{a}_{t-1} + \gamma_t(z_t - p)\mathbf{H}\boldsymbol{\delta}, \qquad \mathbf{a}_0 = 0.$$

Thus $\mathbf{S}_Z = \mathbf{a}_N$, and unrolling gives

$$\mathbf{S}_Z = \sum_{u=1}^{N}(z_u - p)\mathbf{v}_u, \qquad \mathbf{v}_u := \gamma_u\Big(\prod_{j=u+1}^{N}(\mathbf{I}-\gamma_j\mathbf{H})\Big)\mathbf{H}\boldsymbol{\delta}. \tag{56}$$

In the eigenbasis of $\mathbf{H}$,

$$(\mathbf{v}_u)_i = \gamma_u \lambda_i \prod_{j=u+1}^{N} (1-\gamma_j\lambda_i)\delta_i.$$

By the fixed-budget variance identity,

$$\mathbb{E}\|\mathbf{S}_Z\|_{\mathbf{H}}^2 = \frac{p(1-p)}{2(N-1)} \sum_{u=1}^{N}\sum_{r=1}^{N} \|\mathbf{v}_u - \mathbf{v}_r\|_{\mathbf{H}}^2. \tag{57}$$

Let

$$I_- := \{t : \ell_t = 0\}, \qquad I_+ := \{t : \ell_t \geq 7\}.$$

For $N \geq 500$, one has

$$|I_-| \geq \frac{N_{\mathtt{eff}}}{2}, \qquad |I_+| \geq \frac{N}{10}.$$

Restricting (57) to pairs $(u, r) \in I_- \times I_+$ yields

$$\mathbb{E}\|\mathbf{S}_Z\|_{\mathbf{H}}^2 \geq \frac{p(1-p)}{2(N-1)} \sum_{u\in I_-}\sum_{r\in I_+} \|\mathbf{v}_u - \mathbf{v}_r\|_{\mathbf{H}}^2. \tag{58}$$

Fix $i > k^*$. Since $\gamma\lambda_i < 1/N_{\mathtt{eff}} \leq 1/10 < 1/2$, for $u \in I_-$ we have $\gamma_u = \gamma$, and

$$|(\mathbf{v}_u)_i| = \gamma\lambda_i \prod_{j=u+1}^{N} (1-\gamma_j\lambda_i)|\delta_i| \geq e^{-4}\gamma\lambda_i|\delta_i|.$$

For $r \in I_+$, we have $\gamma_r \leq 2^{-7}\gamma$, so

$$|(\mathbf{v}_r)_i| \leq 2^{-7}\gamma\lambda_i|\delta_i|.$$

Therefore

$$|(\mathbf{v}_u - \mathbf{v}_r)_i| \geq (e^{-4} - 2^{-7})\gamma\lambda_i|\delta_i|, \qquad u \in I_-, \quad r \in I_+, \quad i > k^*.$$

Hence

$$\|\mathbf{v}_u - \mathbf{v}_r\|_{\mathbf{H}}^2 \geq (e^{-4} - 2^{-7})^2\gamma^2 \sum_{i>k^*} \lambda_i^3\delta_i^2. \tag{59}$$

Substituting (59) into (58), and using $|I_-| \geq N_{\mathtt{eff}}/2$, $|I_+| \geq N/10$, we obtain

$$\mathbb{E}\|\mathbf{S}_Z\|_{\mathbf{H}}^2 \geq \frac{(e^{-4} - 2^{-7})^2}{40}\, p(1-p)\gamma^2 N_{\mathtt{eff}} \sum_{i>k^*} \lambda_i^3\delta_i^2.$$

That is,

$$\mathbb{E}\|\mathbf{S}_Z\|_{\mathbf{H}}^2 \geq \frac{(e^{-4} - 2^{-7})^2}{40}\, p(1-p)\gamma^2 N_{\mathtt{eff}} \sum_{i>k^*} \lambda_i^3\delta_i^2. \tag{60}$$

Finally, combining (44), (55), and (60), we obtain

$$\left\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}}] \right\rangle \geq \frac{\beta e^{-8}}{400}\, \|\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}\|_{\mathbf{H}_{k^*:\infty}}^2 \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}} + \frac{(e^{-4} - 2^{-7})^2}{40}\, p(1-p)\gamma^2 N_{\mathtt{eff}} \sum_{i>k^*} \lambda_i^3\delta_i^2,$$

This completes the proof. ∎

D.3 Proof of Theorem 2

Proof. Combining Lemma 18 with Eq. 40 yields the desired result. ∎

D.4 Proof of Corollary 3

Proof. Let

$$T := M + N, \qquad \mathbf{P}_T := \prod_{t=1}^{T} (\mathbf{I} - \gamma_t \mathbf{H}).$$

Since the synthetic proportion $p = M/(M+N)$ is fixed, it suffices to study the limit as $T \to \infty$.

We first show that $\mathbf{P}_T$ vanishes in $\mathbf{H}$-norm. Writing the eigendecomposition

$$\mathbf{H} = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i^\top,$$

we have

$$\mathbf{P}_T \mathbf{v}_i = \Big(\prod_{t=1}^{T} (1 - \gamma_t \lambda_i) \Big) \mathbf{v}_i.$$

Under the assumption on the stepsize, for every i with $\lambda_i > 0$,

$$0 \le 1 - \gamma_t \lambda_i < 1.$$

Moreover, under the geometric tail-decay schedule,

$$\sum_{t=1}^{T} \gamma_t \asymp \frac{T}{\log T} \sum_{\ell=0}^{\lfloor \log T \rfloor} 2^{-\ell} \asymp \frac{T}{\log T} \xrightarrow[T \to \infty]{} \infty.$$

Hence for every $\lambda_i > 0$,

$$\prod_{t=1}^{T} (1 - \gamma_t \lambda_i) \longrightarrow 0.$$

Therefore, for any $\mathbf{w}$ with $\|\mathbf{w}\|_{\mathbf{H}} < \infty$,

$$\|\mathbf{P}_T \mathbf{w}\|_{\mathbf{H}}^2 = \sum_i \lambda_i \Big(\prod_{t=1}^{T} (1 - \gamma_t \lambda_i) \Big)^2 \langle \mathbf{w}, \mathbf{v}_i \rangle^2 \longrightarrow 0$$

by dominated convergence.

We now turn to the upper bound in Theorem 1 which gives

$$\mathbb{E}[\mathcal{E}_1(\mathbf{w}_T)] \lesssim \underbrace{\|\mathbf{P}_T(\mathbf{w}_0 - \mathbf{w}_1^*)\|_{\mathbf{H}}^2}_{(I)} + \underbrace{\alpha \|\mathbf{w}_0 - \mathbf{w}_1^*\|^2_{\frac{\mathbf{I}_{0:k^*}}{\gamma \widetilde{N}_{\mathtt{eff}}} + \mathbf{H}_{k^*:\infty}} \frac{\widetilde{D}_{\mathtt{eff}}}{\widetilde{N}_{\mathtt{eff}}}}_{(II)}$$
$$+ \underbrace{\bar{\sigma}^2 \frac{\widetilde{D}_{\mathtt{eff}}}{\widetilde{N}_{\mathtt{eff}}}}_{(III)} + \underbrace{\alpha p \|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \frac{\widetilde{D}_{\mathtt{eff}}}{\widetilde{N}_{\mathtt{eff}}}}_{(IV)} + \underbrace{p^2 \|(\mathbf{I} - \mathbf{P}_T)\boldsymbol{\delta}\|_{\mathbf{H}}^2}_{(V)}.$$

By the assumption $\widetilde{D}_{\mathtt{eff}} = o(\widetilde{N}_{\mathtt{eff}})$, terms (II), (III) and (IV) all vanish as $T \to \infty$. By the argument above, $(I) \to 0$. Finally,

$$\|(\mathbf{I} - \mathbf{P}_T)\boldsymbol{\delta}\|_{\mathbf{H}} \to \|\boldsymbol{\delta}\|_{\mathbf{H}},$$

because $\|\mathbf{P}_T \boldsymbol{\delta}\|_{\mathbf{H}} \to 0$. Therefore,

$$\limsup_{T \to \infty} \mathbb{E}[\mathcal{E}_1(\mathbf{w}_T)] \lesssim p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2.$$

Next, apply the lower bound in Theorem 2:

$$\mathbb{E}[\mathcal{E}_1(\mathbf{w}_T)] \gtrsim \underbrace{\Big\| \mathbf{P}_T(\mathbf{w}_0 - \mathbf{w}_1^*) + p(\mathbf{I} - \mathbf{P}_T)\boldsymbol{\delta} \Big\|_{\mathbf{H}}^2}_{(A)}$$
$$+ \underbrace{\big(\beta \|\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}\|^2_{\mathbf{H}_{k^*:\infty}} + \bar{\sigma}^2 \big) \frac{\widetilde{D}_{\mathtt{eff}}}{\widetilde{N}_{\mathtt{eff}}}}_{(B)} + \underbrace{p(1-p)\gamma^2 \widetilde{N}_{\mathtt{eff}} \|\boldsymbol{\delta}\|^2_{\mathbf{H}^3_{k^*:\infty}}}_{(C)}.$$

Since $(B) \geq 0$ and $(C) \geq 0$, it is enough to keep only (A). Rewrite

$$\mathbf{P}_T(\mathbf{w}_0 - \mathbf{w}_1^*) + p(\mathbf{I} - \mathbf{P}_T)\boldsymbol{\delta} = \mathbf{P}_T(\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}) + p\boldsymbol{\delta}.$$

Again using $\|\mathbf{P}_T\mathbf{w}\|_{\mathbf{H}} \to 0$ for every $\mathbf{w}$ with finite $\mathbf{H}$-norm, we obtain

$$\Big\|\mathbf{P}_T(\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}) + p\boldsymbol{\delta}\Big\|_{\mathbf{H}}^2 \longrightarrow p^2\|\boldsymbol{\delta}\|_{\mathbf{H}}^2.$$

Hence,

$$\liminf_{T\to\infty} \mathbb{E}[\mathcal{E}_1(\mathbf{w}_T)] \gtrsim p^2\|\boldsymbol{\delta}\|_{\mathbf{H}}^2.$$

Combining the upper and lower bounds yields

$$\lim_{M+N\to\infty} \mathbb{E}[\mathcal{E}_1(\mathbf{w}_{M+N})] \eqsim p^2\|\boldsymbol{\delta}\|_{\mathbf{H}}^2.$$

This proves the corollary. ∎

E Two-stage Training Analysis

First, consider the upper bound for training on pure one-source data from [67], which can also be recovered by Theorem 1.

Theorem 19 (Learning with only real data) *Consider last iterate SGD with stepsize scheme* (1)*. Suppose Assumptions 1, 2 and 3 hold. Let* $N_{\texttt{eff}} := N/\log N$*. Suppose* $\gamma < 1/(4\alpha\,\mathrm{tr}(\mathbf{H}))$*. Then we have*

$$\mathbb{E}[\mathcal{R}(\mathbf{w}_N)] - \mathcal{R}(\mathbf{w}_1^*) \lesssim \Big\|(\prod_{t=1}^{N}\mathbf{I} - \gamma_t\mathbf{H})(\mathbf{w}_0^* - \mathbf{w}_1^*)\Big\|_{\mathbf{H}}^2 + \big(\alpha\|\mathbf{w}_0 - \mathbf{w}_1^*\|^2_{\frac{\mathbf{I}_{0:k^*}}{\gamma N_{\texttt{eff}}} + \mathbf{H}_{k^*:\infty}} + \sigma_1^2\big)\frac{D_{\texttt{eff}}}{N_{\texttt{eff}}}$$

with $k^* = \max\{k : \lambda_k \geq \frac{1}{\gamma N_{\texttt{eff}}}\}$ *and* $D_{\texttt{eff}} := k^* + \gamma^2 N_{\texttt{eff}}^2 \sum_{i>k^*}\lambda_i^2$*.*

We now turn to the two-stage training procedure. The key idea is to condition on the first-stage training iterate and then apply the pure real-data bound from Theorem 1 to the real-data training phase.

Suppose that in the first stage we run SGD on M synthetic samples and obtain $\mathbf{w}_M$. In the second stage, starting from $\mathbf{w}_M$, we run SGD on N real samples and output $\mathbf{w}_{M+N}$. Our goal is to upper bound

$$\mathbb{E}\big[\mathcal{R}(\mathbf{w}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*)\big].$$

Let $\mathcal{F}_M := \sigma(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_M, y_M)$ be the sigma-field generated by the first stage. Conditional on $\mathcal{F}_M$, the iterate $\mathbf{w}_M$ is deterministic, and the second-stage procedure is exactly SGD trained on N real samples, initialized at $\mathbf{w}_M$. Therefore, applying Theorem 1 conditionally, we obtain

$$\begin{aligned}\mathbb{E}\big[\mathcal{R}(\mathbf{w}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*) \mid \mathcal{F}_M\big] \lesssim{}& \Big\|\Big(\prod_{t=1}^{N}(\mathbf{I} - \gamma_t\mathbf{H})\Big)(\mathbf{w}_M - \mathbf{w}_1^*)\Big\|_{\mathbf{H}}^2 \\ &+ \left(\alpha\|\mathbf{w}_M - \mathbf{w}_1^*\|^2_{\frac{\mathbf{I}_{0:k^*}}{\gamma N_{\texttt{eff}}} + \mathbf{H}_{k^*:\infty}} + \sigma_1^2\right)\frac{D_{\texttt{eff}}}{N_{\texttt{eff}}}, \qquad (61)\end{aligned}$$

Taking expectation with respect to $\mathcal{F}_M$ on both sides of (61), we get

$$\begin{aligned}\mathbb{E}\big[\mathcal{R}(\mathbf{w}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*)\big] &= \mathbb{E}\Big[\mathbb{E}\big[\mathcal{R}(\mathbf{w}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*) \mid \mathcal{F}_M\big]\Big] \\ &\lesssim \mathbb{E}\left[\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I} - \gamma_t\mathbf{H})\Big)(\mathbf{w}_M - \mathbf{w}_1^*)\Big\|_{\mathbf{H}}^2\right] \\ &\quad + \left(\alpha\mathbb{E}\Big[\|\mathbf{w}_M - \mathbf{w}_1^*\|^2_{\frac{\mathbf{I}_{0:k^*}}{\gamma N_{\texttt{eff}}} + \mathbf{H}_{k^*:\infty}}\Big] + \sigma_1^2\right)\frac{D_{\texttt{eff}}}{N_{\texttt{eff}}}. \qquad (62)\end{aligned}$$

We now relate the terms to the stage-one excess risk under the synthetic distribution. Recall that

$$\mathbf{w}_M - \mathbf{w}_1^* = (\mathbf{w}_M - \mathbf{w}_2^*) + \boldsymbol{\delta}.$$

Hence

$$\begin{aligned}
&\mathbb{E}\left[\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{H})\Big)(\mathbf{w}_M-\mathbf{w}_1^*)\Big\|_{\mathbf{H}}^2\right] \\
&= \mathbb{E}\left[\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{H})\Big)(\mathbf{w}_M-\mathbf{w}_2^*) + \Big(\prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{H})\Big)\boldsymbol{\delta}\Big\|_{\mathbf{H}}^2\right] \\
&\le 2\mathbb{E}\left[\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{H})\Big)(\mathbf{w}_M-\mathbf{w}_2^*)\Big\|_{\mathbf{H}}^2\right] + 2\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{H})\Big)\boldsymbol{\delta}\Big\|_{\mathbf{H}}^2. \qquad (63)
\end{aligned}$$

Now denote

$$\mathbf{P}_N := \prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{H}).$$

Since $\mathbf{P}_N$ is a polynomial in $\mathbf{H}$, it commutes with $\mathbf{H}$. Therefore

$$\|\mathbf{P}_N\mathbf{v}\|_{\mathbf{H}}^2 = \langle \mathbf{v}, \mathbf{P}_N\mathbf{H}\mathbf{P}_N\mathbf{v}\rangle = \langle \mathbf{v}, \mathbf{H}^{1/2}\mathbf{P}_N^2\mathbf{H}^{1/2}\mathbf{v}\rangle \le \|\mathbf{v}\|_{\mathbf{H}}^2,$$

because $0 \preceq \mathbf{P}_N \preceq \mathbf{I}$ under the stepsize assumption $\gamma_t < 1/\lambda_1$. Hence

$$\begin{aligned}
\mathbb{E}\Big[\|\mathbf{P}_N(\mathbf{w}_M-\mathbf{w}_2^*)\|_{\mathbf{H}}^2\Big] &\le \mathbb{E}\Big[\|\mathbf{w}_M-\mathbf{w}_2^*\|_{\mathbf{H}}^2\Big] \\
&= 2\,\mathbb{E}[L_2(\mathbf{w}_M) - L_2(\mathbf{w}_2^*)]. \qquad (64)
\end{aligned}$$

Next, let

$$\mathbf{B} := \frac{\mathbf{I}_{0:k^*}}{\gamma N_{\mathtt{eff}}} + \mathbf{H}_{k^*:\infty}.$$

By the definition of k^*,

$$\lambda_i \ge \frac{1}{\gamma N_{\mathtt{eff}}}, \qquad i \le k^*.$$

Therefore

$$\frac{\mathbf{I}_{0:k^*}}{\gamma N_{\mathtt{eff}}} \preceq \mathbf{H}_{0:k^*},$$

and hence

$$\mathbf{B} = \frac{\mathbf{I}_{0:k^*}}{\gamma N_{\mathtt{eff}}} + \mathbf{H}_{k^*:\infty} \preceq \mathbf{H}_{0:k^*} + \mathbf{H}_{k^*:\infty} = \mathbf{H}. \qquad (65)$$

It follows that

$$\begin{aligned}
\mathbb{E}\Big[\|\mathbf{w}_M-\mathbf{w}_2^*\|_{\mathbf{B}}^2\Big] &\le \mathbb{E}\Big[\|\mathbf{w}_M-\mathbf{w}_2^*\|_{\mathbf{H}}^2\Big] \\
&= 2\,\mathbb{E}[L_2(\mathbf{w}_M) - L_2(\mathbf{w}_2^*)]. \qquad (66)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathbb{E}\Big[\|\mathbf{w}_M-\mathbf{w}_1^*\|_{\mathbf{B}}^2\Big] &= \mathbb{E}\Big[\|(\mathbf{w}_M-\mathbf{w}_2^*)+\boldsymbol{\delta}\|_{\mathbf{B}}^2\Big] \\
&\le 2\mathbb{E}\Big[\|\mathbf{w}_M-\mathbf{w}_2^*\|_{\mathbf{B}}^2\Big] + 2\|\boldsymbol{\delta}\|_{\mathbf{B}}^2 \\
&\le 4\,\mathbb{E}[L_2(\mathbf{w}_M) - L_2(\mathbf{w}_2^*)] + 2\|\boldsymbol{\delta}\|_{\mathbf{B}}^2, \qquad (67)
\end{aligned}$$

where in the last step we used (66).

Substituting (63), (64), and (67) into (62), we obtain

$$\begin{aligned}
\mathbb{E}\big[\mathcal{R}(\mathbf{w}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*)\big] &\lesssim \mathbb{E}[\mathcal{R}_2(\mathbf{w}_M) - \mathcal{R}_2(\mathbf{w}_2^*)] + \big\|\mathbf{P}_N\boldsymbol{\delta}\big\|_{\mathbf{H}}^2 \\
&\quad + \big(\alpha\,\mathbb{E}[\mathcal{R}_2(\mathbf{w}_M) - \mathcal{R}_2(\mathbf{w}_2^*)] + \alpha\|\boldsymbol{\delta}\|_{\mathbf{B}}^2 + \sigma_1^2\big)\,\frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}. \qquad (68)
\end{aligned}$$

Finally, by (65),

$$\|\boldsymbol{\delta}\|_{\mathbf{B}}^2 \le \|\boldsymbol{\delta}\|_{\mathbf{H}}^2.$$

Therefore

$$\mathbb{E}\big[\mathcal{R}(\mathbf{w}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*)\big] \lesssim \mathbb{E}[\mathcal{R}_2(\mathbf{w}_M) - \mathcal{R}_2(\mathbf{w}_2^*)] + \big\|\mathbf{P}_N\boldsymbol{\delta}\big\|_{\mathbf{H}}^2 + \Big(\alpha\,\mathbb{E}[\mathcal{R}_2(\mathbf{w}_M) - \mathcal{R}_2(\mathbf{w}_2^*)] + \alpha\|\boldsymbol{\delta}\|_{\mathbf{H}}^2 + \sigma_1^2\Big)\frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}. \tag{69}$$

Equivalently, writing

$$\mathtt{Excess}_2(\mathbf{w}_M) := \mathbb{E}[\mathcal{R}_2(\mathbf{w}_M) - \mathcal{R}_2(\mathbf{w}_2^*)],$$

we have

$$\mathtt{Excess}(\mathbf{w}_{M+N}) \lesssim \mathtt{Excess}_2(\mathbf{w}_M) + \big\|\mathbf{P}_N\boldsymbol{\delta}\big\|_{\mathbf{H}}^2 + \Big(\alpha\,\mathtt{Excess}_2(\mathbf{w}_M) + \alpha\|\boldsymbol{\delta}\|_{\mathbf{H}}^2 + \sigma_1^2\Big)\frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}. \tag{70}$$

This is the desired two-stage upper bound.

E.1 Two-stage training without learning-rate restart

The two-stage analysis restarts the geometrically decaying stepsize schedule at the beginning of the real-data stage. We briefly discuss how the argument changes when a single uninterrupted schedule is used across both stages.

Let $T := M + N$, and denote the global T-step schedule by

$$\gamma_t^{(T)} := \frac{\gamma}{2^{\lfloor t/(T/\log T)\rfloor}}, \qquad t = 1, \dots, T.$$

Under the restarted protocol analyzed above, the real-data stage instead uses a fresh N-step schedule

$$\gamma_s^{(N)} := \frac{\gamma}{2^{\lfloor s/(N/\log N)\rfloor}}, \qquad s = 1, \dots, N.$$

Thus, without restarting, the contraction operator appearing in the two-stage bound changes from $\mathbf{P}_N^{\mathrm{rst}} := \prod_{s=1}^{N}(\mathbf{I} - \gamma_s^{(N)}\mathbf{H})$ to the real-data tail of the global schedule $\mathbf{P}_{M,N}^{\mathrm{tail}} := \prod_{t=M+1}^{M+N}(\mathbf{I} - \gamma_t^{(T)}\mathbf{H})$.

The proof follows the same conditional argument as above. Conditional on the first-stage iterate $\mathbf{w}_M$, the second stage remains single-source SGD on real data. The decomposition $\mathbf{w}_M - \mathbf{w}_1^* = (\mathbf{w}_M - \mathbf{w}_2^*) + \boldsymbol{\delta}$ is unchanged, as are the subsequent comparisons between the stage-one excess risk and the corresponding $\mathbf{H}$-weighted errors. The only modification is that the proof of the single-source bound must be applied to the shifted stepsize sequence $\{\gamma_{M+1}^{(T)}, \dots, \gamma_{M+N}^{(T)}\}$ rather than to a freshly restarted sequence $\{\gamma_1^{(N)}, \dots, \gamma_N^{(N)}\}$. Consequently, every occurrence of $\mathbf{P}_N^{\mathrm{rst}}$ is replaced by $\mathbf{P}_{M,N}^{\mathrm{tail}}$, and the spectral cutoff and effective-dimension factor are defined using the real-stage tail schedule.

In particular, the resulting bound has the same qualitative decomposition:

$$\mathbb{E}[\mathcal{E}_1(\mathbf{w}_{M+N})] \lesssim \mathbb{E}[\mathcal{E}_2(\mathbf{w}_M)] + \big\|\mathbf{P}_{M,N}^{\mathrm{tail}}\boldsymbol{\delta}\big\|_{\mathbf{H}}^2 + \Big(\alpha\mathbb{E}[\mathcal{E}_2(\mathbf{w}_M)] + \alpha\|\boldsymbol{\delta}\|_{\mathbf{H}}^2 + \sigma_1^2\Big)\frac{D_{\mathrm{eff}}^{\mathrm{tail}}}{T_{\mathrm{eff}}},$$

where $T_{\mathrm{eff}} := T/\log T$, $k_{\mathrm{tail}}^* := \max\{k : \lambda_k \ge 1/(\gamma_{M+1}^{(T)} T_{\mathrm{eff}})\}$, and $D_{\mathrm{eff}}^{\mathrm{tail}} := k_{\mathrm{tail}}^* + (\gamma_{M+1}^{(T)} T_{\mathrm{eff}})^2 \sum_{i > k_{\mathrm{tail}}^*} \lambda_i^2$.

Nevertheless, the qualitative asymptotic conclusion is unchanged. For every eigendirection with $\lambda_i > 0$,

$$\prod_{t=M+1}^{M+N}\big(1 - \gamma_t^{(T)}\lambda_i\big) \le \exp\left(-\lambda_i \sum_{t=M+1}^{M+N}\gamma_t^{(T)}\right).$$

When $p = M/(M+N) \in (0,1)$ is fixed, the real-data tail of the geometric schedule satisfies $\sum_{t=M+1}^{M+N}\gamma_t^{(T)} \to \infty$. It follows that $\big\|\mathbf{P}_{M,N}^{\mathrm{tail}}\boldsymbol{\delta}\big\|_{\mathbf{H}}^2 \to 0$. Moreover, under the corresponding effective-dimension consistency condition, $D_{\mathrm{eff}}^{\mathrm{tail}}/T \to 0$. Therefore, using an uninterrupted schedule slows the finite-sample forgetting of the synthetic initialization, but it does not introduce a persistent synthetic drift during the real-data stage and does not create a non-vanishing excess-risk floor.

F Random Sketch Model

Following [45], we begin by decomposing the population risk into three components: irreducible risk, approximation error, and excess risk:

$$\mathcal{R}_D(\mathbf{v}_{M+N}) = \underbrace{\min \mathcal{R}(\cdot)}_{\textsf{Irreducible}} + \underbrace{\min \mathcal{R}_D(\cdot) - \min \mathcal{R}(\cdot)}_{\textsf{Approx}} + \underbrace{\mathcal{R}_D(\mathbf{v}_{M+N}) - \min \mathcal{R}_D(\cdot)}_{\textsf{Excess}}. \tag{71}$$

For the Irreducible risk, under the well-specified model Assumption 4,

$$\textsf{Irreducible} = \mathcal{R}(\mathbf{w}_1^*) = \frac{1}{2}\sigma_1^2.$$

For Approximation Error, as established in Lemma C.4 in [45], under Assumption 4, with probability at least $1 - e^{\Omega(D)}$,

$$\mathbb{E}_{\mathbf{w}_1^*}\textsf{Approx} \eqsim D^{1-a}.$$

So in the following subsections, we focus on derive the bound for $\mathbb{E}\textsf{Excess}$ using risk upper bound under Assumption 4 and 5. Unless other illustration, expectations are conditioned on $\mathbf{S}$.

F.1 Mixed Training Upper Bound

Recall the upper bound in Theorem 1, where the data covariance becomes $\mathbf{SHS}^\top$ and the optimal parameters becomes $\mathbf{v}_1^*, \mathbf{v}_2^*$.

Suppose Assumptions 2, 4 hold. Let $\widetilde{N}_{\mathtt{eff}} := (M+N)/\log(M+N)$, $\bar{\sigma}^2 = (1-p)\sigma_1^2 + p\sigma_2^2$ and $\tilde{\lambda}_j$ be the eigenvalue of $\mathbf{SHS}^\top$. Suppose $\gamma < 1/(4\alpha\,\mathrm{tr}(\mathbf{SHS}^\top))$. Then we have

$$\textsf{Excess} \lesssim \mathrm{Bias} + \mathrm{Var} + \underbrace{\alpha\, p\|\mathbf{v}_2^* - \mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2 \frac{\widetilde{D}_{\mathtt{eff}}}{\widetilde{N}_{\mathtt{eff}}}}_{\textsf{Fluct}} + \underbrace{p^2\big\|\mathbf{v}_2^* - \mathbf{v}_1^*\big\|_{\mathbf{SHS}^\top}^2}_{\textsf{Drift}},$$

with

$$\mathbf{v}_\ell^* := (\mathbf{SHS}^\top)^{-1}\mathbf{SH}\mathbf{w}_\ell^*, \quad \text{and} \quad D_{\mathtt{eff}} := \#\{\tilde{\lambda}_j \geq 1/(\widetilde{N}_{\mathtt{eff}}\gamma)\} + (\widetilde{N}_{\mathtt{eff}}\gamma)^2 \sum_{\tilde{\lambda}_j < 1/(\widetilde{N}_{\mathtt{eff}}\gamma)} \tilde{\lambda}_j^2.$$

Note that when $\sigma_1^2, \sigma_2^2 \eqsim 1$, Theorem A.4 in [45] shows that

$$\mathrm{Bias} + \mathrm{Var} \lesssim \Big\| \prod_{t=1}^{M+N} (\mathbf{I} - \gamma_t \mathbf{SHS}^\top)(\mathbf{v}_0 - \mathbf{v}_1^*)\Big\|_{\mathbf{SHS}^\top}^2 + \frac{D_{\mathtt{eff}}}{\widetilde{N}_{\mathtt{eff}}}$$

Future assume that Assumption 5 holds and zero initialiazation where $\mathbf{v}_0 = \mathbf{0}$, [45](Appendix D and E) shows that with probability at least $1 - e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$

$$\mathbb{E}_{\mathbf{w}_1^*}\Big\| \prod_{t=1}^{M+N} (\mathbf{I} - \gamma_t \mathbf{SHS}^\top)(\mathbf{v}_1^*)\Big\|_{\mathbf{SHS}^\top}^2 \lesssim \max\big\{D^{1-a},\ (\widetilde{N}_{\mathtt{eff}}\gamma)^{1/a-1}\big\},$$

$$\mathbb{E}_{\mathbf{w}_1^*}\Big\| \prod_{t=1}^{M+N} (\mathbf{I} - \gamma_t \mathbf{SHS}^\top)(\mathbf{v}_1^*)\Big\|_{\mathbf{SHS}^\top}^2 \gtrsim (\widetilde{N}_{\mathtt{eff}}\gamma)^{1/a-1} \text{ when } (\widetilde{N}_{\mathtt{eff}}\gamma)^{1/a} \leq D/c \quad \text{for some constant } c > 0,$$

and

$$\frac{D_{\mathtt{eff}}}{\widetilde{N}_{\mathtt{eff}}} \eqsim \min\big\{D,\ (\widetilde{N}_{\mathtt{eff}}\gamma)^{1/a}\big\}/\widetilde{N}_{\mathtt{eff}}(\widetilde{N}_{\mathtt{eff}}\gamma)^{1/a-1}.$$

For the Fluct term, we have

$$\mathrm{Fluct} \lesssim \|\mathbf{v}_2^* - \mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2 \cdot \frac{D_{\mathtt{eff}}}{\widetilde{N}_{\mathtt{eff}}}$$

We verify that

$$\begin{aligned}\|\mathbf{v}_2^* - \mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2 &= \|\mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top(\mathbf{v}_2^* - \mathbf{v}_1^*)\|^2 \\ &= \|\mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top(\mathbf{SHS}^\top)^{-1}\mathbf{SH}\boldsymbol{\delta}\|^2 \\ &\leq \|\mathbf{H}^{\frac{1}{2}}\boldsymbol{\delta}\|^2 = \|\boldsymbol{\delta}\|_{\mathbf{H}}^2,\end{aligned}$$

which implies that

$$\mathbb{E}_{\boldsymbol{\delta}}\|\mathbf{v}_2^* - \mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2 \lesssim \mathbb{E}\|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \eqsim \sum_i i^{-b} \eqsim 1 \tag{72}$$

So that with probability at least $1 - e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$

$$\mathbb{E}_{\boldsymbol{\delta}}\text{Fluct} \lesssim \frac{D_{\mathtt{eff}}}{\widetilde{N}_{\mathtt{eff}}} \eqsim \min\left\{D,\ (\widetilde{N}_{\mathtt{eff}}\gamma)^{1/a}\right\}/\widetilde{N}_{\mathtt{eff}} \lesssim D^{1-a} + (\widetilde{N}_{\mathtt{eff}}\gamma)^{1/a-1}.$$

For the Drift term $p^2\|\mathbf{v}_2^* - \mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2$,

$$\begin{aligned}\|\mathbf{v}_2^* - \mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2 &= \|(\mathbf{SHS}^\top)^{-1}\mathbf{SH}(\mathbf{w}_2^* - \mathbf{w}_1^*)\|_{\mathbf{SHS}^\top}^2 \\ &= \boldsymbol{\delta}^\top\mathbf{HS}^\top(\mathbf{SHS}^\top)^{-1}(\mathbf{SHS}^\top)(\mathbf{SHS}^\top)^{-1}\mathbf{SH}\boldsymbol{\delta} \\ &= \boldsymbol{\delta}^\top\mathbf{HS}^\top(\mathbf{SHS}^\top)^{-1}\mathbf{SH}\boldsymbol{\delta} + \boldsymbol{\delta}^\top\mathbf{H}\boldsymbol{\delta} - \boldsymbol{\delta}\top\mathbf{H}\boldsymbol{\delta} \\ &= \boldsymbol{\delta}^\top\mathbf{H}\boldsymbol{\delta} - \boldsymbol{\delta}^\top(\mathbf{H} - \mathbf{HS}^\top(\mathbf{SHS}^\top)^{-1}\mathbf{SH})\boldsymbol{\delta} \\ &= \|\boldsymbol{\delta}\|_{\mathbf{H}}^2 - \left\|\left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top(\mathbf{SHS}^\top)^{-1}\mathbf{S}^\top\mathbf{H}^{\frac{1}{2}}\right)\mathbf{H}^{\frac{1}{2}}\boldsymbol{\delta}\right\|^2\end{aligned}$$

For the second minus term, [45](Lemma C.5) shows that for any random vector $\mathbf{x}$, let $(\lambda_i, \mathbf{v}_i)_{i\geq 1}$ be the eigenvalue-eigenvector pairs of $\mathbf{H}$, if the Assumption 5 (power-law decay) satisfies and $\mathbb{E}[\langle\mathbf{v}_i, \mathbf{x}\rangle\langle\mathbf{v}_j, \mathbf{x}\rangle] = 0$ for $i \neq j$ and $\mathbb{E}[\lambda_i\langle\mathbf{v}_i, \mathbf{x}\rangle^2] \eqsim i^{-b}$ for some $b > 1$, Then with probability at least $1 - e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$

$$\mathbb{E}_{\mathbf{x}}\mathsf{Approx}(\mathbf{S}, \mathbf{H}, \mathbf{x}) := \left\|\left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top(\mathbf{SHS}^\top)^{-1}\mathbf{S}^\top\mathbf{H}^{\frac{1}{2}}\right)\mathbf{H}^{\frac{1}{2}}\mathbf{x}\right\|^2 \eqsim D^{1-b}$$

So that under Assumption 5,

$$\mathbb{E}_{\boldsymbol{\delta}}\left\|\left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top(\mathbf{SHS}^\top)^{-1}\mathbf{S}^\top\mathbf{H}^{\frac{1}{2}}\right)\mathbf{H}^{\frac{1}{2}}\boldsymbol{\delta}\right\|^2 \eqsim D^{1-b} \tag{73}$$

Taking expectation in the decomposition and using (73), we obtain

$$\mathbb{E}_{\boldsymbol{\delta}}\|\mathbf{v}_2^* - \mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2 = \mathbb{E}_{\boldsymbol{\delta}}\|\boldsymbol{\delta}\|_{\mathbf{H}}^2 - \Theta(D^{1-b}). \tag{74}$$

Since $\mathbb{E}_{\boldsymbol{\delta}}\|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \eqsim 1$, we suppress this problem-dependent constant in the scaling-law statement and write, for sufficiently large D,

$$\mathbb{E}_{\boldsymbol{\delta}}\|\mathbf{v}_2^* - \mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2 \eqsim 1 - D^{1-b}. \tag{75}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\delta}}\mathsf{Drift} &= p^2\mathbb{E}_{\boldsymbol{\delta}}\|\mathbf{v}_2^* - \mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2 \\ &\eqsim p^2(1 - D^{1-b}).\end{aligned}$$

Finally, put $\mathsf{Irreducible}$, Approx and Excess together and choose $\gamma \eqsim 1$, we have that with probability at least $1 - e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$:

$$\mathbb{E}\mathcal{R}_D(\mathbf{v}_{M+N}) \lesssim \sigma_1^2 + \frac{1}{D^{a-1}} + \frac{1}{(\widetilde{N}_{\mathtt{eff}})^{1-1/a}} + \left(\frac{M}{M+N}\right)^2(1 - \frac{1}{D^{b-1}})$$

F.2 Mixed Training Lower Bound

We now establish the matching lower bound in the optimization-saturated regime $\gamma\widetilde{N}_{\texttt{eff}} \gtrsim D^a$.

Applying Theorem 2 to the sketched problem, with $\mathbf{H}$ replaced by $\mathbf{SHS}^\top$ and using $\mathbf{v}_0 = \mathbf{0}$, gives

$$\mathbb{E}\mathsf{Excess} \gtrsim \mathbb{E}_{\mathbf{w}_1^*,\boldsymbol{\delta}}\left\| -\left(\prod_{t=1}^{M+N}(\mathbf{I}-\gamma_t\mathbf{SHS}^\top)\right)\mathbf{v}_1^* + p\left(\mathbf{I}-\prod_{t=1}^{M+N}(\mathbf{I}-\gamma_t\mathbf{SHS}^\top)\right)(\mathbf{v}_2^*-\mathbf{v}_1^*)\right\|_{\mathbf{SHS}^\top}^2,$$

where we have dropped the remaining nonnegative terms in Theorem 2.

By Assumption 4,

$$\begin{aligned}\mathbb{E}_{\mathbf{w}_1^*,\boldsymbol{\delta}}\left[\mathbf{v}_1^*(\mathbf{v}_2^*-\mathbf{v}_1^*)^\top\right] &= (\mathbf{SHS}^\top)^{-1}\mathbf{SH}\mathbb{E}[\mathbf{w}_1^*\boldsymbol{\delta}^\top]\mathbf{HS}^\top(\mathbf{SHS}^\top)^{-1}\\ &= \mathbf{0}.\end{aligned}$$

Therefore, the cross term vanishes after taking expectation over $(\mathbf{w}_1^*, \boldsymbol{\delta})$, and hence

$$\mathbb{E}\mathsf{Excess} \gtrsim p^2\mathbb{E}_{\boldsymbol{\delta}}\left\|\left(\mathbf{I}-\prod_{t=1}^{M+N}(\mathbf{I}-\gamma_t\mathbf{SHS}^\top)\right)(\mathbf{v}_2^*-\mathbf{v}_1^*)\right\|_{\mathbf{SHS}^\top}^2. \tag{76}$$

It remains to control the spectral filter in (76). By Lemma 6.2 of Lin et al. [45], with probability at least $1-e^{-\Omega(D)}$, the eigenvalues of $\mathbf{SHS}^\top$ satisfy

$$\widetilde{\lambda}_j \eqsim j^{-a}, \qquad j = 1,\dots,D.$$

Moreover, the geometric stepsize schedule satisfies $\sum_{t=1}^{M+N}\gamma_t \eqsim \gamma\widetilde{N}_{\texttt{eff}}$. Thus, for every $j = 1,\dots,D$,

$$\begin{aligned}\prod_{t=1}^{M+N}(1-\gamma_t\widetilde{\lambda}_j) &\le \exp\left(-\widetilde{\lambda}_j\sum_{t=1}^{M+N}\gamma_t\right)\\ &\le \exp\left(-c\gamma\widetilde{N}_{\texttt{eff}}D^{-a}\right) \le e^{-c},\end{aligned}$$

where the last inequality follows from $\gamma\widetilde{N}_{\texttt{eff}} \gtrsim D^a$. Let $(\widetilde{\lambda}_j, \widetilde{\mathbf{v}}_j)_{j=1}^D$ be the eigenvalue-eigenvector pairs of $\mathbf{SHS}^\top$. Expanding the norm in this eigenbasis gives

$$\begin{aligned}&\left\|\left(\mathbf{I}-\prod_{t=1}^{M+N}(\mathbf{I}-\gamma_t\mathbf{SHS}^\top)\right)(\mathbf{v}_2^*-\mathbf{v}_1^*)\right\|_{\mathbf{SHS}^\top}^2\\ &= \sum_{j=1}^{D}\widetilde{\lambda}_j\left(1-\prod_{t=1}^{M+N}(1-\gamma_t\widetilde{\lambda}_j)\right)^2\langle\widetilde{\mathbf{v}}_j,\mathbf{v}_2^*-\mathbf{v}_1^*\rangle^2\\ &\ge (1-e^{-c})^2\sum_{j=1}^{D}\widetilde{\lambda}_j\langle\widetilde{\mathbf{v}}_j,\mathbf{v}_2^*-\mathbf{v}_1^*\rangle^2\\ &= (1-e^{-c})^2\|\mathbf{v}_2^*-\mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2.\end{aligned}$$

Therefore,

$$\mathbb{E}_{\boldsymbol{\delta}}\left\|\left(\mathbf{I}-\prod_{t=1}^{M+N}(\mathbf{I}-\gamma_t\mathbf{SHS}^\top)\right)(\mathbf{v}_2^*-\mathbf{v}_1^*)\right\|_{\mathbf{SHS}^\top}^2 \gtrsim \mathbb{E}_{\boldsymbol{\delta}}\|\mathbf{v}_2^*-\mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2.$$

The upper-bound analysis above has already established that

$$\mathbb{E}_{\boldsymbol{\delta}}\|\mathbf{v}_2^*-\mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2 \eqsim 1-D^{1-b}.$$

Substituting this result into (76) gives $\mathbb{E}\mathsf{Excess} \gtrsim p^2(1-D^{1-b})$. In addition, the approximation-error result stated above gives $\mathbb{E}_{\mathbf{w}_1^*}\mathsf{Approx} \eqsim D^{1-a}$. Combining this with (71), we obtain

$$\mathbb{E}\mathcal{R}_D(\mathbf{v}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*) \gtrsim D^{1-a} + p^2(1-D^{1-b}).$$

Finally, since $\gamma \eqsim 1$ and $\gamma\widetilde{N}_{\texttt{eff}} \gtrsim D^a$, we have $\widetilde{N}_{\texttt{eff}}^{1/a-1} \lesssim D^{1-a}$. Therefore,

$$\mathbb{E}\mathcal{R}_D(\mathbf{v}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*) \gtrsim D^{1-a} + \widetilde{N}_{\texttt{eff}}^{1/a-1} + p^2(1-D^{1-b}),$$

which matches the upper bound in Theorem 6 up to constants.

F.3 Two-stage Training Analysis

First, Recall the upper bound in Theorem 4. For a tighter analysis, we need a intermediate result from its proof. The results are for specific $\mathbf{w}_1^*, \mathbf{w}_2^*$ and the expectation is taken with respect to the one-stage M-step SGD.

Suppose Assumptions 1, 2 and 4 hold. Let $N_{\mathtt{eff}} := N/\log N$ and $M_{\mathtt{eff}} = M/\log M$. Suppose $\gamma < 1/(4\alpha\,\mathrm{tr}(\mathbf{SHS}^\top))$. Then we have

$$\mathtt{Excess}(\mathbf{v}_{M+N}) \lesssim \mathbb{E}\left[\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{SHS}^\top)\Big)(\mathbf{v}_M-\mathbf{v}_2^*)\Big\|_{\mathbf{SHS}^\top}^2\right] + \big\|(\prod_{t=1}^{N}\mathbf{I}-\gamma_t\mathbf{SHS}^\top)(\mathbf{v}_2^*-\mathbf{v}_1^*)\big\|_{\mathbf{SHS}^\top}^2 + \alpha\big(\mathsf{Excess}_2(\mathbf{v}_M) + \|\mathbf{v}_2^*-\mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2\big)\frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}} + \mathsf{Var}$$

where

$$\mathbf{v}_\ell^* := (\mathbf{SHS}^\top)^{-1}\mathbf{SH}\mathbf{w}_\ell^*, \quad \text{and} \quad D_{\mathtt{eff}} := \#\{\tilde\lambda_j \ge 1/(N_{\mathtt{eff}}\gamma)\} + (N_{\mathtt{eff}}\gamma)^2\sum_{\tilde\lambda_j < 1/(N_{\mathtt{eff}}\gamma)}\tilde\lambda_j^2.$$

Suppose $\sigma_1^2, \sigma_2^2 \eqsim 1$. For the Var term, similar to the analysis in mixed training analysis and [45](Lemma E.1), with probability at least $1-e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$

$$\mathsf{Var} \eqsim \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}} \eqsim \min\big\{D,\ (N_{\mathtt{eff}}\gamma)^{1/a}\big\}/N_{\mathtt{eff}}.$$

Analysis of $\mathsf{Excess}_2(\mathbf{v}_M)$. We now analyze the first-stage excess risk when the teacher is $\mathbf{w}_2^*$. Conditioning on the sketch matrix $\mathbf{S}$, define

$$\mathbf{v}_2^* := (\mathbf{SHS}^\top)^{-1}\mathbf{SH}\mathbf{w}_2^*, \qquad M_{\mathtt{eff}} := M/\log M.$$

Using Theorem 6.1 in [45], for $\gamma < 1/(4\alpha\ \mathrm{tr}(\mathbf{SHS}^\top))$ and zero initialization,

$$\mathbb{E}\mathsf{Excess}_2(\mathbf{v}_M) \lesssim \mathsf{Bias}_2 + \frac{D_{\mathtt{eff}1}}{M_{\mathtt{eff}}},$$

where

$$D_{\mathtt{eff}1} := \#\{\tilde\lambda_j \ge 1/(M_{\mathtt{eff}}\gamma)\} + (M_{\mathtt{eff}}\gamma)^2\sum_{\tilde\lambda_j < 1/(M_{\mathtt{eff}}\gamma)}\tilde\lambda_j^2,$$

and the expectation is over the randomness of $\mathbf{w}_2^*$ and the first-stage SGD.

By Assumption 5, let $(\lambda_i, \mathbf{v}_i)_{i\ge1}$ be the eigenvalue-eigenvector pairs of $\mathbf{H}$, and $\mathbf{\Sigma}$ shares the eigenvectors $(\mathbf{v}_i)_{i\ge1}$ with $\mathbf{H}$, i.e.,

$$\mathbf{\Sigma}\mathbf{v}_i = \tau_i\mathbf{v}_i, \qquad \lambda_i \eqsim i^{-a}, \qquad \lambda_i\tau_i \eqsim i^{-b},$$

for some $a > 1$ and $b > 1$. Then

$$\mathbb{E}\langle\mathbf{v}_i, \mathbf{w}_2^*\rangle^2 = \mathbb{E}\langle\mathbf{v}_i, \mathbf{w}_1^*\rangle^2 + \mathbb{E}\langle\mathbf{v}_i, \boldsymbol{\delta}\rangle^2 = 1+\tau_i,$$

where we used $\mathbb{E}[\mathbf{w}_1^*\boldsymbol{\delta}^\top] = 0$.

For the bias term, applying Lemma D.1 in [45] with a cutoff $k \le D/3$ yields

$$\mathbb{E}\mathsf{Bias}_2 \lesssim \frac{\sum_{i\le k}\mathbb{E}\langle\mathbf{v}_i, \mathbf{w}_2^*\rangle^2}{M_{\mathtt{eff}}\gamma} + \sum_{i>k}\lambda_i\mathbb{E}\langle\mathbf{v}_i, \mathbf{w}_2^*\rangle^2.$$

Substituting $\mathbb{E}\langle\mathbf{v}_i, \mathbf{w}_2^*\rangle^2 = 1+\tau_i$ gives

$$\mathbb{E}\mathsf{Bias}_2 \lesssim \frac{\sum_{i\le k}(1+\tau_i)}{M_{\mathtt{eff}}\gamma} + \sum_{i>k}\lambda_i(1+\tau_i).$$

Using $\tau_i \eqsim i^{a-b}$, we obtain

$$\sum_{i\le k}(1+\tau_i) \eqsim k + k^{1+a-b}, \qquad \sum_{i>k}\lambda_i(1+\tau_i) \eqsim k^{1-a} + k^{1-b},$$

and hence

$$\mathbb{E}\mathsf{Bias}_2 \lesssim \frac{k + k^{1+a-b}}{M_{\mathtt{eff}}\gamma} + k^{1-a} + k^{1-b}.$$

Let $c := \min\{a, b\}$. and distinguish two cases. If $b \geq a$, then $k^{1+a-b} \lesssim k$ and $k^{1-b} \lesssim k^{1-a}$, so

$$\mathbb{E}\mathsf{Bias}_2 \lesssim \frac{k}{M_{\mathtt{eff}}\gamma} + k^{1-a}.$$

If $b < a$, then $k \lesssim k^{1+a-b}$ and $k^{1-a} \lesssim k^{1-b}$, so

$$\mathbb{E}\mathsf{Bias}_2 \lesssim \frac{k^{1+a-b}}{M_{\mathtt{eff}}\gamma} + k^{1-b}.$$

Hence, in both cases,

$$\mathbb{E}\mathsf{Bias}_2 \lesssim \frac{k^{1+a-c}}{M_{\mathtt{eff}}\gamma} + k^{1-c}, \qquad c = \min\{a, b\}.$$

Choosing

$$k \eqsim \min\{D, (M_{\mathtt{eff}}\gamma)^{1/a}\},$$

yields

$$\mathbb{E}\mathsf{Bias}_2 \lesssim \max\{D^{1-c}, (M_{\mathtt{eff}}\gamma)^{(1-c)/a}\}.$$

For the lower bound, note that $\mathbb{E}[\mathbf{w}_2^*(\mathbf{w}_2^*)^\top] = \mathbf{I} + \mathbf{\Sigma}$, under the assumption $\mathbb{E}[\mathbf{w}_1^*\boldsymbol{\delta}^\top] = 0$. Since $\mathbf{\Sigma}$ shares the eigenvectors of $\mathbf{H}$ and $\lambda_i\tau_i \eqsim i^{-b}$, the eigenvalues of $\mathbf{H}(\mathbf{I} + \mathbf{\Sigma})$ satisfy

$$\lambda_i(1 + \tau_i) \eqsim i^{-a} + i^{-b} \eqsim i^{-c}, \qquad c = \min\{a, b\}.$$

Therefore, applying the same argument as in Lemma D.2 and Lemma D.4 of [45], we obtain

$$\mathbb{E}\mathsf{Bias}_2 \gtrsim (M_{\mathtt{eff}}\gamma)^{(1-c)/a} \qquad \text{when} \qquad (M_{\mathtt{eff}}\gamma)^{1/a} \leq D/c_0$$

for some constant $c_0 > 0$. Consequently, with probability at least $1 - e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$, the upper bound is

$$\mathbb{E}\mathsf{Bias}_2 \lesssim \max\{D^{1-c}, (M_{\mathtt{eff}}\gamma)^{(1-c)/a}\}, \qquad c = \min\{a, b\},$$

Moreover, in the data-limited regime $(M_{\mathtt{eff}}\gamma)^{1/a} \leq D/c_0$, this upper bound is tight up to constants.

Finally, by Lemma E.1 in [45],

$$\frac{D_{\mathtt{eff}\,1}}{M_{\mathtt{eff}}} \eqsim \frac{\min\{D, (M_{\mathtt{eff}}\gamma)^{1/a}\}}{M_{\mathtt{eff}}}$$

with probability at least $1 - e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$. Combining the above estimates and choose $\gamma \eqsim 1$, we conclude that with probability at least $1 - e^{-\Omega(D)}$ over $\mathbf{S}$,

$$\mathbb{E}\mathsf{Excess}_2(\mathbf{v}_M) \lesssim \max\{D^{1-a\wedge b}, (M_{\mathtt{eff}}\gamma)^{(1-a\wedge b)/a}\} + \frac{\min\{D, (M_{\mathtt{eff}}\gamma)^{1/a}\}}{M_{\mathtt{eff}}} \lesssim 1$$

For the term $\alpha\big(\mathsf{Excess}_2(\mathbf{v}_M) + \|\mathbf{v}_2^* - \mathbf{v}_1^*\|^2_{\mathbf{SHS}^\top}\big)\frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}$, we claim that it is of the same order as $D_{\mathtt{eff}}/N_{\mathtt{eff}}$. Indeed, by the analysis in mixed-training-analysis, with probability at least $1 - e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$,

$$\mathbb{E}_{\boldsymbol{\delta}}\|\mathbf{v}_2^* - \mathbf{v}_1^*\|^2_{\mathbf{SHS}^\top} \lesssim 1.$$

Therefore, on the intersection of the above high-probability events,

$$\alpha\Big(\mathbb{E}\mathsf{Excess}_2(\mathbf{v}_M) + \mathbb{E}_{\boldsymbol{\delta}}\|\mathbf{v}_2^* - \mathbf{v}_1^*\|^2_{\mathbf{SHS}^\top}\Big)\frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}} \lesssim \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}} \eqsim \frac{\min\{D, (N_{\mathtt{eff}}\gamma)^{1/a}\}}{N_{\mathtt{eff}}}.$$

For the term $\|(\prod_{t=1}^{N}\mathbf{I}-\gamma_t\mathbf{S}\mathbf{H}\mathbf{S}^\top)(\mathbf{v}_2^*-\mathbf{v}_1^*)\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2$, under Assumption 5, we can apply Lemma D.4 in [45] directly and get that probability at least $1-e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$:

$$\mathbb{E}_{\boldsymbol{\delta}}\|(\prod_{t=1}^{N}\mathbf{I}-\gamma_t\mathbf{S}\mathbf{H}\mathbf{S}^\top)(\mathbf{v}_2^*-\mathbf{v}_1^*)\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2\lesssim\max\big\{D^{1-b},(N_{\mathtt{eff}}\gamma)^{(1-b)/a}\big\},$$

and

$$\mathbb{E}_{\boldsymbol{\delta}}\|(\prod_{t=1}^{N}\mathbf{I}-\gamma_t\mathbf{S}\mathbf{H}\mathbf{S}^\top)(\mathbf{v}_2^*-\mathbf{v}_1^*)\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2\gtrsim(N_{\mathtt{eff}}\gamma)^{(1-b)/a}$$

when $(N_{\mathtt{eff}}\gamma)^{1/a}\le D/c$ for some constant $c>0$. Moreover, when $b\ge a+1$

$$\mathbb{E}_{\boldsymbol{\delta}}\|(\prod_{t=1}^{N}\mathbf{I}-\gamma_t\mathbf{S}\mathbf{H}\mathbf{S}^\top)(\mathbf{v}_2^*-\mathbf{v}_1^*)\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2\lesssim\log N_{\mathtt{eff}}\cdot\max\big\{(N_{\mathtt{eff}}\gamma)^{(1-b)/a},\ D^{1-b}\big\}$$

For the last term

$$\mathbb{E}\bigg[\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{S}\mathbf{H}\mathbf{S}^\top)\Big)(\mathbf{v}_M-\mathbf{v}_2^*)\Big\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2\bigg],$$

let $\mathbf{P}_N:=\prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{S}\mathbf{H}\mathbf{S}^\top)$. Conditioning on the sketch matrix $\mathbf{S}$, we diagonalize $\mathbf{S}\mathbf{H}\mathbf{S}^\top=\mathbf{U}\mathrm{diag}(\tilde\lambda_1,\dots,\tilde\lambda_D)\mathbf{U}^\top$ with $\tilde\lambda_1\ge\cdots\ge\tilde\lambda_D>0$. Then

$$\mathbf{P}_N^\top\mathbf{S}\mathbf{H}\mathbf{S}^\top\mathbf{P}_N=\mathbf{U}\mathrm{diag}\Big(\tilde\lambda_i\prod_{t=1}^{N}(1-\gamma_t\tilde\lambda_i)^2\Big)\mathbf{U}^\top\preceq\Big(\prod_{t=1}^{N}(1-\gamma_t\tilde\lambda_D)\Big)^2\mathbf{S}\mathbf{H}\mathbf{S}^\top.$$

Hence

$$\mathbb{E}\big[\|\mathbf{P}_N(\mathbf{v}_M-\mathbf{v}_2^*)\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2\,\big|\,\mathbf{S}\big]\le\Big(\prod_{t=1}^{N}(1-\gamma_t\tilde\lambda_D)\Big)^2\mathbb{E}\big[\|\mathbf{v}_M-\mathbf{v}_2^*\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2\,\big|\,\mathbf{S}\big].$$

Using $1-u\le e^{-u}$ and $\sum_{t=1}^{N}\gamma_t\eqsim N_{\mathtt{eff}}\gamma$, we further obtain

$$\mathbb{E}\big[\|\mathbf{P}_N(\mathbf{v}_M-\mathbf{v}_2^*)\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2\,\big|\,\mathbf{S}\big]\lesssim\exp\big(-c\,N_{\mathtt{eff}}\gamma\,\mu_D(\mathbf{S}\mathbf{H}\mathbf{S}^\top)\big)\,\mathbb{E}\big[\|\mathbf{v}_M-\mathbf{v}_2^*\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2\,\big|\,\mathbf{S}\big]$$

for some constant $c>0$. Since

$$\mathbb{E}\big[\|\mathbf{v}_M-\mathbf{v}_2^*\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2\,\big|\,\mathbf{S}\big]=\mathbb{E}[\mathsf{Excess}_2(\mathbf{v}_M)\,|\,\mathbf{S}],$$

it follows that

$$\mathbb{E}\big[\|\mathbf{P}_N(\mathbf{v}_M-\mathbf{v}_2^*)\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2\,\big|\,\mathbf{S}\big]\lesssim\exp\big(-c\,N_{\mathtt{eff}}\gamma\,\mu_D(\mathbf{S}\mathbf{H}\mathbf{S}^\top)\big)\,\mathbb{E}[\mathsf{Excess}_2(\mathbf{v}_M)\,|\,\mathbf{S}].$$

Now let $\mathcal{E}_{\mathrm{spec}}$ be the event that

$$\mu_D(\mathbf{S}\mathbf{H}\mathbf{S}^\top)\eqsim D^{-a},$$

and let $\mathcal{E}_{\mathrm{ex}}$ be the event that

$$\mathbb{E}[\mathsf{Excess}_2(\mathbf{v}_M)\,|\,\mathbf{S}]\lesssim\max\{D^{1-a\wedge b},(M_{\mathtt{eff}}\gamma)^{(1-a\wedge b)/a}\}.$$

By Lemma 6.2 in [45] and the above analysis of $\mathsf{Excess}_2(\mathbf{v}_M)$, both events hold with probability at least $1-e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$. Therefore, on the intersection event $\mathcal{E}_{\mathrm{spec}}\cap\mathcal{E}_{\mathrm{ex}}$, we have

$$\mathbb{E}\bigg[\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{S}\mathbf{H}\mathbf{S}^\top)\Big)(\mathbf{v}_M-\mathbf{v}_2^*)\Big\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2\bigg]\lesssim\exp\bigg(-c\,\frac{N_{\mathtt{eff}}\gamma}{D^a}\bigg)\max\{D^{1-a\wedge b},(M_{\mathtt{eff}}\gamma)^{(1-a\wedge b)/a}\}.$$

Consequently, with probability at least $1-e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$,

$$\mathbb{E}\bigg[\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{S}\mathbf{H}\mathbf{S}^\top)\Big)(\mathbf{v}_M-\mathbf{v}_2^*)\Big\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2\bigg]\lesssim\exp\bigg(-c\,\frac{N_{\mathtt{eff}}\gamma}{D^a}\bigg)\max\{D^{1-a\wedge b},(M_{\mathtt{eff}}\gamma)^{(1-a\wedge b)/a}\}.$$

Finally, put Irreducible, Approx and Excess together. Choose $\gamma \eqsim 1$ and assume $b < a + 1$ for simplicity and clarity, we have that with probability at least $1 - e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$:

$$\mathbb{E}\mathcal{R}_D(\mathbf{v}_{M+N}) \lesssim \sigma_1^2 + \frac{1}{D^{a-1}} + \underbrace{\frac{\min\{D, (N_{\texttt{eff}})^{1/a}\}}{N_{\texttt{eff}}}}_{\textsf{Var}} + \underbrace{e^{-\Omega(\frac{N_{\texttt{eff}}}{D^a})} \cdot \max\Big\{\frac{1}{D^{a\wedge b-1}}, \frac{1}{(M_{\texttt{eff}})^{\frac{a\wedge b-1}{a}}}\Big\} + \max\Big\{\frac{1}{D^{b-1}}, \frac{1}{(N_{\texttt{eff}})^{\frac{b-1}{a}}}\Big\}}_{\textsf{Bias}}$$

G Exact Synthetic-Benefit Condition

Proof. [Proof of Theorem 8] Reindex the N real-stage samples as $\{(\mathbf{x}_t, y_t)\}_{t=1}^N$, where

$$y_t = \langle \mathbf{x}_t, \mathbf{w}_1^* \rangle + \xi_{1,t}.$$

For any real-stage iterate, the error recursion is

$$\mathbf{w}_t - \mathbf{w}_1^* = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)(\mathbf{w}_{t-1} - \mathbf{w}_1^*) + \gamma_t \xi_{1,t} \mathbf{x}_t.$$

Iterating this recursion for real-only training gives

$$\mathbf{w}_N^{\text{real}} - \mathbf{w}_1^* = \Phi_N(\mathbf{w}_0 - \mathbf{w}_1^*) + \sum_{t=1}^N \gamma_t \Big[(\mathbf{I} - \gamma_N \mathbf{x}_N \mathbf{x}_N^\top) \cdots (\mathbf{I} - \gamma_{t+1} \mathbf{x}_{t+1} \mathbf{x}_{t+1}^\top)\Big] \xi_{1,t} \mathbf{x}_t, \tag{77}$$

where the product inside the summation is interpreted as $\mathbf{I}$ when $t = N$.

The two-stage procedure uses exactly the same real samples, label noises, and stepsizes, but starts the real stage from $\mathbf{w}_M$. Therefore,

$$\mathbf{w}_{M+N} - \mathbf{w}_1^* = \Phi_N(\mathbf{w}_M - \mathbf{w}_1^*) + \sum_{t=1}^N \gamma_t \Big[(\mathbf{I} - \gamma_N \mathbf{x}_N \mathbf{x}_N^\top) \cdots (\mathbf{I} - \gamma_{t+1} \mathbf{x}_{t+1} \mathbf{x}_{t+1}^\top)\Big] \xi_{1,t} \mathbf{x}_t. \tag{78}$$

In particular, the real-stage noise accumulation in (77) and (78) is identical.

Let $\boldsymbol{\nu}_N$ denote the common real-stage noise-accumulation term in (77) and (78). These two recursions can be written as

$$\mathbf{w}_N^{\text{real}} - \mathbf{w}_1^* = \Phi_N(\mathbf{w}_0 - \mathbf{w}_1^*) + \boldsymbol{\nu}_N, \qquad \mathbf{w}_{M+N} - \mathbf{w}_1^* = \Phi_N(\mathbf{w}_M - \mathbf{w}_1^*) + \boldsymbol{\nu}_N.$$

Conditional on $(\mathbf{w}_M, \mathbf{x}_1, \ldots, \mathbf{x}_N)$, the coefficients of all $\xi_{1,t}$ in $\boldsymbol{\nu}_N$ are fixed. Since the real label noises are independent of the covariates and the synthetic stage, and satisfy $\mathbb{E}[\xi_{1,t}] = 0$, we have

$$\mathbb{E}[\boldsymbol{\nu}_N \mid \mathbf{w}_M, \mathbf{x}_1, \ldots, \mathbf{x}_N] = \mathbf{0}.$$

Both $\Phi_N(\mathbf{w}_0 - \mathbf{w}_1^*)$ and $\Phi_N(\mathbf{w}_M - \mathbf{w}_1^*)$ are measurable under this conditioning. Therefore, by the tower property,

$$\mathbb{E}\Big[\big(\Phi_N(\mathbf{w}_0 - \mathbf{w}_1^*)\big)^\top \mathbf{H} \boldsymbol{\nu}_N\Big] = 0,$$
$$\mathbb{E}\Big[\big(\Phi_N(\mathbf{w}_M - \mathbf{w}_1^*)\big)^\top \mathbf{H} \boldsymbol{\nu}_N\Big] = 0.$$

Expanding the two squared errors consequently gives

$$\mathbb{E}\|\mathbf{w}_N^{\text{real}} - \mathbf{w}_1^*\|_{\mathbf{H}}^2 = \mathbb{E}\|\Phi_N(\mathbf{w}_0 - \mathbf{w}_1^*)\|_{\mathbf{H}}^2 + \mathbb{E}\|\boldsymbol{\nu}_N\|_{\mathbf{H}}^2,$$
$$\mathbb{E}\|\mathbf{w}_{M+N} - \mathbf{w}_1^*\|_{\mathbf{H}}^2 = \mathbb{E}\|\Phi_N(\mathbf{w}_M - \mathbf{w}_1^*)\|_{\mathbf{H}}^2 + \mathbb{E}\|\boldsymbol{\nu}_N\|_{\mathbf{H}}^2.$$

The final terms are identical because the two procedures use the same real samples, label noises, and stepsizes. They therefore cancel in the risk difference.

Under Assumption 3, for any $\mathbf{w}$, $\mathcal{E}_1(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{w}_1^*\|_{\mathbf{H}}^2$. Hence,

$$\begin{aligned}
&2\left(\mathbb{E}[\mathcal{E}_1(\mathbf{w}_N^{\mathrm{real}})] - \mathbb{E}[\mathcal{E}_1(\mathbf{w}_{M+N})]\right) \\
&\quad = \mathbb{E}\|\Phi_N(\mathbf{w}_0 - \mathbf{w}_1^*)\|_{\mathbf{H}}^2 - \mathbb{E}\|\Phi_N(\mathbf{w}_M - \mathbf{w}_1^*)\|_{\mathbf{H}}^2.
\end{aligned}$$

Since $\mathbf{w}_0$ is deterministic, the definition of $\mathbf{K}_N$ gives

$$\begin{aligned}
\mathbb{E}\|\Phi_N(\mathbf{w}_0 - \mathbf{w}_1^*)\|_{\mathbf{H}}^2 &= (\mathbf{w}_0 - \mathbf{w}_1^*)^\top \mathbb{E}_{\mathbf{x}_1,\dots,\mathbf{x}_N}[\Phi_N^\top \mathbf{H} \Phi_N](\mathbf{w}_0 - \mathbf{w}_1^*) \\
&= \|\mathbf{w}_0 - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2.
\end{aligned}$$

Similarly, because $\mathbf{w}_M$ is independent of the real-stage covariates, the tower property yields

$$\begin{aligned}
\mathbb{E}\|\Phi_N(\mathbf{w}_M - \mathbf{w}_1^*)\|_{\mathbf{H}}^2 &= \mathbb{E}_{\mathbf{w}_M}\left[(\mathbf{w}_M - \mathbf{w}_1^*)^\top \mathbb{E}_{\mathbf{x}_1,\dots,\mathbf{x}_N}[\Phi_N^\top \mathbf{H} \Phi_N](\mathbf{w}_M - \mathbf{w}_1^*)\right] \\
&= \mathbb{E}\|\mathbf{w}_M - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2.
\end{aligned}$$

Finally, decompose $\mathbf{w}_M - \mathbf{w}_1^* = \big(\mathbf{w}_M - \mathbb{E}[\mathbf{w}_M]\big) + \big(\mathbb{E}[\mathbf{w}_M] - \mathbf{w}_1^*\big)$ and Expandthe corresponding quadratic form to get

$$\begin{aligned}
\mathbb{E}\|\mathbf{w}_M - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2 &= \mathbb{E}\|\mathbf{w}_M - \mathbb{E}[\mathbf{w}_M]\|_{\mathbf{K}_N}^2 + \|\mathbb{E}[\mathbf{w}_M] - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2 \\
&\quad + 2\mathbb{E}\Big[\big(\mathbf{w}_M - \mathbb{E}[\mathbf{w}_M]\big)^\top \mathbf{K}_N \big(\mathbb{E}[\mathbf{w}_M] - \mathbf{w}_1^*\big)\Big].
\end{aligned}$$

The last term is zero because $\mathbb{E}[\mathbf{w}_M - \mathbb{E}[\mathbf{w}_M]] = \mathbf{0}$. Moreover,

$$\begin{aligned}
\mathbb{E}\|\mathbf{w}_M - \mathbb{E}[\mathbf{w}_M]\|_{\mathbf{K}_N}^2 &= \mathbb{E}\,\mathrm{tr}\big(\mathbf{K}_N(\mathbf{w}_M - \mathbb{E}[\mathbf{w}_M])(\mathbf{w}_M - \mathbb{E}[\mathbf{w}_M])^\top\big) \\
&= \mathrm{tr}(\mathbf{K}_N \,\mathrm{Cov}(\mathbf{w}_M)).
\end{aligned}$$

Combining the preceding identities yields

$$\begin{aligned}
&2\left(\mathbb{E}[\mathcal{E}_1(\mathbf{w}_N^{\mathrm{real}})] - \mathbb{E}[\mathcal{E}_1(\mathbf{w}_{M+N})]\right) \\
&\quad = \|\mathbf{w}_0 - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2 - \|\mathbb{E}[\mathbf{w}_M] - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2 - \mathrm{tr}(\mathbf{K}_N \,\mathrm{Cov}(\mathbf{w}_M)).
\end{aligned}$$

Therefore, $\mathbb{E}[\mathcal{E}_1(\mathbf{w}_{M+N})] < \mathbb{E}[\mathcal{E}_1(\mathbf{w}_N^{\mathrm{real}})]$ if and only if

$$\|\mathbb{E}[\mathbf{w}_M] - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2 + \mathrm{tr}(\mathbf{K}_N \,\mathrm{Cov}(\mathbf{w}_M)) < \|\mathbf{w}_0 - \mathbf{w}_1^*\|_{\mathbf{K}_N}^2,$$

which proves Theorem 8. ∎

Proof. [Proof of Corollary 9] Since $x_s^2 = 1$, the s-th synthetic update with stepsize $1/s$ satisfies

$$\begin{aligned}
w_s &= w_{s-1} - \frac{1}{s}\big(w_{s-1}x_s - (\theta + \delta)x_s - \xi_{2,s}\big)x_s \\
&= \left(1 - \frac{1}{s}\right) w_{s-1} + \frac{1}{s}\big(\theta + \delta + \xi_{2,s}x_s\big).
\end{aligned}$$

Induction over s therefore gives

$$w_M = \theta + \delta + \frac{1}{M}\sum_{s=1}^{M} \xi_{2,s}x_s.$$

Because the synthetic noises are independent, centered, and independent of the Rademacher covariates,

$$\mathbb{E}[w_M] = \theta + \delta, \qquad \mathrm{Var}(w_M) = \frac{\sigma_2^2}{M}. \tag{79}$$

In one dimension, $\mathbf{H} = \mathbb{E}[x^2] = 1$. Moreover, every real-stage update has multiplicative operator $1 - \gamma_t x_t^2 = 1 - \gamma_t$. Thus,

$$\Phi_N = \prod_{t=1}^{N}(1 - \gamma_t), \qquad \mathbf{K}_N = \left(\prod_{t=1}^{N}(1 - \gamma_t)\right)^2.$$

Applying Theorem 8 and using (79), we obtain

$$2\left(\mathbb{E}[\mathcal{E}_1(w_N^{\text{real}})] - \mathbb{E}[\mathcal{E}_1(w_{M+N})]\right) = \left(\prod_{t=1}^{N}(1-\gamma_t)\right)^2 \left[(w_0-\theta)^2 - \delta^2 - \frac{\sigma_2^2}{M}\right].$$

Since $0 < \gamma_t < 1$, the multiplicative factor is strictly positive. Therefore, two-stage training has strictly smaller expected risk if and only if

$$\delta^2 + \frac{\sigma_2^2}{M} < (w_0-\theta)^2.$$

If $\delta^2 \geq (w_0-\theta)^2$, this inequality cannot hold for any M. Otherwise, it is equivalent to

$$M > \frac{\sigma_2^2}{(w_0-\theta)^2 - \delta^2},$$

which completes the proof. ∎

H constant stepsize SGD with iterate averaging

In this section, we also denote total sample size as N, where synthetic sample size is pN and real sample size is $(1-p)N$. This light abuse of notation will achieve more simplicity and clarity.

H.1 Upper bound analysis

We use the same definition in Appendix C.

$$\mathbf{B}_t = \mathbb{E}[\boldsymbol{\eta}_t^{\text{bias}} \otimes \boldsymbol{\eta}_t^{\text{bias}}], \quad \mathbf{V}_t = \mathbb{E}[\boldsymbol{\eta}_t^{\text{var}} \otimes \boldsymbol{\eta}_t^{\text{var}}],$$
$$\mathbf{D}_t = \mathbb{E}[\boldsymbol{\eta}_t^{\text{drift}} \otimes \boldsymbol{\eta}_t^{\text{drift}}], \quad \mathbf{F}_t = \mathbb{E}[\boldsymbol{\eta}_t^{\text{fluct}} \otimes \boldsymbol{\eta}_t^{\text{fluct}}].$$

Using the definition of $\mathbf{B}_t, \mathbf{V}_t, \mathbf{D}_t, \mathbf{F}_t$, we can then decompose Excess Risk using Cauchy–Schwarz Inequality:

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(\bar{\mathbf{w}}_N) - \mathcal{R}(\mathbf{w}_1^*)] &= \frac{1}{2}\big\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N \otimes \bar{\boldsymbol{\eta}}_N]\big\rangle \\
&= \frac{1}{2}\big\langle \mathbf{H}, \mathbb{E}[(\bar{\boldsymbol{\eta}}_N^{\text{bias}} + \bar{\boldsymbol{\eta}}_N^{\text{var}} + \bar{\boldsymbol{\eta}}_N^{\text{drift}} + \bar{\boldsymbol{\eta}}_N^{\text{fluct}}) \otimes (\bar{\boldsymbol{\eta}}_N^{\text{bias}} + \bar{\boldsymbol{\eta}}_N^{\text{var}} + \bar{\boldsymbol{\eta}}_N^{\text{drift}} + \bar{\boldsymbol{\eta}}_N^{\text{fluct}})]\big\rangle \\
&\leq 2\big\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\text{bias}} \otimes \bar{\boldsymbol{\eta}}_N^{\text{bias}}]\big\rangle + 2\big\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\text{var}} \otimes \bar{\boldsymbol{\eta}}_N^{\text{var}}]\big\rangle \\
&\quad + 2\big\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\text{drift}} \otimes \bar{\boldsymbol{\eta}}_N^{\text{drift}}]\big\rangle + 2\big\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\text{fluct}} \otimes \bar{\boldsymbol{\eta}}_N^{\text{fluct}}]\big\rangle \\
&= 2\big\langle \mathbf{H}, \bar{\mathbf{B}}_N\big\rangle + 2\big\langle \mathbf{H}, \bar{\mathbf{V}}_N\big\rangle + 2\big\langle \mathbf{H}, \bar{\mathbf{D}}_N\big\rangle + 2\big\langle \mathbf{H}, \bar{\mathbf{F}}_N\big\rangle.
\end{aligned}$$

Bias upper bound Using the defined operators, the update rule of the iterates imply the following recursive form of $\mathbf{B}_t$:

$$\mathbf{B}_t = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{B}_{t-1}, \quad \mathbf{B}_0 = \boldsymbol{\eta}_0 \otimes \boldsymbol{\eta}_0, \tag{80}$$

Note that this bias term $\big\langle \mathbf{H}, \bar{\mathbf{B}}_N \big\rangle$ is the same as that in [73], so we can apply Lemma B.11 in [73] to bound it directly:

Lemma 20 (A bias upper bound for Avg-SGD) *Suppose Assumptions 1 and 2 hold. Let* $N_{\texttt{eff}} = N$. *Consider* (80). *Suppose* $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$. *We have*

$$\frac{1}{2}\big\langle \mathbf{H}, \bar{\mathbf{B}}_N \big\rangle \leq \frac{1}{\gamma^2 N_{\texttt{eff}}^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{0:k^*}^{-1}} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} + \frac{2\alpha\big(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + N_{\texttt{eff}}\gamma\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big)}{N_{\texttt{eff}}\gamma(1 - \gamma\alpha\operatorname{tr}(\mathbf{H}))} \cdot \frac{D_{\texttt{eff}}}{N_{\texttt{eff}}}$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{\gamma N_{\texttt{eff}}}\}$ *and* $D_{\texttt{eff}} := k^* + \gamma^2 N_{\texttt{eff}}^2 \sum_{i>k^*} \lambda_i^2$.

Proof. See proof of Lemma B.11 in [73]. ∎

Variance upper bound Using the defined operators, the update rule of the variance process implies

$$\mathbf{V}_t = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{V}_{t-1} + \gamma^2 \mathbf{\Sigma}_t, \qquad \mathbf{V}_0 = \mathbf{0}, \tag{81}$$

where

$$\mathbf{\Sigma}_t := \mathbb{E}[\xi_t^2 \mathbf{x}_t \mathbf{x}_t^\top].$$

Since $\mathbf{x}_t$ is independent of z_t, conditioning on the source indicator gives

$$\mathbf{\Sigma}_t = \mathbb{E}\big[\mathbb{E}[\xi_t^2 \mathbf{x}_t \mathbf{x}_t^\top \mid z_t]\big] = (1-p)\mathbf{\Sigma}_1 + p\mathbf{\Sigma}_2 \preceq \big((1-p)\sigma_1^2 + p\sigma_2^2\big)\mathbf{H}.$$

Define

$$\bar{\sigma}^2 := (1-p)\sigma_1^2 + p\sigma_2^2.$$

Then (81) has exactly the same form as the variance recursion for constant-stepsize SGD with iterate averaging analyzed in [73], with noise level upper bounded by σ_{mix}^2. Therefore, by applying the averaged-SGD variance bound in [73], we obtain the following result. The corresponding sharp variance characterization for iterate averaging is stated in Theorem 2.1 and developed in Lemma B.6 of [73].

Lemma 21 (A variance upper bound for Avg-SGD) *Suppose Assumptions 1, 2 and 3 hold. Consider* (81)*. Let $N_{\mathtt{eff}} = N$ and $\bar{\sigma}^2 = (1-p)\sigma_1^2 + p\sigma_2^2$. Suppose $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$. Then*

$$\frac{1}{2}\big\langle \mathbf{H}, \bar{\mathbf{V}}_N \big\rangle \le \frac{\bar{\sigma}^2}{1 - \gamma\alpha \operatorname{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}},$$

where $k^* = \max\left\{k : \lambda_k \ge \frac{1}{\gamma N_{\mathtt{eff}}}\right\}$, and $D_{\mathtt{eff}} := k^* + \gamma^2 N_{\mathtt{eff}}^2 \sum_{i>k^*} \lambda_i^2$.

Proof. The recursion (81) is identical to the variance recursion of constant-stepsize SGD with iterate averaging, except that the noise covariance is now $\mathbf{\Sigma}_t$ instead of a single-source covariance matrix. Since

$$\mathbf{\Sigma}_t \preceq \bar{\sigma}^2 \mathbf{H} \qquad \text{for all } t,$$

the proof of the variance upper bound in [73] applies verbatim with σ^2 replaced by σ_{mix}^2. Hence

$$\frac{1}{2}\big\langle \mathbf{H}, \bar{\mathbf{V}}_N \big\rangle \le \frac{\bar{\sigma}^2}{1 - \gamma\alpha \operatorname{tr}(\mathbf{H})} \left(\frac{k^*}{N_{\mathtt{eff}}} + N_{\mathtt{eff}} \gamma^2 \sum_{i>k^*} \lambda_i^2 \right) = \frac{\bar{\sigma}^2}{1 - \gamma\alpha \operatorname{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}.$$

This completes the proof. ∎

Drift upper bound We now analyze the drift contribution

$$\frac{1}{2}\big\langle \mathbf{H}, \bar{\mathbf{D}}_N \big\rangle = \frac{1}{2}\big\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{drift}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{drift}}] \big\rangle, \qquad \bar{\boldsymbol{\eta}}_N^{\mathrm{drift}} := \frac{1}{N} \sum_{t=0}^{N-1} \boldsymbol{\eta}_t^{\mathrm{drift}}.$$

Lemma 22 (A drift upper bound for Avg-SGD) *Suppose Assumptions 1, 2 and 3 hold. Let $N_{\mathtt{eff}} = N$. Suppose $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$. Then*

$$\begin{aligned}\frac{1}{2}\big\langle \mathbf{H}, \bar{\mathbf{D}}_N \big\rangle \le \frac{p^2}{2} \left\| \left(\mathbf{I} - \frac{1}{\gamma N_{\mathtt{eff}}} \mathbf{H}^{-1} \big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N_{\mathtt{eff}}} \big) \right) \boldsymbol{\delta} \right\|_{\mathbf{H}}^2 \\ + \frac{\alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1 - \gamma\alpha \operatorname{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}},\end{aligned}$$

where

$$k^* = \max\left\{k : \lambda_k \ge \frac{1}{\gamma N_{\mathtt{eff}}}\right\}, \qquad D_{\mathtt{eff}} := k^* + \gamma^2 N_{\mathtt{eff}}^2 \sum_{i>k^*} \lambda_i^2.$$

Proof. Recall that the drift process satisfies

$$\boldsymbol{\eta}_t^{\mathrm{drift}} = (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top) \boldsymbol{\eta}_{t-1}^{\mathrm{drift}} + \gamma p \mathbf{H} \boldsymbol{\delta}, \qquad \boldsymbol{\eta}_0^{\mathrm{drift}} = \mathbf{0}.$$

We first separate the drift process into its mean part and centered fluctuation part. Define

$$\mathbf{m}_t := \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{drift}}].$$

Since $\mathbf{x}_t$ is independent of $\boldsymbol{\eta}_{t-1}^{\mathrm{drift}}$ and $\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top] = \mathbf{H}$, taking expectation on both sides yields

$$\mathbf{m}_t = (\mathbf{I} - \gamma\mathbf{H})\mathbf{m}_{t-1} + \gamma p\mathbf{H}\boldsymbol{\delta}, \qquad \mathbf{m}_0 = \mathbf{0}. \tag{82}$$

Let

$$\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}} := \boldsymbol{\eta}_t^{\mathrm{drift}} - \mathbf{m}_t.$$

Subtracting (82) from the recursion of $\boldsymbol{\eta}_t^{\mathrm{drift}}$ gives

$$\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}} = (\mathbf{I} - \gamma\mathbf{x}_t\mathbf{x}_t^\top)\widetilde{\boldsymbol{\eta}}_{t-1}^{\mathrm{drift}} + \gamma(\mathbf{H} - \mathbf{x}_t\mathbf{x}_t^\top)\mathbf{m}_{t-1}, \qquad \widetilde{\boldsymbol{\eta}}_0^{\mathrm{drift}} = \mathbf{0}. \tag{83}$$

Moreover, by construction,

$$\mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}}] = \mathbf{0}, \qquad \forall t \geq 0.$$

Now define the averaged quantities

$$\bar{\mathbf{m}}_N := \frac{1}{N}\sum_{t=0}^{N-1}\mathbf{m}_t, \qquad \bar{\widetilde{\boldsymbol{\eta}}}_N^{\mathrm{drift}} := \frac{1}{N}\sum_{t=0}^{N-1}\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}}.$$

Then

$$\bar{\boldsymbol{\eta}}_N^{\mathrm{drift}} = \bar{\mathbf{m}}_N + \bar{\widetilde{\boldsymbol{\eta}}}_N^{\mathrm{drift}}.$$

Since $\mathbb{E}[\bar{\widetilde{\boldsymbol{\eta}}}_N^{\mathrm{drift}}] = \mathbf{0}$, the cross term vanishes and

$$\bar{\mathbf{D}}_N = \bar{\mathbf{m}}_N \otimes \bar{\mathbf{m}}_N + \bar{\widetilde{\mathbf{D}}}_N, \qquad \bar{\widetilde{\mathbf{D}}}_N := \mathbb{E}\Big[\bar{\widetilde{\boldsymbol{\eta}}}_N^{\mathrm{drift}} \otimes \bar{\widetilde{\boldsymbol{\eta}}}_N^{\mathrm{drift}}\Big]. \tag{84}$$

Therefore,

$$\frac{1}{2}\big\langle \mathbf{H}, \bar{\mathbf{D}}_N\big\rangle = \frac{1}{2}\|\bar{\mathbf{m}}_N\|_{\mathbf{H}}^2 + \frac{1}{2}\big\langle \mathbf{H}, \bar{\widetilde{\mathbf{D}}}_N\big\rangle. \tag{85}$$

Unrolling (82), we obtain

$$\mathbf{m}_t = p\big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^t\big)\boldsymbol{\delta}, \qquad t \geq 0.$$

Hence

$$\begin{aligned}\bar{\mathbf{m}}_N &= \frac{p}{N}\sum_{t=0}^{N-1}\big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^t\big)\boldsymbol{\delta} \\ &= p\left(\mathbf{I} - \frac{1}{N}\sum_{t=0}^{N-1}(\mathbf{I} - \gamma\mathbf{H})^t\right)\boldsymbol{\delta}.\end{aligned}$$

Using the matrix geometric-series identity

$$\sum_{t=0}^{N-1}(\mathbf{I} - \gamma\mathbf{H})^t = (\gamma\mathbf{H})^{-1}\big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^N\big),$$

we get

$$\bar{\mathbf{m}}_N = p\left(\mathbf{I} - \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^N\big)\right)\boldsymbol{\delta}. \tag{86}$$

Therefore,

$$\frac{1}{2}\|\bar{\mathbf{m}}_N\|_{\mathbf{H}}^2 = \frac{p^2}{2}\left\|\left(\mathbf{I} - \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^N\big)\right)\boldsymbol{\delta}\right\|_{\mathbf{H}}^2. \tag{87}$$

We next bound

$$\frac{1}{2}\big\langle \mathbf{H}, \bar{\widetilde{\mathbf{D}}}_N\big\rangle.$$

Define

$$\widetilde{\mathbf{D}}_t := \mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}} \otimes \widetilde{\boldsymbol{\eta}}_t^{\mathrm{drift}}].$$

From (83), we obtain the recursion

$$\widetilde{\mathbf{D}}_t = (\mathcal{I} - \gamma\mathcal{T}) \circ \widetilde{\mathbf{D}}_{t-1} + \gamma^2 \widetilde{\boldsymbol{\Sigma}}_t^{\mathrm{drift}}, \qquad \widetilde{\mathbf{D}}_0 = \mathbf{0}, \tag{88}$$

where

$$\widetilde{\boldsymbol{\Sigma}}_t^{\mathrm{drift}} := \mathbb{E}\big[\big((\mathbf{H} - \mathbf{x}_t\mathbf{x}_t^\top)\mathbf{m}_{t-1}\big) \otimes \big((\mathbf{H} - \mathbf{x}_t\mathbf{x}_t^\top)\mathbf{m}_{t-1}\big)\big].$$

Indeed, the cross term vanishes because

$$\mathbb{E}[(\mathbf{H} - \mathbf{x}_t\mathbf{x}_t^\top)\mathbf{m}_{t-1}] = (\mathbf{H} - \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top])\mathbf{m}_{t-1} = \mathbf{0}.$$

Let

$$\mathbf{A}_t := \mathbf{m}_{t-1} \otimes \mathbf{m}_{t-1}.$$

Then

$$\begin{aligned}
\widetilde{\boldsymbol{\Sigma}}_t^{\mathrm{drift}} &= \mathbb{E}\big[(\mathbf{H} - \mathbf{x}_t\mathbf{x}_t^\top)\mathbf{A}_t(\mathbf{H} - \mathbf{x}_t\mathbf{x}_t^\top)\big] \\
&= \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top \mathbf{A}_t \mathbf{x}_t\mathbf{x}_t^\top] - \mathbf{H}\mathbf{A}_t\mathbf{H} \\
&= (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{A}_t.
\end{aligned}$$

Since $\mathcal{M} - \widetilde{\mathcal{M}}$ is a PSD mapping, we have $\widetilde{\boldsymbol{\Sigma}}_t^{\mathrm{drift}} \succeq \mathbf{0}$. On the other hand, by Assumption 22,

$$\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top \mathbf{A}_t \mathbf{x}_t\mathbf{x}_t^\top] = \mathcal{M} \circ \mathbf{A}_t \preceq \alpha\ \mathrm{tr}(\mathbf{H}\mathbf{A}_t)\,\mathbf{H} = \alpha\,\|\mathbf{m}_{t-1}\|_{\mathbf{H}}^2\,\mathbf{H}.$$

Therefore,

$$\widetilde{\boldsymbol{\Sigma}}_t^{\mathrm{drift}} \preceq \alpha\,\|\mathbf{m}_{t-1}\|_{\mathbf{H}}^2\,\mathbf{H}. \tag{89}$$

Next, by the explicit formula for $\mathbf{m}_t$,

$$\mathbf{m}_{t-1} = p\big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{t-1}\big)\boldsymbol{\delta},$$

and hence

$$\begin{aligned}
\|\mathbf{m}_{t-1}\|_{\mathbf{H}}^2 &= p^2 \sum_i \lambda_i \left(1 - (1 - \gamma\lambda_i)^{t-1}\right)^2 \delta_i^2 \\
&\le p^2 \sum_i \lambda_i \delta_i^2 = p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2.
\end{aligned}$$

Substituting this into (89) gives the uniform bound

$$\widetilde{\boldsymbol{\Sigma}}_t^{\mathrm{drift}} \preceq \alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2\,\mathbf{H}, \qquad \forall t. \tag{90}$$

Now (88) has exactly the same form as the variance recursion for constant-stepsize SGD with iterate averaging, with noise level upper bounded by

$$\sigma_{\mathrm{drift}}^2 := \alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2.$$

Therefore, applying Lemma 21 with $\bar{\sigma}^2$ replaced by $\sigma_{\mathrm{drift}}^2$, we obtain

$$\frac{1}{2}\big\langle \mathbf{H}, \bar{\widetilde{\mathbf{D}}}_N \big\rangle \le \frac{\alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1 - \gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N}. \tag{91}$$

Finally, combining (85), (87) and (91), we conclude that

$$\begin{aligned}
\frac{1}{2}\big\langle \mathbf{H}, \bar{\mathbf{D}}_N \big\rangle \le\ & \frac{p^2}{2} \left\| \left(\mathbf{I} - \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^N\big)\right)\boldsymbol{\delta} \right\|_{\mathbf{H}}^2 \\
&+ \frac{\alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1 - \gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N}.
\end{aligned}$$

This completes the proof. ∎

Fluctuation upper bound Recall that the fluctuation process satisfies

$$\boldsymbol{\eta}_t^{\mathrm{fluct}} = (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{fluct}} + \gamma z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} - \gamma p \mathbf{H}\boldsymbol{\delta}, \qquad \boldsymbol{\eta}_0^{\mathrm{fluct}} = \mathbf{0}.$$

We decompose the driving term into two parts:

$$z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} - p\mathbf{H}\boldsymbol{\delta} = z_t(\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta} + (z_t - p)\mathbf{H}\boldsymbol{\delta}.$$

Accordingly, define

$$\boldsymbol{\eta}_t^{\mathrm{fluct}} = \boldsymbol{\eta}_t^{\mathrm{cov}} + \boldsymbol{\eta}_t^{\mathrm{sch}},$$

where

$$\begin{cases} \boldsymbol{\eta}_t^{\mathrm{cov}} = (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{cov}} + \gamma z_t(\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}, \\ \boldsymbol{\eta}_0^{\mathrm{cov}} = \mathbf{0}, \end{cases} \qquad \begin{cases} \boldsymbol{\eta}_t^{\mathrm{sch}} = (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{sch}} + \gamma(z_t - p)\mathbf{H}\boldsymbol{\delta}, \\ \boldsymbol{\eta}_0^{\mathrm{sch}} = \mathbf{0}. \end{cases}$$

Thus, for the averaged iterates,

$$\bar{\boldsymbol{\eta}}_N^{\mathrm{fluct}} = \bar{\boldsymbol{\eta}}_N^{\mathrm{cov}} + \bar{\boldsymbol{\eta}}_N^{\mathrm{sch}}.$$

Hence, by $(a+b)^2 \le 2a^2 + 2b^2$,

$$\frac{1}{2}\big\langle \mathbf{H}, \bar{\mathbf{F}}_N \big\rangle = \frac{1}{2}\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_N^{\mathrm{fluct}}\|_{\mathbf{H}}^2\big] \le \mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_N^{\mathrm{cov}}\|_{\mathbf{H}}^2\big] + \mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_N^{\mathrm{sch}}\|_{\mathbf{H}}^2\big]. \tag{92}$$

Define $\mathbf{F}_t^{\mathrm{cov}} := \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{cov}} \otimes \boldsymbol{\eta}_t^{\mathrm{cov}}]$. Then $\mathbf{F}_t^{\mathrm{cov}}$ satisfies the recursion

$$\mathbf{F}_t^{\mathrm{cov}} = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{F}_{t-1}^{\mathrm{cov}} + \gamma^2 \boldsymbol{\Sigma}_t^{\mathrm{cov}}, \qquad \mathbf{F}_0^{\mathrm{cov}} = \mathbf{0}, \tag{93}$$

where

$$\boldsymbol{\Sigma}_t^{\mathrm{cov}} := \mathbb{E}\big[\big(z_t(\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big) \otimes \big(z_t(\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big)\big].$$

Indeed, the cross term vanishes as we have proved in last-iterate analysis.

Next, since $z_t \in \{0,1\}$ and $\mathbb{E}[z_t] = p$, we have

$$\begin{aligned} \boldsymbol{\Sigma}_t^{\mathrm{cov}} &= \mathbb{E}\big[z_t^2\big((\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big) \otimes \big((\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big)\big] \\ &= p\,\mathbb{E}\big[\big((\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big) \otimes \big((\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big)\big]. \end{aligned}$$

Let $\mathbf{A} := \boldsymbol{\delta} \otimes \boldsymbol{\delta}$. Then

$$\begin{aligned} \mathbb{E}\big[\big((\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big) \otimes \big((\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big)\big] &= \mathbb{E}\big[(\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\mathbf{A}(\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{H})\big] \\ &= \mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top \mathbf{A} \mathbf{x}_t \mathbf{x}_t^\top] - \mathbf{H}\mathbf{A}\mathbf{H} \\ &= (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{A}. \end{aligned}$$

Since $\mathcal{M} - \widetilde{\mathcal{M}}$ is a PSD mapping, and by Assumption 2,

$$\mathcal{M} \circ \mathbf{A} = \mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top \mathbf{A} \mathbf{x}_t \mathbf{x}_t^\top] \preceq \alpha\ \mathrm{tr}(\mathbf{H}\mathbf{A})\,\mathbf{H} = \alpha \|\boldsymbol{\delta}\|_{\mathbf{H}}^2\,\mathbf{H},$$

we obtain the upper bound

$$\boldsymbol{\Sigma}_t^{\mathrm{cov}} \preceq p\alpha \|\boldsymbol{\delta}\|_{\mathbf{H}}^2\,\mathbf{H}. \tag{94}$$

Therefore, the recursion (93) has exactly the same form as the variance recursion for constant-stepsize SGD with iterate averaging, with noise level upper bounded by

$$\sigma_{\mathrm{cov}}^2 := p\alpha \|\boldsymbol{\delta}\|_{\mathbf{H}}^2.$$

Applying Lemma 21 with $\bar{\sigma}^2$ replaced by σ_{cov}^2, we get

$$\frac{1}{2}\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_N^{\mathrm{cov}}\|_{\mathbf{H}}^2\big] = \frac{1}{2}\big\langle \mathbf{H}, \bar{\mathbf{F}}_N^{\mathrm{cov}} \big\rangle \le \frac{p\alpha \|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1 - \gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N}. \tag{95}$$

We now bound the second component

$$\boldsymbol{\eta}_t^{\mathrm{sch}} = (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{sch}} + \gamma(z_t - p)\mathbf{H}\boldsymbol{\delta}, \qquad \boldsymbol{\eta}_0^{\mathrm{sch}} = \mathbf{0}.$$

Its averaged iterate is

$$\bar{\boldsymbol{\eta}}_N^{\mathrm{sch}} := \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\eta}_t^{\mathrm{sch}}.$$

Then we use the following lemma to bound $\langle \mathbf{H}, \bar{\boldsymbol{\eta}}_N^{\mathrm{sch}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{sch}} \rangle$.

Lemma 23 (A schedule-fluctuation upper bound for Avg-SGD) *Suppose Assumptions 1 and 2 hold. Assume the source indicators $(z_1, \dots, z_N)$ follow the fixed-budget model with exactly m ones, and let $p := \frac{m}{N}$. Consider the schedule fluctuation recursion*

$$\boldsymbol{\eta}_t^{\mathrm{sch}} = (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\mathrm{sch}} + \gamma(z_t - p)\mathbf{H}\boldsymbol{\delta}, \qquad \boldsymbol{\eta}_0^{\mathrm{sch}} = \mathbf{0},$$

and define its averaged iterate by $\bar{\boldsymbol{\eta}}_N^{\mathrm{sch}} := \frac{1}{N}\sum_{t=1}^N \boldsymbol{\eta}_t^{\mathrm{sch}}$. Suppose $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$. Let

$$k^* := \max\left\{k : \lambda_k \geq \frac{1}{\gamma N}\right\}, \qquad D_{\mathtt{eff}} := k^* + \gamma^2 N^2 \sum_{i>k^*} \lambda_i^2.$$

Then

$$\frac{1}{2}\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_N^{\mathrm{sch}}\|_{\mathbf{H}}^2\big] \leq \frac{p(1-p)\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{N-1}\left(1 + \frac{2\alpha\gamma \operatorname{tr}(\mathbf{H})}{1 - \gamma\alpha \operatorname{tr}(\mathbf{H})} D_{\mathtt{eff}}\right). \tag{96}$$

Proof. Let

$$\mathbf{B}_t := \mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top.$$

By repeated substitution of the recursion

$$\boldsymbol{\eta}_t^{\mathrm{sch}} = \mathbf{B}_t \boldsymbol{\eta}_{t-1}^{\mathrm{sch}} + \gamma(z_t - p)\mathbf{H}\boldsymbol{\delta}, \qquad \boldsymbol{\eta}_0^{\mathrm{sch}} = 0,$$

we obtain, for every $t \in \{1, \dots, N\}$,

$$\boldsymbol{\eta}_t^{\mathrm{sch}} = \gamma \sum_{s=1}^t \left(\prod_{j=s+1}^t \mathbf{B}_j\right)(z_s - p)\mathbf{H}\boldsymbol{\delta}, \tag{97}$$

where the empty product is interpreted as the identity operator.

Therefore,

$$\begin{aligned}
\bar{\boldsymbol{\eta}}_N^{\mathrm{sch}} &= \frac{1}{N}\sum_{t=1}^N \boldsymbol{\eta}_t^{\mathrm{sch}} \\
&= \frac{\gamma}{N}\sum_{t=1}^N \sum_{s=1}^t \left(\prod_{j=s+1}^t \mathbf{B}_j\right)(z_s - p)\mathbf{H}\boldsymbol{\delta} \\
&= \sum_{s=1}^N (z_s - p)\mathbf{v}_s,
\end{aligned} \tag{98}$$

where

$$\mathbf{v}_s := \frac{\gamma}{N}\sum_{t=s}^N \left(\prod_{j=s+1}^t \mathbf{B}_j\right)\mathbf{H}\boldsymbol{\delta}. \tag{99}$$

Condition on the sample sequence $(\mathbf{x}_1, \dots, \mathbf{x}_N)$. Then $\mathbf{v}_1, \dots, \mathbf{v}_N$ are deterministic vectors. Since the feature sequence is independent of the fixed-budget schedule, the indicators $(z_1, \dots, z_N)$ still follow the same fixed-budget law conditional on $(\mathbf{x}_1, \dots, \mathbf{x}_N)$.

Under the fixed-budget model, for every s,

$$\mathbb{E}[z_s] = p, \qquad \mathbb{E}[(z_s - p)^2] = p(1-p),$$

and for every $s \neq r$,

$$\mathbb{E}[(z_s - p)(z_r - p)] = -\frac{p(1-p)}{N-1}.$$

Hence, conditioned on $(\mathbf{x}_1, \dots, \mathbf{x}_N)$,

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{s=1}^N (z_s - p)\mathbf{v}_s\right\|_{\mathbf{H}}^2 \,\middle|\, \mathbf{x}_1, \dots, \mathbf{x}_N\right] &= \sum_{s=1}^N \sum_{r=1}^N \mathbb{E}[(z_s - p)(z_r - p)]\langle \mathbf{v}_s, \mathbf{v}_r\rangle_{\mathbf{H}} \\
&= p(1-p)\sum_{s=1}^N \|\mathbf{v}_s\|_{\mathbf{H}}^2 - \frac{p(1-p)}{N-1}\sum_{s\neq r}\langle \mathbf{v}_s, \mathbf{v}_r\rangle_{\mathbf{H}}.
\end{aligned} \tag{100}$$

Let

$$\bar{\mathbf{v}} := \frac{1}{N}\sum_{s=1}^{N}\mathbf{v}_s.$$

Using the identity

$$\sum_{s=1}^{N}\|\mathbf{v}_s-\bar{\mathbf{v}}\|_{\mathbf{H}}^2=\sum_{s=1}^{N}\|\mathbf{v}_s\|_{\mathbf{H}}^2-\frac{1}{N}\left\|\sum_{s=1}^{N}\mathbf{v}_s\right\|_{\mathbf{H}}^2,$$

one checks that the right-hand side of (100) equals

$$\frac{Np(1-p)}{N-1}\sum_{s=1}^{N}\|\mathbf{v}_s-\bar{\mathbf{v}}\|_{\mathbf{H}}^2\le\frac{Np(1-p)}{N-1}\sum_{s=1}^{N}\|\mathbf{v}_s\|_{\mathbf{H}}^2.$$

Combining this with (98), and taking expectation over the samples, we obtain

$$\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_N^{\mathrm{sch}}\|_{\mathbf{H}}^2\big]\le\frac{Np(1-p)}{N-1}\sum_{s=1}^{N}\mathbb{E}\big[\|\mathbf{v}_s\|_{\mathbf{H}}^2\big]. \tag{101}$$

Fix $s\in\{1,\dots,N\}$ and let $n_s := N-s+1$. Define an auxiliary bias-only process initialized at $\gamma\mathbf{H}\boldsymbol{\delta}$:

$$\mathbf{u}_0^{(s)}:=\gamma\mathbf{H}\boldsymbol{\delta},\qquad \mathbf{u}_r^{(s)}:=\mathbf{B}_{s+r}\mathbf{u}_{r-1}^{(s)},\qquad r=1,\dots,n_s-1.$$

Then

$$\mathbf{u}_r^{(s)}=\left(\prod_{j=s+1}^{s+r}\mathbf{B}_j\right)\gamma\mathbf{H}\boldsymbol{\delta}.$$

Therefore,

$$\begin{aligned}\mathbf{v}_s&=\frac{\gamma}{N}\sum_{t=s}^{N}\left(\prod_{j=s+1}^{t}\mathbf{B}_j\right)\mathbf{H}\boldsymbol{\delta}\\&=\frac{1}{N}\sum_{r=0}^{n_s-1}\mathbf{u}_r^{(s)}=\frac{n_s}{N}\,\bar{\mathbf{u}}_{n_s}^{(s)},\end{aligned} \tag{102}$$

where

$$\bar{\mathbf{u}}_{n_s}^{(s)}:=\frac{1}{n_s}\sum_{r=0}^{n_s-1}\mathbf{u}_r^{(s)}.$$

Since the samples are i.i.d., the distribution of

$$(\mathbf{u}_0^{(s)},\dots,\mathbf{u}_{n_s-1}^{(s)})$$

coincides with the distribution of a length-n_s bias-only SGD trajectory with constant stepsize γ and initialization $\gamma\mathbf{H}\boldsymbol{\delta}$. Hence

$$\mathbb{E}[\|\mathbf{v}_s\|_{\mathbf{H}}^2]=\left(\frac{n_s}{N}\right)^2\mathbb{E}\big[\|\bar{\mathbf{u}}_{n_s}(\gamma\mathbf{H}\boldsymbol{\delta})\|_{\mathbf{H}}^2\big], \tag{103}$$

where $\bar{\mathbf{u}}_n(\gamma\mathbf{H}\boldsymbol{\delta})$ denotes the averaged bias-only SGD iterate of horizon n initialized at $\gamma\mathbf{H}\boldsymbol{\delta}$.

Substituting (103) into (101), and changing variables from s to $n=n_s$, yields

$$\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_N^{\mathrm{sch}}\|_{\mathbf{H}}^2\big]\le\frac{Np(1-p)}{N-1}\sum_{n=1}^{N}\left(\frac{n}{N}\right)^2\mathbb{E}\big[\|\bar{\mathbf{u}}_n(\gamma\mathbf{H}\boldsymbol{\delta})\|_{\mathbf{H}}^2\big]. \tag{104}$$

For each $n\in\{1,\dots,N\}$, let

$$k_n:=\max\left\{k:\lambda_k\ge\frac{1}{\gamma n}\right\},\qquad D_n:=k_n+n^2\gamma^2\sum_{i>k_n}\lambda_i^2.$$

Applying Lemma 20 with horizon n and initialization $\mathbf{u}_0 = \gamma\mathbf{H}\boldsymbol{\delta}$, we obtain

$$\frac{1}{2}\mathbb{E}\big[\|\bar{\mathbf{u}}_n(\gamma\mathbf{H}\boldsymbol{\delta})\|_{\mathbf{H}}^2\big] \le \frac{1}{\gamma^2 n^2}\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{0:k_n}^{-1}}^2 + \|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{k_n:\infty}}^2 + \frac{2\alpha\big(\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{I}_{0:k_n}}^2 + n\gamma\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{k_n:\infty}}^2\big)}{n\gamma(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))} \cdot \frac{D_n}{n}. \tag{105}$$

We now simplify each term in the eigenbasis of $\mathbf{H}$. Writing

$$\boldsymbol{\delta} = \sum_i \delta_i \mathbf{v}_i, \qquad \gamma\mathbf{H}\boldsymbol{\delta} = \gamma\sum_i \lambda_i\delta_i\mathbf{v}_i,$$

we have

$$\frac{1}{\gamma^2 n^2}\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{0:k_n}^{-1}}^2 = \frac{1}{n^2}\sum_{i\le k_n}\lambda_i\delta_i^2, \tag{106}$$

$$\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{k_n:\infty}}^2 = \gamma^2\sum_{i>k_n}\lambda_i^3\delta_i^2, \tag{107}$$

$$\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{I}_{0:k_n}}^2 = \gamma^2\sum_{i\le k_n}\lambda_i^2\delta_i^2. \tag{108}$$

For the first two terms, since $\lambda_i < 1/(\gamma n)$ for every $i > k_n$,

$$\gamma^2\lambda_i^3 \le \frac{\lambda_i}{n^2}.$$

Therefore, by (106) and (107),

$$\begin{aligned}\frac{1}{\gamma^2 n^2}\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{0:k_n}^{-1}}^2 + \|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{k_n:\infty}}^2 &\le \frac{1}{n^2}\sum_{i\le k_n}\lambda_i\delta_i^2 + \frac{1}{n^2}\sum_{i>k_n}\lambda_i\delta_i^2 \\ &= \frac{\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{n^2}.\end{aligned} \tag{109}$$

For the third term, using again $\lambda_i < 1/(\gamma n)$ for $i > k_n$, we have

$$n\gamma\lambda_i^3 \le \lambda_i^2.$$

Hence

$$\begin{aligned}\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{I}_{0:k_n}}^2 + n\gamma\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{k_n:\infty}}^2 &= \gamma^2\sum_{i\le k_n}\lambda_i^2\delta_i^2 + n\gamma^3\sum_{i>k_n}\lambda_i^3\delta_i^2 \\ &\le \gamma^2\sum_i\lambda_i^2\delta_i^2 \\ &\le \gamma^2\,\mathrm{tr}(\mathbf{H})\sum_i\lambda_i\delta_i^2 = \gamma^2\,\mathrm{tr}(\mathbf{H})\|\boldsymbol{\delta}\|_{\mathbf{H}}^2.\end{aligned} \tag{110}$$

Substituting (109) and (110) into (105), we obtain

$$\frac{1}{2}\mathbb{E}\big[\|\bar{\mathbf{u}}_n(\gamma\mathbf{H}\boldsymbol{\delta})\|_{\mathbf{H}}^2\big] \le \frac{\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{n^2}\left(1 + \frac{2\alpha\gamma\,\mathrm{tr}(\mathbf{H})}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\,D_n\right). \tag{111}$$

Substituting (111) into (104) yields

$$\begin{aligned}\frac{1}{2}\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_N^{\mathrm{sch}}\|_{\mathbf{H}}^2\big] &\le \frac{Np(1-p)}{N-1}\sum_{n=1}^{N}\left(\frac{n}{N}\right)^2\cdot\frac{\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{n^2}\left(1 + \frac{2\alpha\gamma\,\mathrm{tr}(\mathbf{H})}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\,D_n\right) \\ &= \frac{p(1-p)\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{N(N-1)}\sum_{n=1}^{N}\left(1 + \frac{2\alpha\gamma\,\mathrm{tr}(\mathbf{H})}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\,D_n\right).\end{aligned} \tag{112}$$

It remains to compare D_n with the final effective dimension $D_{\mathtt{eff}}$. Observe that

$$D_n = k_n + n^2\gamma^2 \sum_{i>k_n} \lambda_i^2 = \sum_i \min\{1, n^2\gamma^2\lambda_i^2\}.$$

Since $n \le N$,

$$D_n \le \sum_i \min\{1, N^2\gamma^2\lambda_i^2\}.$$

Now let

$$k^* := \max\left\{k : \lambda_k \ge \frac{1}{\gamma N}\right\}.$$

Then

$$\sum_i \min\{1, N^2\gamma^2\lambda_i^2\} \le k^* + \gamma^2 N^2 \sum_{i>k^*} \lambda_i^2 = D_{\mathtt{eff}}.$$

Hence

$$D_n \le D_{\mathtt{eff}}, \qquad \forall n \le N.$$

Applying this to (112) gives

$$\begin{aligned}\frac{1}{2}\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_N^{\mathrm{sch}}\|_{\mathbf{H}}^2\big] &\le \frac{p(1-p)\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{N(N-1)} \sum_{n=1}^{N} \left(1 + \frac{2\alpha\gamma\,\mathrm{tr}(\mathbf{H})}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\, D_{\mathtt{eff}}\right) \\ &= \frac{p(1-p)\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{N-1} \left(1 + \frac{2\alpha\gamma\,\mathrm{tr}(\mathbf{H})}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\, D_{\mathtt{eff}}\right),\end{aligned}$$

which proves (96). ∎

An upper bound for excess risk. Combining the bounds for the bias, variance, drift, and fluctuation terms derived above, we obtain the following upper bound on the excess risk.

Theorem 24 (Excess risk upper bound for constant-stepsize SGD with iterate averaging) *Suppose Assumptions 1, 2, and 3 hold. Let $\bar{\sigma}^2 := (1-p)\sigma_1^2 + p\sigma_2^2$. Assume $\gamma < 1/(\alpha\,\mathrm{tr}(\mathbf{H}))$. Then*

$$\begin{aligned}\mathbb{E}[\mathcal{R}(\bar{\mathbf{w}}_N) - \mathcal{R}(\mathbf{w}_1^*)] \lesssim\ & \frac{1}{\gamma^2 N^2}\|\mathbf{w}_0 - \mathbf{w}_1^*\|_{\mathbf{H}_{0:k^*}^{-1}}^2 + \|\mathbf{w}_0 - \mathbf{w}_1^*\|_{\mathbf{H}_{k^*:\infty}}^2 \\ &+ \big(\alpha\|\mathbf{w}_0 - \mathbf{w}_1^*\|_{\mathbf{H}}^2 + \bar{\sigma}^2 + \alpha p\|\boldsymbol{\delta}\|_{\mathbf{H}}^2\big) \cdot \frac{D_{\mathtt{eff}}}{N} + \frac{p(1-p)\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{N}, \\ &+ p^2 \left\| \left(\mathbf{I} - \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N\big)\right)\boldsymbol{\delta}\right\|_{\mathbf{H}}^2\end{aligned}$$

where $k^* = \max\left\{k : \lambda_k \ge \frac{1}{\gamma N}\right\}$, and $D_{\mathtt{eff}} := k^* + \gamma^2 N^2 \sum_{i>k^*} \lambda_i^2$.

H.2 Lower bound analysis

The starting point is again the error recursion

$$\boldsymbol{\eta}_t = (\mathbf{I} - \gamma\mathbf{x}_t\mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1} + \gamma z_t\mathbf{x}_t\mathbf{x}_t^\top\boldsymbol{\delta} + \gamma\xi_t\mathbf{x}_t, \qquad \boldsymbol{\eta}_0 = \mathbf{w}_0 - \mathbf{w}_1^*, \tag{113}$$

where

$$\boldsymbol{\eta}_t = \mathbf{w}_t - \mathbf{w}_1^*.$$

Recall that the averaged iterate is

$$\bar{\mathbf{w}}_N := \frac{1}{N}\sum_{t=0}^{N-1} \mathbf{w}_t,$$

and correspondingly

$$\bar{\boldsymbol{\eta}}_N := \frac{1}{N}\sum_{t=0}^{N-1} \boldsymbol{\eta}_t = \bar{\mathbf{w}}_N - \mathbf{w}_1^*.$$

By definition of the excess risk,

$$\mathbb{E}[\mathcal{R}(\bar{\mathbf{w}}_N) - \mathcal{R}(\mathbf{w}_1^*)] = \frac{1}{2}\left\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N \otimes \bar{\boldsymbol{\eta}}_N] \right\rangle.$$

To derive a lower bound, we center the averaged error around its mean. Define

$$\bar{\boldsymbol{\mu}}_N := \mathbb{E}[\bar{\boldsymbol{\eta}}_N], \qquad \widetilde{\bar{\boldsymbol{\eta}}}_N := \bar{\boldsymbol{\eta}}_N - \bar{\boldsymbol{\mu}}_N.$$

Then

$$\bar{\boldsymbol{\eta}}_N = \bar{\boldsymbol{\mu}}_N + \widetilde{\bar{\boldsymbol{\eta}}}_N, \qquad \mathbb{E}[\widetilde{\bar{\boldsymbol{\eta}}}_N] = \mathbf{0}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[\bar{\boldsymbol{\eta}}_N \otimes \bar{\boldsymbol{\eta}}_N] &= \mathbb{E}\Big[(\bar{\boldsymbol{\mu}}_N + \widetilde{\bar{\boldsymbol{\eta}}}_N) \otimes (\bar{\boldsymbol{\mu}}_N + \widetilde{\bar{\boldsymbol{\eta}}}_N)\Big] \\
&= \bar{\boldsymbol{\mu}}_N \otimes \bar{\boldsymbol{\mu}}_N + \mathbb{E}[\widetilde{\bar{\boldsymbol{\eta}}}_N \otimes \widetilde{\bar{\boldsymbol{\eta}}}_N] + \bar{\boldsymbol{\mu}}_N \otimes \mathbb{E}[\widetilde{\bar{\boldsymbol{\eta}}}_N] + \mathbb{E}[\widetilde{\bar{\boldsymbol{\eta}}}_N] \otimes \bar{\boldsymbol{\mu}}_N \\
&= \bar{\boldsymbol{\mu}}_N \otimes \bar{\boldsymbol{\mu}}_N + \mathbb{E}[\widetilde{\bar{\boldsymbol{\eta}}}_N \otimes \widetilde{\bar{\boldsymbol{\eta}}}_N].
\end{aligned}$$

Hence

$$\mathbb{E}[\mathcal{R}(\bar{\mathbf{w}}_N) - \mathcal{R}(\mathbf{w}_1^*)] = \frac{1}{2}\|\bar{\boldsymbol{\mu}}_N\|_{\mathbf{H}}^2 + \frac{1}{2}\big\langle \mathbf{H}, \mathbb{E}[\widetilde{\bar{\boldsymbol{\eta}}}_N \otimes \widetilde{\bar{\boldsymbol{\eta}}}_N] \big\rangle. \tag{114}$$

Since the second term is nonnegative, we immediately obtain

$$\mathbb{E}[\mathcal{R}(\bar{\mathbf{w}}_N) - \mathcal{R}(\mathbf{w}_1^*)] \geq \frac{1}{2}\|\bar{\boldsymbol{\mu}}_N\|_{\mathbf{H}}^2. \tag{115}$$

We now characterize the mean trajectory of the SGD error process. Let $\boldsymbol{\mu}_t := \mathbb{E}[\boldsymbol{\eta}_t]$. Taking expectation on both sides of (113), and using

$$\mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top] = \mathbf{H}, \qquad \mathbb{E}[z_t] = p, \qquad \mathbb{E}[\xi_t \mathbf{x}_t] = \mathbf{0},$$

we obtain

$$\boldsymbol{\mu}_t = (\mathbf{I} - \gamma\mathbf{H})\boldsymbol{\mu}_{t-1} + \gamma p \mathbf{H}\boldsymbol{\delta}, \qquad \boldsymbol{\mu}_0 = \mathbf{w}_0 - \mathbf{w}_1^*.$$

Unrolling the recursion gives

$$\boldsymbol{\mu}_t = (\mathbf{I} - \gamma\mathbf{H})^t(\mathbf{w}_0 - \mathbf{w}_1^*) + p\sum_{s=1}^{t} \gamma(\mathbf{I} - \gamma\mathbf{H})^{t-s}\mathbf{H}\boldsymbol{\delta}. \tag{116}$$

Since $(\mathbf{I} - \gamma\mathbf{H})$ is a polynomial in $\mathbf{H}$, it commutes with $\mathbf{H}$. Therefore,

$$\sum_{s=1}^{t} \gamma(\mathbf{I} - \gamma\mathbf{H})^{t-s}\mathbf{H} = \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^t.$$

Substituting this identity into (116), we obtain the closed form

$$\boldsymbol{\mu}_t = (\mathbf{I} - \gamma\mathbf{H})^t(\mathbf{w}_0 - \mathbf{w}_1^*) + p\big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^t\big)\boldsymbol{\delta}.$$

Now average over $t = 0, \ldots, N-1$. Since

$$\bar{\boldsymbol{\mu}}_N = \mathbb{E}[\bar{\boldsymbol{\eta}}_N] = \frac{1}{N}\sum_{t=0}^{N-1} \boldsymbol{\mu}_t,$$

we have

$$\bar{\boldsymbol{\mu}}_N = \frac{1}{N}\sum_{t=0}^{N-1} (\mathbf{I} - \gamma\mathbf{H})^t(\mathbf{w}_0 - \mathbf{w}_1^*) + \frac{p}{N}\sum_{t=0}^{N-1} \big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^t\big)\boldsymbol{\delta}.$$

Using the matrix geometric-series identity

$$\sum_{t=0}^{N-1} (\mathbf{I} - \gamma\mathbf{H})^t = (\gamma\mathbf{H})^{-1}\big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^N\big),$$

we obtain

$$\bar{\boldsymbol{\mu}}_N = \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N\big)(\mathbf{w}_0 - \mathbf{w}_1^*) + p\left(\mathbf{I} - \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N\big)\right)\boldsymbol{\delta}.$$

Consequently,

$$\|\bar{\boldsymbol{\mu}}_N\|_{\mathbf{H}}^2 = \left\|\frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N\big)(\mathbf{w}_0 - \mathbf{w}_1^*) + p\left(\mathbf{I} - \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N\big)\right)\boldsymbol{\delta}\right\|_{\mathbf{H}}^2. \tag{117}$$

Combining (115) and (117), we arrive at the lower bound

$$\mathbb{E}[\mathcal{R}(\bar{\mathbf{w}}_N) - \mathcal{R}(\mathbf{w}_1^*)] \gtrsim \left\|\frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N\big)(\mathbf{w}_0 - \mathbf{w}_1^*) + p\left(\mathbf{I} - \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N\big)\right)\boldsymbol{\delta}\right\|_{\mathbf{H}}^2.$$

H.3 Strong model collapse behavior

Corollary 25 (Strong model collapse in Avg-SGD) *Consider averaged SGD with constant stepsize γ. Suppose Assumptions 1, 2 and 3 hold. Suppose $\gamma < 1/\alpha\,\mathrm{tr}(\mathbf{H})$ and $D_{\mathtt{eff}} = o(M+N)$. Further assume that $\|\mathbf{w}_0 - \mathbf{w}_1^*\|_2^2$ is finite. As the total sample size $M+N$ scales to infinity with a fixed synthetic proportion p,*

$$\lim_{M+N\to\infty} \mathbb{E}[\mathcal{E}_1(\bar{\mathbf{w}}_{M+N})] \eqsim p^2\|\boldsymbol{\delta}\|_{\mathbf{H}}^2.$$

Proof. Let

$$T := M+N, \qquad \mathbf{A}_T := \frac{1}{\gamma T}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^T\big).$$

Since the synthetic proportion $p = M/(M+N)$ is fixed, it suffices to study the limit as $T\to\infty$.

We first show that $\mathbf{A}_T$ vanishes in $\mathbf{H}$-norm. Writing the eigendecomposition

$$\mathbf{H} = \sum_i \lambda_i \mathbf{v}_i\mathbf{v}_i^\top,$$

we have

$$\mathbf{A}_T\mathbf{v}_i = \frac{1-(1-\gamma\lambda_i)^T}{\gamma T\lambda_i}\,\mathbf{v}_i.$$

Since $\gamma < 1/(\alpha\,\mathrm{tr}(\mathbf{H}))$, we have $0 \le \gamma\lambda_i \le \gamma\,\mathrm{tr}(\mathbf{H}) < 1$, hence

$$0 \le 1-\gamma\lambda_i < 1.$$

For every i with $\lambda_i > 0$, it follows that

$$\frac{1-(1-\gamma\lambda_i)^T}{\gamma T\lambda_i} \longrightarrow 0 \qquad \text{as } T\to\infty.$$

Moreover, using the elementary inequality

$$1-(1-a)^T \le Ta, \qquad 0 \le a \le 1,$$

we get

$$0 \le \frac{1-(1-\gamma\lambda_i)^T}{\gamma T\lambda_i} \le 1.$$

Therefore, for any $\mathbf{w}$ with $\|\mathbf{w}\|_{\mathbf{H}} < \infty$,

$$\|\mathbf{A}_T\mathbf{w}\|_{\mathbf{H}}^2 = \sum_i \lambda_i \left(\frac{1-(1-\gamma\lambda_i)^T}{\gamma T\lambda_i}\right)^2 \langle \mathbf{w}, \mathbf{v}_i\rangle^2 \longrightarrow 0$$

by dominated convergence.

We now turn to the upper bound. By the upper bound theorem for averaged SGD,

$$\mathbb{E}[\mathcal{E}_1(\bar{\mathbf{w}}_T)] \lesssim \underbrace{\frac{1}{\gamma^2 T^2}\|\mathbf{w}_0 - \mathbf{w}_1^*\|^2_{\mathbf{H}^{-1}_{0:k^*}}}_{(I)} + \underbrace{\|\mathbf{w}_0 - \mathbf{w}_1^*\|^2_{\mathbf{H}_{k^*:\infty}}}_{(II)}$$
$$+ \underbrace{\left(\alpha\|\mathbf{w}_0 - \mathbf{w}_1^*\|^2_{\mathbf{H}} + \bar{\sigma}^2 + \alpha p\|\boldsymbol{\delta}\|^2_{\mathbf{H}}\right)\cdot\frac{D_{\texttt{eff}}}{T}}_{(III)} + \underbrace{\frac{p(1-p)\|\boldsymbol{\delta}\|^2_{\mathbf{H}}}{T}}_{(IV)}$$
$$+ \underbrace{p^2\|(\mathbf{I} - \mathbf{A}_T)\boldsymbol{\delta}\|^2_{\mathbf{H}}}_{(V)},$$

where

$$k^* = \max\left\{k : \lambda_k \geq \frac{1}{\gamma T}\right\}, \qquad D_{\texttt{eff}} := k^* + \gamma^2 T^2 \sum_{i>k^*}\lambda_i^2.$$

By the assumption $D_{\texttt{eff}} = o(T)$, terms (III) and (IV) vanish as $T \to \infty$. For term (I), note that

$$\frac{1}{\gamma^2 T^2}\|\mathbf{w}_0 - \mathbf{w}_1^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} \leq \|\mathbf{w}_0 - \mathbf{w}_1^*\|_2^2 \cdot \frac{1}{\gamma T},$$

because for every $i \leq k^*$ we have $\lambda_i \geq 1/(\gamma T)$, hence $\lambda_i^{-1} \leq \gamma T$. Therefore $(I) \to 0$. For term (II), since $k^* \to \infty$ as $T \to \infty$ and $\|\mathbf{w}_0 - \mathbf{w}_1^*\|_{\mathbf{H}} < \infty$, we have

$$\|\mathbf{w}_0 - \mathbf{w}_1^*\|^2_{\mathbf{H}_{k^*:\infty}} = \sum_{i>k^*}\lambda_i\langle\mathbf{w}_0 - \mathbf{w}_1^*, \mathbf{v}_i\rangle^2 \longrightarrow 0.$$

Finally,

$$\|(\mathbf{I} - \mathbf{A}_T)\boldsymbol{\delta}\|_{\mathbf{H}} \to \|\boldsymbol{\delta}\|_{\mathbf{H}},$$

because

$$\|(\mathbf{I} - \mathbf{A}_T)\boldsymbol{\delta} - \boldsymbol{\delta}\|_{\mathbf{H}} = \|\mathbf{A}_T\boldsymbol{\delta}\|_{\mathbf{H}} \to 0.$$

Therefore,

$$\limsup_{T\to\infty}\mathbb{E}[\mathcal{E}_1(\bar{\mathbf{w}}_T)] \lesssim p^2\|\boldsymbol{\delta}\|^2_{\mathbf{H}}.$$

Next, we apply the lower bound obtained from the mean term:

$$\mathbb{E}[\mathcal{E}_1(\bar{\mathbf{w}}_T)] \gtrsim \|\mathbf{A}_T(\mathbf{w}_0 - \mathbf{w}_1^*) + p(\mathbf{I} - \mathbf{A}_T)\boldsymbol{\delta}\|^2_{\mathbf{H}}.$$

Rewrite the term inside the norm as

$$\mathbf{A}_T(\mathbf{w}_0 - \mathbf{w}_1^*) + p(\mathbf{I} - \mathbf{A}_T)\boldsymbol{\delta} = \mathbf{A}_T(\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}) + p\boldsymbol{\delta}.$$

Again using $\|\mathbf{A}_T\mathbf{w}\|_{\mathbf{H}} \to 0$ for every $\mathbf{w}$ with finite $\mathbf{H}$-norm, we obtain

$$\|\mathbf{A}_T(\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}) + p\boldsymbol{\delta}\|^2_{\mathbf{H}} \longrightarrow p^2\|\boldsymbol{\delta}\|^2_{\mathbf{H}}.$$

Hence

$$\liminf_{T\to\infty}\mathbb{E}[\mathcal{E}_1(\bar{\mathbf{w}}_T)] \gtrsim p^2\|\boldsymbol{\delta}\|^2_{\mathbf{H}}.$$

Combining the upper and lower bounds yields

$$\lim_{M+N\to\infty}\mathbb{E}[\mathcal{E}_1(\bar{\mathbf{w}}_{M+N})] \eqsim p^2\|\boldsymbol{\delta}\|^2_{\mathbf{H}}.$$

This proves the corollary. ∎

I Additional Experiments

I.1 CIFAR10 CNN Experiments

CIFAR10 experiment setup. We conduct experiments on the `CIFAR10` [37] dataset, which consists of 50,000 training images and 10,000 test images across 10 classes. All images are of size 32×32. All experiments are run on a single NVIDIA RTX 3090 GPU.

We use a CNN with five convolutional layers. Each block consists of a convolution layer followed by batch normalization and ReLU activation. Model size is controlled by a width parameter D, which scales the number of channels in all convolutional layers proportionally. Specifically, the channel dimensions follow the pattern $(D, D, 2D, 2D, 4D)$. We vary D in $\{8, 16, 32, 64, 128, 256\}$ to study the effect of model capacity.

Synthetic data shares the same input distribution as the real data, but labels are generated by a weak teacher model. The teacher is a smaller CNN (width 16) trained on a random subset of 5,000 training samples for 5 epochs. For each input x, the teacher produces a soft label distribution $q_{\text{teacher}}(y \mid x)$ with temperature scaling. The synthetic label is then defined as $\tilde{y} = q_{\text{teacher}}(y \mid x)$, , corresponding to fully synthetic labels.

Similar to the simulation experiments, we consider two training protocols as described in Section 2. For a fixed total training budget T, we allocate pT synthetic samples and $(1-p)T$ real samples, where $p \in [0, 1]$ is the synthetic data proportion. We vary p over $\{0.0, 0.1, \ldots, 0.9\}$ and T over a grid ranging from 200 to 20,000. To examine how the synthetic-induced floor depends on model size, we fix $p = 0.9$ and vary the model width D (while also varying T across the same grid).

Evaluation. We evaluate performance on the real test set using both classification error and cross-entropy loss. Each experiment is repeated with 3 random seeds, and we report the mean and standard deviation. To isolate the effect of synthetic data from baseline model performance, we report the metrics difference relative to $p = 0$ under the same T:

$$\Delta_{error} = Error(p = 0.9) - Error(p = 0), \quad \Delta_{loss} = Loss(p = 0.9) - Loss(p = 0),$$

which removes the intrinsic advantage of larger models and highlights the synthetic-data-induced degradation.

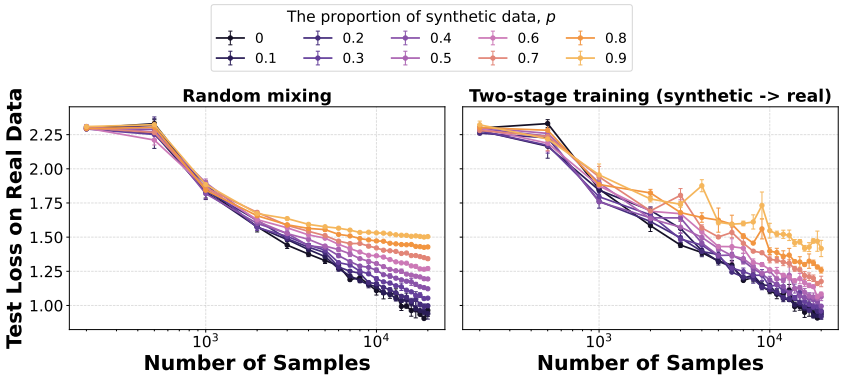
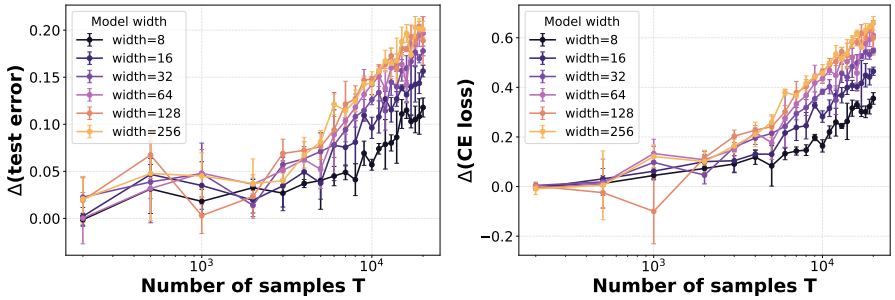


Figure 3: **Real-data experiments of different training protocols.** Test cross-entropy loss on the real distribution as a function of the number of training samples T, under different synthetic proportions p. **Left:** random mixing of real and synthetic data. **Right:** two-stage training (synthetic-data training followed by real-data training).

From Figure 3, we observe the same qualitative behavior predicted by our theory. Under random mixing, the test loss quickly saturates as T increases, forming a clear floor that depends on the synthetic proportion p. In particular, larger values of p lead to systematically higher asymptotic error, indicating that synthetic data introduces an irreducible bias that cannot be removed by additional samples. In contrast, two-stage training mitigates this phenomenon: the test loss continues to decrease with T, and consistently achieves slightly better performance than mixing in the large-sample regime.

These results confirm that the synthetic-data-induced floor is a consequence of mixing, rather than an inherent limitation of the data itself.



(a) Test error difference (b) Test cross-entropy difference

Figure 4: **Effect of model size on synthetic-data-induced degradation.** We plot the performance gap between high synthetic proportion ($p = 0.9$) and pure real-data training ($p = 0$), as a function of the number of samples T, for different model widths. Across both test error and cross-entropy loss, the gap consistently increases with model size, especially in the large-sample regime. This indicates that larger models are more sensitive to synthetic data, leading to a stronger irreducible bias under mixing.

Figure 4 shows how the impact of synthetic data varies with model size on real data. We measure the performance gap between high synthetic proportion ($p = 0.9$) and pure real-data training ($p = 0$), thereby isolating the effect of synthetic data. We observe a clear and consistent trend: as model size increases, the gap becomes significantly larger, particularly at large sample sizes. While all models exhibit some degradation due to synthetic data, larger models suffer substantially more, indicating that the synthetic-data-induced bias is amplified with model capacity. This behavior aligns with our theoretical predictions, suggesting that the floor effect induced by mixing is not only persistent but also worsens with increasing model size.

I.2 Language Model Experiments

Experiment setup. We conduct language-modeling experiments on WikiText-103, tokenized with the GPT-2 tokenizer. All models are decoder-only Transformers trained with context length 256 using next-token cross-entropy loss, and evaluated by validation loss on the real validation set. Unless otherwise stated, we use batch size 64 and AdamW with learning rate 6×10^{-4}, weight decay 0.1, training on a single NVIDIA RTX 3090 GPU.

Synthetic text is generated from a DistilGPT-2 teacher. For the synthetic corpus, we sample continuations of length 256 from real WikiText-103 prompts of length 32, using temperature 1.5, top-$p = 0.95$, top-$k = 0$, and store only the generated continuations.

We first compare the same two training protocols as before. In *mixed* training, each step uses real or synthetic data so that the cumulative synthetic fraction tracks p. In *two-stage* training, for each total budget T, the model is first trained for approximately pT synthetic steps and then for the remaining real steps. Our main experiment uses a 4-layer Transformer with 4 attention heads and embedding dimension 256, containing 16.1M parameters. We vary $p \in \{0, 0.3, 0.6, 0.9\}$ and evaluate 12 logarithmically spaced budgets from 100 to 24,000 steps, corresponding to up to 393M trained tokens. Results are shown in Figure 5.

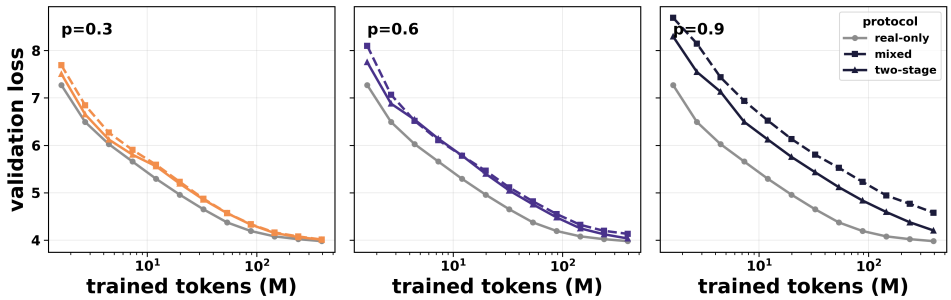


Figure 5: **Language-model protocol comparison.** We compare mixed training and two-stage training on WikiText-103 for synthetic fractions $p \in \{0.3, 0.6, 0.9\}$, with the real-only baseline shown in gray. Under mixed training, validation loss remains consistently worse than the real-only curve, especially at large p, indicating a persistent degradation from training on a fixed fraction of synthetic data. Two-stage training reduces this degradation and approaches the real-only curve more closely at large budgets. This matches the theoretical prediction that mixed training induces a non-vanishing synthetic-data bias floor, whereas using synthetic data only as an initialization stage can mitigate this floor.

To study the effect of model capacity, we fix a high synthetic fraction $p = 0.9$ under the mixed protocol and compare against the matched real-only baseline $p = 0$. We train three Transformer models with sizes 4/256/4 (16.1M params), 6/384/6 (30.0M params), and 8/512/8 (51.1M params), where the three numbers denote layers, embedding dimension, and attention heads. We evaluate budgets from 3,000 to 48,000 steps with batch size 32, corresponding to 24.6M to 393.2M trained tokens. We report the degradation $\Delta\text{loss} = \text{ValLoss}(p = 0.9) - \text{ValLoss}(p = 0)$, which isolates the synthetic-data-induced effect from the baseline improvement due to model size. Results are shown in Figure 6.

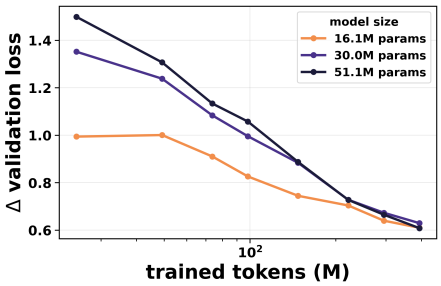
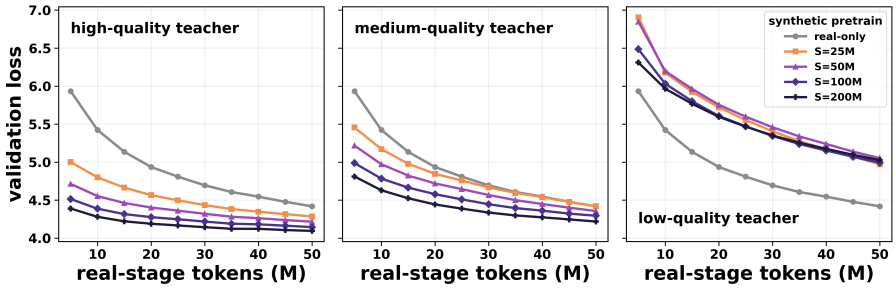
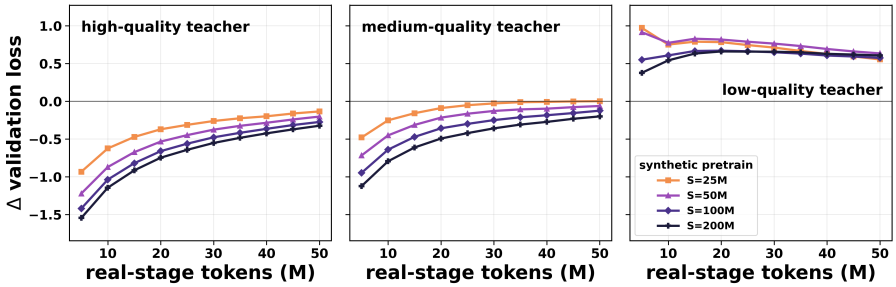


Figure 6: **Effect of model size on synthetic-data-induced degradation.** We plot the validation-loss gap $\Delta\text{loss} = \text{ValLoss}(p = 0.9) - \text{ValLoss}(p = 0)$ for models with 16.1M, 30.0M, and 51.1M parameters. The larger models exhibit a larger degradation gap at small and intermediate budgets, showing that the effect of synthetic mismatch is amplified by model capacity. This supports the theoretical prediction that higher-capacity models can fit the biased synthetic signal more strongly, leading to a larger synthetic-induced degradation under mixed training.

Finally, we test how teacher quality affects the usefulness of synthetic pretraining under a fixed real-data budget. We use the same 4-layer student and compare three synthetic teachers: a high-quality DistilGPT-2 corpus sampled with temperature 0.9, top-$p = 0.95$, top-$k = 0$, and prompt length 32; the medium-quality corpus from the main experiment; and a low-quality corpus sampled with temperature 4.0, top-$p = 1.0$, top-$k = 0$, and prompt length 4. We pretrain on synthetic budgets of approximately 25M, 50M, 100M, and 200M tokens, then reset the optimizer and train on real-data budgets up to 50M tokens. Performance is evaluated by real validation loss and by the gap relative to the real-only curve at the same real-token budget. Figure 7.



(a) Validation loss after synthetic pretraining followed by fixed real-data training.



(b) Validation-loss gap relative to the real-only curve at the same real-token budget.

Figure 7: **Effect of synthetic teacher quality under a fixed real-data budget.** We compare high-, medium-, and low-quality synthetic corpora while keeping the downstream real-data budget fixed. High-quality synthetic pretraining consistently improves validation loss relative to real-only training, and the improvement grows with the synthetic pretraining budget. Medium-quality synthetic data provides weaker but still positive gains at larger synthetic budgets. In contrast, low-quality synthetic data leads to a positive loss gap, meaning that it hurts performance even after subsequent real-data training. These results validate the theoretical bias-quality trade-off: synthetic data can help when its induced bias is small enough to reduce optimization error, but poor-quality synthetic data introduces a mismatch that dominates and causes degradation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. See the Abstract and Introduction sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work in the conclusion section, which are also our future research directions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide the assumptions and main results in the main paper, and state all the complete proof in Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the configuration of the experiments in Section 4 and Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper includes experiments based on standard synthetic Gaussian simulations and the public `CIFAR10` dataset, with details provided in Section 4 and Appendix I. We provide the code in Supplementary Material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the experimental details in Section 4 and Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We repeat the experiments over multiple independent random seeds and report mean performance with error bars indicating variability across runs. See details in Section 4 and Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide all the experimental details in Section 4 and Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper is a theoretical analysis paper on the generalization of stochastic gradient descent whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the creators of all external datasets and assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[N/A]

    Justification: This paper does not release new assets

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This paper utilized LLM as a writing-assistive tool. Its application was strictly limited to improving the language and presentation of the paper

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.