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

Jichu Li LIJICHU52@GMAIL.COM
Renmin University of China

Difan Zou DZOU@HKU.HK
The University of Hong Kong

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* [17], where any fixed fraction of synthetic data prevents model performance from improving under data scaling, leaving a non-vanishing excess risk floor. In this paper, we studied how synthetic data affects the generalization of one-pass SGD in high-dimensional linear regression with model shift. We show that mixed training induces strong model collapse while two-stage training avoids this by using synthetic data only in the first stage, followed by real-data training in the second stage, showing that strong model collapse is not inevitable through a simple data curriculum. We further establish scaling-law upper bounds for both protocols under a random sketch model, showing that larger models may amplify synthetic-induced degradation in mixed training and giving an explicit characterization of how high-quality synthetic training may reduce bias in two-stage training. 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 achieved remarkable performance, largely driven by scaling both model size and high-quality training data [9, 28, 32, 53]. However, the supply of high-quality human-generated data is limited and increasingly scarce [49, 59], posing a major bottleneck for further scaling. As a result, synthetic data generated by existing models has emerged as a promising alternative, especially in label scarce regimes [10, 27, 46].

A central question is whether synthetic data truly improves learning. While prior work shows that synthetic data can effectively increase data volume and improve generalization [23, 44, 47], it can also induce *model collapse*, where training on generated data leads to performance degradation [1, 24, 55, 56]. Recent theoretical studies formalize this phenomenon as regression problem under distribution mismatch, showing that synthetic data can degrade performance or require careful design to be beneficial [16, 18, 21, 54]. In particular, Dohmatob et al. [17] introduce *strong model collapse*, where even a small fraction of synthetic data induces a non-vanishing error floor under data scaling. However, existing analyses largely focus on static estimators or asymptotic regimes, and do not capture the dynamics of practical iterate optimization algorithms such as SGD.

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 covariates but differ in their labeling functions, and analyze SGD under both mixed and two-stage training protocols. We establish finite-sample upper and lower bounds

© J. Li & D. Zou.

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. Beyond this, we derive scaling-law upper bounds under a random sketch model, characterizing how model size affects synthetic-induced degradation and how synthetic data may benefit learning through bias reduction.

Our contributions can be summarized as follows:

- In Section 3.1, we show that last-iterate one-pass SGD under mixed training will induce 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 upper bounds for the two training protocols under a random sketch model, 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.

2. 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 $\mathcal{H} \times \mathbb{R}$.

In this paper, motivated by [17], 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 capturing the prevalent self-training and knowledge distillation paradigms.

Assumption 1 (Data Covariance) *Let $\mathbf{H} := \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ be the data covariance matrix and assume that $\operatorname{tr}(\mathbf{H})$ 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 $\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top] \preceq \alpha \operatorname{tr}(\mathbf{H}\mathbf{A})\mathbf{H}$; (2) There is a constant $\beta > 0$, such that for every PSD matrix $\mathbf{A}$, we have $\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top] - \mathbf{H}\mathbf{A}\mathbf{H} \succeq \beta \operatorname{tr}(\mathbf{H}\mathbf{A})\mathbf{H}$.*

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 $p := M/(M+N)$ denote the fraction of synthetic data. The training samples are presented in a random order. *(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 train using stochastic gradient descent (SGD) with updates $\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma_t(\langle \mathbf{x}_t, \mathbf{w}_t\rangle - y_t)\mathbf{x}_t$, where $\{(\mathbf{x}_t, y_t)\}_{t=1}^{M+N}$ are samples drawn from either the mixed or two-stage protocol. We use a geometrically decaying stepsize $\gamma_t = \gamma/2^{\ell_t}$ with $\ell_t = \lfloor t/(T/\log T)\rfloor$ [42, 43, 62, 63], where T is the total number of iterations ($T = M + N$ for mixed training, and applied decay schedule separately per stage for two-stage training). We output the last iterate $\mathbf{w}_{M+N}$ and evaluate the excess risk on the real distribution, $\mathsf{Excess} := \mathcal{E}_1(\mathbf{w}_{M+N}) = \mathcal{R}(\mathbf{w}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*)$..

3. Main Results

In this section, we only present our core results due to space constraints. Detailed analyses including finite-sample upper/lower bounds and are deferred to Appendix C, D, E and F.

3.1. Mixed training

Corollary 1 (Strong model collapse) *Given M synthetic and N real samples, consider the last iterate of SGD with geometrically decaying step sizes. Under the assumptions of Theorems 6 and 13, suppose $\gamma < 1/4\alpha\,\mathrm{tr}(\mathbf{H})$ and $\widetilde{D}_{\mathtt{eff}} = o(\widetilde{N}_{\mathtt{eff}})$. Then 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 1 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 synthetic proportion 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. More importantly, we show that *strong model collapse* is fundamentally induced by sample mixing in the training protocol, rather than an artifact of the optimization scheme. In particular, the same phenomenon arises for constant stepsize SGD with iterate averaging, where $\gamma_t = \gamma$ and the output is $\bar{\mathbf{w}}_{M+N} = \frac{1}{M+N}\sum_{i=0}^{M+N-1}\mathbf{w}_i$. In this case, we have $\lim_{M+N\to\infty}\mathbb{E}[\mathcal{E}_1(\bar{\mathbf{w}}_{M+N})] \eqsim p^2\|\boldsymbol{\delta}\|_{\mathbf{H}}^2$, demonstrating that the same non-vanishing error floor persists; see Appendix G for details.

This naturally raises an important question: *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

Corollary 2 (No strong model collapse) *Given M synthetic and N real samples, consider the last iterate of SGD with geometrically decaying step sizes. Under the assumptions of Theorem 16. As the*

total sample size $M+N$ scales to infinity with a fixed synthetic proportion $p = M/(M+N) \in (0,1)$, suppose that $D_{\texttt{eff}} = o(N_{\texttt{eff}})$, then $\mathbb{E}[\mathcal{E}_2(\mathbf{w}_M)] \to 0$ as $M \to \infty$, and we have

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

Corollary 2 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.

3.3. Random Sketch Model & Scaling laws

To reveal how the effect of synthetic data interacts with model size, following Lin et al. [42], 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 $\mathbf{Sx}$. We then train a linear predictor in the sketched space, $f_{\mathbf{v}}(\mathbf{x}) = \langle \mathbf{v}, \mathbf{Sx}\rangle$, $\mathbf{v} \in \mathbb{R}^D$, 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 $\mathcal{R}_D(\mathbf{v}) := \frac{1}{2}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_1}\big(\langle \mathbf{v}, \mathbf{Sx}\rangle - y\big)^2$, where the expectation is conditioned on $\mathbf{S}$. The corresponding optimal parameters are $\mathbf{v}_\ell^* := (\mathbf{SHS}^\top)^{-1}\mathbf{SHw}_\ell^*$, $\ell \in \{1,2\}$. 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. Then 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 [37, 42, 43, 66], we further make the following additional assumptions:

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 3 (Scaling law upper bound for mixed training) *Given M synthetic and N real samples. Suppose that Assumptions 1, 2, 3, 4 and 5 hold. Let $\widetilde{N}_{\texttt{eff}} = (M+N)/\log(M+N)$. 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*

$$\mathbb{E}\mathcal{R}_D(\mathbf{v}_{M+N}) \lesssim \sigma_1^2 + \underbrace{\frac{1}{D^{a-1}} + \frac{1}{(\widetilde{N}_{\texttt{eff}})^{1-1/a}}}_{\text{Approx+Bias}} + \underbrace{\Big(\frac{M}{M+N}\Big)^2\Big(1 - \frac{1}{D^{b-1}}\Big)}_{\text{Drift Floor}}$$

Interpretation. Theorem 3 provides a clean upper-bound scaling picture for mixed training under the random sketch model. Synthetic data improves the reducible part of the risk at the pre-asymptotic level. Compared with the real-only scaling $\frac{1}{(N_{\mathtt{eff}})^{1-1/a}}$ in Lin et al. [42], the term $\frac{1}{(\widetilde{N}_{\mathtt{eff}})^{1-1/a}}$ shows that synthetic and real samples contribute jointly, effectively increasing the sample size and reducing $\mathsf{Approx} + \mathsf{Bias}$. However, while increasing the sketch dimension D decreases the approximation term D^{1-a} and therefore improves $\mathsf{Approx} + \mathsf{Bias}$, 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, revealing a fundamental trade-off in mixed training.

Theorem 4 (Scaling law upper bound for two-stage training) *Given M synthetic and N real samples. Suppose 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*

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

Interpretation. The variance term is of the same order as in the real-only scaling law of Lin et al. [42], namely $\frac{\min\{D, N_{\mathtt{eff}}^{1/a}\}}{N_{\mathtt{eff}}}$, and therefore synthetic data does not improve the variance. 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 $\max\{D^{1-a}, N_{\mathtt{eff}}^{-(a-1)/a}\}$[42], 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 collapse.

4. Conclusion

We studied how synthetic data affects the generalization of one-pass SGD in high-dimensional linear regression with model shift. We show that mixed training induces *strong model collapse*: even a small fixed synthetic-data proportion p leaves a non-vanishing excess-risk floor under data scaling. In contrast, two-stage training avoids this floor, showing that collapse is not inevitable through a simple data curriculum. Our random sketch model analysis further establishes scaling-law upper bounds for both protocols, showing that larger models may amplify synthetic-induced degradation in mixed training and giving an explicit characterization of how high-quality synthetic training may reduce bias in two-stage training. Overall, synthetic data is not inherently harmful or beneficial; its effect depends critically on both its quality and how it is incorporated during training.

References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

[14] 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.

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

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

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

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

[42] 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.

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

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

[62] 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.

[63] 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.

[64] 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.

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

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

[67] 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.

[68] 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.

[69] 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 Problem Setup **2**

3 Main Results **3**
3.1 Mixed training . . . 3
3.2 Two-stage Training . . . 3
3.3 Random Sketch Model & Scaling laws . . . 4

4 Conclusion **5**

A Related Work **12**

B Preliminaries **13**

C Mixed Training Upper Bound Analysis **14**
C.1 Excess Risk Decomposition . . . 14
C.2 Bias upper bound . . . 15
C.3 Variance upper bound . . . 16
C.4 Drift upper bound . . . 16
C.5 Fluctuation upper bound . . . 19
C.6 Proof of Theorem 6 . . . 26

D Mixed Training Lower Bound Analysis **26**
D.1 The mean recursion . . . 27
D.2 The centered fluctuation term . . . 28
D.3 Proof of Theorem 13 . . . 36
D.4 Proof of Corollary 1 . . . 36

E Two-stage Training Analysis **37**

F Random Sketch Model **40**
F.1 Mixed Training Analysis . . . 40
F.2 Two-stage Training Analysis . . . 42

G constant stepsize SGD with iterate averaging **46**
G.1 Upper bound analysis . . . 46
G.2 Lower bound analysis . . . 57
G.3 Strong model collapse behavior . . . 59

H Experiments **61**

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$\langle \mathbf{A}, \mathbf{B} \rangle := \operatorname{tr}(\mathbf{A}^\top \mathbf{B})$. 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 $\otimes$.

Let $\mathbf{H} = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ be the eigen-decomposition of $\mathbf{H}$, where $\{\lambda_i\}_{i=1}^\infty$ are the eigenvalues of $\mathbf{H}$ sorted in non-increasing order and $\mathbf{v}_i$ are the corresponding eigenvectors. Following [62, 69], we denote:

$$\mathbf{H}_{0:k} := \sum_{i=1}^{k} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \quad \mathbf{H}_{k:\infty} := \sum_{i>k} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \quad \boldsymbol{\delta} = \sum_i \delta_i \mathbf{v}_i.$$

Appendix A. Related Work

Synthetic data and model collapse. Training with synthetic data has been shown to induce *model collapse* [55], 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 [1, 7, 8, 24, 26, 56]. Theoretical studies of model collapse typically frame it as a regression problem under distribution mismatch between real and synthetic data [16, 18, 54], showing how synthetic data degrade performance, while others explore how to avoid collapse or leverage synthetic data to benefit learning [3, 20, 21, 54]. In particular, Dohmatob et al. [17] 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 ([19, 29]) 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 [4, 6, 25, 33, 34, 36, 39, 40, 50, 57, 61]. 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. [62] study pretraining-finetuning under covariate shift via SGD and establish excess risk bounds. Ding et al. [15] extend the analysis to continual learning settings and show how covariate shifts across tasks affect SGD dynamics and forgetting. Liu et al. [45] characterize the minimax optimality of SGD-type methods for high-dimensional linear regression under covariate shift. Beyond standard covariate shift, Deng and Kpotufe [12] 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 [2, 11, 14, 30, 31, 51, 52]. One-pass SGD under different iterate schemes (e.g., last iterate, averaging) has also been rigorously studied in the overparameterized setting [5, 13, 22, 58, 63, 65, 67, 69], providing framework for analyzing the generalization of SGD in high-dimensional regime. In parallel, Several works have extended the analysis to multipass SGD [36, 41, 48, 51, 68]. More recently, a line of research has focused on the theoretical scaling laws of SGD in high-dimensional linear regression. Lin et al. [42] 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. [43] further extended their analysis to multi-epoch SGD, followed by an even finer-grained characterization from Yan et al. [64]. Li et al. [37] established functional scaling laws and analyzed optimal learning rate and batch size schedules [38, 60]. 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.

Appendix B. Preliminaries

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

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

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 $\mathbf{A}$, [69]:

$$\mathcal{I} \circ \mathbf{A} = \mathbf{A}, \quad \mathcal{M} \circ \mathbf{A} = \mathbb{E}[(\mathbf{x}^\top \mathbf{A} \mathbf{x})\mathbf{x}\mathbf{x}^\top], \quad \widetilde{\mathcal{M}} \circ \mathbf{A} = \mathbf{H}\mathbf{A}\mathbf{H},$$
$$(\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \mathbf{A} = \mathbb{E}[(\mathbf{I} - \gamma_t \mathbf{x}\mathbf{x}^\top)\mathbf{A}(\mathbf{I} - \gamma_t \mathbf{x}\mathbf{x}^\top)], \quad (\mathcal{I} - \gamma_t \widetilde{\mathcal{T}}_t) \circ \mathbf{A} = (\mathbf{I} - \gamma_t \mathbf{H})\mathbf{A}(\mathbf{I} - \gamma_t \mathbf{H}). \tag{1}$$

For the linear operators we have the following technical lemma from Zou et al. [69].

Lemma 5 (Lemma B.1, Zou et al. [69]) *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. [69]. ■

Appendix C. Mixed Training Upper Bound Analysis

Theorem 6 (An upper bound for mixed training) *Given M synthetic and N real samples. Consider last iterate SGD with geometrically decaying stepsizes. Suppose Assumptions 1, 2 and 3 hold. Let $\widetilde{N}_{\texttt{eff}} := (M+N)/\log(M+N)$ and $\bar{\sigma}^2 = (1-p)\sigma_1^2 + p\sigma_2^2$. Suppose $\gamma < 1/(4\alpha\,\mathrm{tr}(\mathbf{H}))$. We have*

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

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

First, we introduce source indicators to help analysis.

Source indicators. At iteration $t \geq 1$ the algorithm receives one sample $(\mathbf{x}_t, y_t)$ together with a source indicator $z_t \in \{0,1\}$, where

$$z_t = \begin{cases} 0, & \text{real sample,} \\ 1, & \text{synthetic sample.} \end{cases}$$

Conditional on $z_t = a$, the sample $(\mathbf{x}_t, y_t)$ is drawn according to the source-a population.

Fix an integer M and define $p := \frac{M}{M+N}$. The vector

$$\mathbf{z}_{1:M+N} := (z_1, \ldots, z_{M+N})$$

is uniformly distributed over all binary sequences in $\{0,1\}^{M+N}$ containing exactly M ones. Directly, $\sum_{i=1}^{M+N} z_i = M$.

Note that in this section we 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.

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 (2) serves as the starting point of the analysis:

$$\boldsymbol{\eta}_t = (\mathbf{I} - \gamma_t\mathbf{x}_t\mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1} + \gamma_t z_t\mathbf{x}_t\mathbf{x}_t^\top\boldsymbol{\delta} + \gamma_t\xi_t\mathbf{x}_t. \tag{2}$$

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

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

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

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

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:

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

C.2. Bias upper bound

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

$$\mathbf{B}_t = (\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \mathbf{B}_{t-1}, \quad \mathbf{B}_0 = \boldsymbol{\eta}_0 \otimes \boldsymbol{\eta}_0, \tag{3}$$

Note that this bias term $\big\langle \mathbf{H}, \mathbf{B}_N \big\rangle$ is the same as that in [63], so we can apply Lemma in [63] to bound it directly:

Lemma 7 (A bias upper bound) *Suppose Assumptions 1 and 2 hold. Let $N_{\texttt{eff}} = N/\log N$. Consider (3). Suppose $\gamma < 1/(4\alpha\,\mathrm{tr}(\mathbf{H}))$. We have*

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

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 See proof of Theorem D.1 in [62]. ■

C.3. Variance upper bound

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

$$\mathbf{V}_t = (\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \mathbf{V}_{t-1} + \gamma_t^2 \mathbf{\Sigma}_t, \qquad \mathbf{V}_0 = \mathbf{0}. \tag{4}$$

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

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

Then we can apply Lemma C.2 in [63] to bound variance term $\langle \mathbf{H}, \mathbf{V}_N \rangle$ by simply replace σ^2 with $(1-p)\sigma_1^2 + p\sigma_2^2$.

Lemma 8 (A variance bound) *Suppose Assumptions 1, 2 and 3 hold. Consider (4). Let $N_{\texttt{eff}} = N/\log N$. Suppose $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$. We have*

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

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 Theorem C.2 in [63]. ■

C.4. Drift upper bound

We now analyze the drift term

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

Lemma 9 (A drift bound) *Suppose Assumptions 1, 2 and 3 hold. Consider (4). Let $N_{\texttt{eff}} = N/\log N$. Suppose $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$. We have*

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

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

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

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

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

$$\overline{\boldsymbol{\eta}}_t^{\text{drift}} = (\mathbf{I} - \gamma_t \mathbf{H}) \overline{\boldsymbol{\eta}}_{t-1}^{\text{drift}} + \gamma_t p \mathbf{H} \boldsymbol{\delta}, \qquad \overline{\boldsymbol{\eta}}_0^{\text{drift}} = \mathbf{0}. \tag{5}$$

Define the centered drift fluctuation $\widetilde{\boldsymbol{\eta}}_t^{\text{drift}} := \boldsymbol{\eta}_t^{\text{drift}} - \overline{\boldsymbol{\eta}}_t^{\text{drift}}$ and subtracting (5) from the recursion of $\boldsymbol{\eta}_t^{\text{drift}}$, we obtain

$$\begin{aligned}\widetilde{\boldsymbol{\eta}}_t^{\text{drift}} &= (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{drift}} - (\mathbf{I} - \gamma_t \mathbf{H})\overline{\boldsymbol{\eta}}_{t-1}^{\text{drift}} \\ &= (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\widetilde{\boldsymbol{\eta}}_{t-1}^{\text{drift}} + \gamma_t(\mathbf{H} - \mathbf{x}_t \mathbf{x}_t^\top)\overline{\boldsymbol{\eta}}_{t-1}^{\text{drift}}, \qquad \widetilde{\boldsymbol{\eta}}_0^{\text{drift}} = \mathbf{0}. \end{aligned} \tag{6}$$

By definition,

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

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

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

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

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

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

$$\langle \mathbf{H}, \mathbf{D}_N \rangle = \|\overline{\boldsymbol{\eta}}_N^{\text{drift}}\|_{\mathbf{H}}^2 + \langle \mathbf{H}, \widetilde{\mathbf{D}}_N \rangle. \tag{7}$$

In the following, we will bound the two terms in (7) separately.

Mean part of the drift term. We now derive the explicit form of $\|\overline{\boldsymbol{\eta}}_N^{\text{drift}}\|_{\mathbf{H}}^2$. Recall from (5) that

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

Unrolling the recursion gives

$$\overline{\boldsymbol{\eta}}_N^{\text{drift}} = p \sum_{s=1}^{N} \gamma_s \left(\prod_{t=s+1}^{N} (\mathbf{I} - \gamma_t \mathbf{H}) \right) \mathbf{H}\boldsymbol{\delta}. \tag{8}$$

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 (8), we obtain

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

As a consequence,

$$\|\overline{\boldsymbol{\eta}}_N^{\text{drift}}\|_{\mathbf{H}}^2 = p^2 \Big\| \Big(\mathbf{I} - \prod_{t=1}^{N} (\mathbf{I} - \gamma_t \mathbf{H})\Big)\boldsymbol{\delta} \Big\|_{\mathbf{H}}^2. \tag{9}$$

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

Fluctuation part of the drift term. We now bound the second term in (7), namely

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

Recall from (6) that

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

Define

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

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

$$\widetilde{\mathbf{D}}_t = (\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \widetilde{\mathbf{D}}_{t-1} + \gamma_t^2 \widetilde{\boldsymbol{\Sigma}}_t^{\text{drift}}, \qquad \widetilde{\mathbf{D}}_0 = \mathbf{0}. \tag{10}$$

Indeed, the cross term vanishes because

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

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

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

Since $\mathcal{M} - \widetilde{\mathcal{M}}$ is a PSD mapping by Lemma 5, we have $\widetilde{\boldsymbol{\Sigma}}_t^{\text{drift}} \succeq 0$. On the other hand, by Assumption 2,

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

Therefore,

$$\widetilde{\boldsymbol{\Sigma}}_t^{\text{drift}} \preceq \alpha \|\overline{\boldsymbol{\eta}}_{t-1}^{\text{drift}}\|_{\mathbf{H}}^2 \mathbf{H}. \tag{11}$$

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

$$\begin{aligned}
\|\overline{\boldsymbol{\eta}}_{t-1}^{\text{drift}}\|_{\mathbf{H}}^2 &= p^2 \sum_i \lambda_i \left(1 - \prod_{s=1}^{t-1}(1 - \gamma_s\lambda_i)\right)^2 \delta_i^2 \\
&\le p^2 \sum_i \lambda_i \delta_i^2 = p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2.
\end{aligned} \tag{12}$$

Combining (11) and (12), we obtain

$$\widetilde{\boldsymbol{\Sigma}}_t^{\text{drift}} \preceq \alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \mathbf{H}. \tag{13}$$

The recursion (10) has exactly the same form as the variance recursion (4). Therefore, applying Lemma 8 with

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

we immediately get

$$\left\langle \mathbf{H}, \widetilde{\mathbf{D}}_N \right\rangle \le \frac{8\alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}, \tag{14}$$

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 (9) and (14), we conclude that

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

■

C.5. Fluctuation upper bound

We now analyze the drift term

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

Lemma 10 (A fluctuation bound) *Suppose Assumptions 1, 2 and 3 hold. Consider (4). Let $N_{\mathtt{eff}} = N/\log N$. Suppose $\gamma < 1/(\alpha\,\mathrm{tr}(\mathbf{H}))$. We have*

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

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,

$$\begin{aligned}\big\langle \mathbf{H}, \mathbf{F}_N \big\rangle &= \big\langle \mathbf{H}, \mathbb{E}\big[(\boldsymbol{\eta}_N^{\text{fluct},1} + \boldsymbol{\eta}_N^{\text{fluct},2}) \otimes (\boldsymbol{\eta}_N^{\text{fluct},1} + \boldsymbol{\eta}_N^{\text{fluct},2})\big]\big\rangle \\ &\le 2\big\langle \mathbf{H}, \mathbf{F}_N^{(1)} \big\rangle + 2\big\langle \mathbf{H}, \mathbf{F}_N^{(2)} \big\rangle, \end{aligned} \tag{15}$$

where

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

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

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

Define

$$\boldsymbol{\Sigma}_t^{\text{fluct},1} := \mathbb{E}\Big[z_t^2\big((\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big) \otimes \big((\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\boldsymbol{\delta}\big)\Big]. \tag{16}$$

Then

$$\mathbf{F}_t^{(1)} = (\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \mathbf{F}_{t-1}^{(1)} + \gamma_t^2 \boldsymbol{\Sigma}_t^{\text{fluct},1} + \mathbf{C}_t^{(1)} + (\mathbf{C}_t^{(1)})^\top, \tag{17}$$

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

$$\mathbb{E}\big[\boldsymbol{\eta}_t^{\text{fluct},1} \mid \mathcal{Z}\big] = \mathbf{0}, \qquad \forall t \ge 0. \tag{18}$$

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 (18) 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 (18). Hence

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

Taking transpose also gives

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

Therefore (17) reduces to

$$\mathbf{F}_t^{(1)} = (\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \mathbf{F}_{t-1}^{(1)} + \gamma_t^2 \boldsymbol{\Sigma}_t^{\text{fluct},1}, \qquad \mathbf{F}_0^{(1)} = \mathbf{0}. \tag{19}$$

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

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

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

$$\boldsymbol{\Sigma}_t^{\text{fluct},1} \preceq p\, \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top \mathbf{A} \mathbf{x}_t\mathbf{x}_t^\top]. \tag{20}$$

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

$$\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top \mathbf{A} \mathbf{x}_t\mathbf{x}_t^\top] \preceq \alpha\, \operatorname{tr}(\mathbf{H}\mathbf{A})\, \mathbf{H} = \alpha\, \|\boldsymbol{\delta}\|_{\mathbf{H}}^2\, \mathbf{H}. \tag{21}$$

Combining (20) and (21), we obtain

$$\mathbf{\Sigma}_t^{\text{fluct},1} \preceq p\,\alpha\|\boldsymbol{\delta}\|_{\mathbf{H}}^2\,\mathbf{H}. \tag{22}$$

Comparing (19) with the variance recursion (4), we see that $\mathbf{F}_t^{(1)}$ satisfies exactly the same recursion form as the variance term, with noise covariance $\mathbf{\Sigma}_t^{\text{fluct},1}$ in place of $\mathbf{\Sigma}_t$. By (22), we may apply Lemma 8 with

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

Therefore,

$$\langle \mathbf{H}, \mathbf{F}_N^{(1)} \rangle \leq \frac{8p\alpha\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \frac{D_{\texttt{eff}}}{N_{\texttt{eff}}}, \tag{23}$$

where

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

Consequently, by (15),

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

and the first term is controlled by (23). The second term $\langle \mathbf{H}, \mathbf{F}_N^{(2)} \rangle$ is be bounded by Lemma 12. ■

Lemma 11 (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

$$\mathbb{E}\left\|\sum_{s=1}^{N}(z_s - p)\boldsymbol{v}_s\right\|_{\mathbf{H}}^2 = \frac{Np(1-p)}{N-1}\sum_{s=1}^{N}\|\boldsymbol{v}_s - \bar{\boldsymbol{v}}\|_{\mathbf{H}}^2. \tag{24}$$

In particular,

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

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

$$\mathbb{E}\left\|\sum_{s=1}^{N}(z_s - p)\boldsymbol{v}_s\right\|_{\mathbf{H}}^2 = \frac{p(1-p)}{N-1}\left[(N-1)\sum_{s=1}^{N}\|\boldsymbol{v}_s\|_{\mathbf{H}}^2 - \sum_{s\neq r}\langle \boldsymbol{v}_s, \boldsymbol{v}_r\rangle_{\mathbf{H}}\right]. \tag{26}$$

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

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

Substituting this identity into (26) 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 (24).

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 (25). ■

Lemma 12 (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* $D_{\mathtt{eff}} := k^* + \gamma^2 N_{\mathtt{eff}}^2\sum_{i>k^*}\lambda_i^2$.

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 $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ are deterministic. Since the source indicators are independent of the features and follow the fixed-budget law, Lemma 11 gives

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

Taking expectation over the features gives

$$\big\langle \mathbf{H}, \mathbf{F}_N^{(2)} \big\rangle = \mathbb{E}\Big[\|\boldsymbol{\eta}_N^{\text{fluct},2}\|_{\mathbf{H}}^2\Big] \le \frac{Np(1-p)}{N-1} \sum_{s=1}^{N} \mathbb{E}\big[\|\boldsymbol{v}_s\|_{\mathbf{H}}^2\big]. \tag{27}$$

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

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

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

$$\mathbf{O} \preceq \|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \, \mathbf{H}. \tag{28}$$

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 5, 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 (28) holds, an induction on t shows that

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

Consequently,

$$\sum_{s=1}^{N} \mathbb{E}\big[\|\boldsymbol{v}_s\|_{\mathbf{H}}^2\big] = \langle \mathbf{H}, \bar{\mathbf{S}}_N \rangle \le \langle \mathbf{H}, \mathbf{C}_N^\delta \rangle. \tag{29}$$

The recursion defining $\mathbf{C}_t^\delta$ has exactly the same form as the variance recursion (4), with noise covariance $\boldsymbol{\Sigma}^\delta := \|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \mathbf{H}$. Applying Lemma 8 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 (27) and (29), 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 6

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

Appendix D. Mixed Training Lower Bound Analysis

Theorem 13 (A lower bound for mixed training) *Given M synthetic and N real samples. Consider last iterate SGD with geometrically decaying stepsizes. 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 $M + N \ge 500$ and $\gamma < 1/\lambda_1$, then*

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

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

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

$$\boldsymbol{\eta}_t = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1} + \gamma_t z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} + \gamma_t \xi_t \mathbf{x}_t, \qquad \boldsymbol{\eta}_0 = \mathbf{w}_0 - \mathbf{w}_1^*. \tag{30}$$

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

$$\mathbb{E}[\mathcal{R}(\mathbf{w}_N) - \mathcal{R}(\mathbf{w}_1^*)] = \frac{1}{2}\|\bar{\boldsymbol{\eta}}_N\|_{\mathbf{H}}^2 + \frac{1}{2}\left\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N \otimes \widetilde{\boldsymbol{\eta}}_N] \right\rangle. \tag{31}$$

D.1. The mean recursion

Taking expectation on both sides of (30), 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

$$\bar{\boldsymbol{\eta}}_N = \prod_{t=1}^{N}(\mathbf{I} - \gamma_t \mathbf{H})(\mathbf{w}_0 - \mathbf{w}_1^*) + p\sum_{s=1}^{N}\gamma_s \left(\prod_{t=s+1}^{N}(\mathbf{I} - \gamma_t \mathbf{H})\right)\mathbf{H}\boldsymbol{\delta}. \tag{32}$$

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 (32), 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,

$$\|\bar{\boldsymbol{\eta}}_N\|_{\mathbf{H}}^2 = \Big\| \prod_{t=1}^{N}(\mathbf{I} - \gamma_t \mathbf{H})(\mathbf{w}_0 - \mathbf{w}_1^*) + p\big(\mathbf{I} - \prod_{t=1}^{N}(\mathbf{I} - \gamma_t \mathbf{H})\big)\boldsymbol{\delta}\Big\|_{\mathbf{H}}^2. \tag{33}$$

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 $\big\langle \mathbf{H}, \mathbf{V}_N \big\rangle$ in [63], which will be used in our analysis.

Lemma 14 (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*

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

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

Proof See proof of Theorem C.2 in [63]. ■

We now continue the analysis of the centered fluctuation term

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

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

$$\widetilde{\boldsymbol{\eta}}_t = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\widetilde{\boldsymbol{\eta}}_{t-1} + \gamma_t \xi_t \mathbf{x}_t + \gamma_t\big(z_t \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\delta} - p\mathbf{H}\boldsymbol{\delta}\big) + \gamma_t(\mathbf{H} - \mathbf{x}_t \mathbf{x}_t^\top)\bar{\boldsymbol{\eta}}_{t-1}. \tag{34}$$

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

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

Thus

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

Expanding the second moment gives

$$\begin{aligned}\mathbb{E}[\widetilde{\boldsymbol{\eta}}_N \otimes \widetilde{\boldsymbol{\eta}}_N] &= \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{noise}}] + \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{res}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{res}}] \\ &\quad + \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{res}}] + \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{res}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{noise}}].\end{aligned} \tag{35}$$

We now prove that

$$\mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_t^{\text{res}}] = \mathbf{0}, \qquad \forall t \geq 0. \tag{36}$$

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

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

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 (37).

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 (36). 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

$$\big\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N \otimes \widetilde{\boldsymbol{\eta}}_N]\big\rangle = \big\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{noise}}]\big\rangle + \big\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{res}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{res}}]\big\rangle. \tag{38}$$

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 14, with σ_1^2 replaced by $\bar{\sigma}^2$. Hence

$$\big\langle \mathbf{H}, \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\text{noise}} \otimes \widetilde{\boldsymbol{\eta}}_N^{\text{noise}}]\big\rangle = \big\langle \mathbf{H}, \mathbf{V}_N^{\text{mix}}\big\rangle \geq \frac{\bar{\sigma}^2}{400} \cdot \frac{D_{\texttt{eff}}}{N_{\texttt{eff}}}. \tag{39}$$

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 15 (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*

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

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

$$\widetilde{\boldsymbol{\eta}}_t^{\mathrm{res}} = (\mathbf{I} - \gamma_t\mathbf{x}_t\mathbf{x}_t^\top)\widetilde{\boldsymbol{\eta}}_{t-1}^{\mathrm{res}} + \gamma_t(z_t - p)\mathbf{x}_t\mathbf{x}_t^\top\boldsymbol{\delta} + \gamma_t(\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\mathbf{c}_{t-1}, \qquad \widetilde{\boldsymbol{\eta}}_0^{\mathrm{res}} = \mathbf{0}. \tag{40}$$

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

$$\mathbf{S}_Z := \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}} \mid \mathcal{Z}], \qquad \mathbf{S}_X := \mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}} \mid \mathcal{X}], \qquad \mathbf{R}^\circ := \widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}} - \mathbf{S}_Z - \mathbf{S}_X. \tag{41}$$

We first show that

$$\mathbb{E}[\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}}] = \mathbf{0}. \tag{42}$$

Indeed, taking expectation in (40) 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 $\widetilde{\boldsymbol{\eta}}_0^{\mathrm{res}} = \mathbf{0}$, (42) 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 (41),

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

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,

$$\mathbb{E}\|\widetilde{\boldsymbol{\eta}}_N^{\mathrm{res}}\|_{\mathbf{H}}^2 = \mathbb{E}\|\mathbf{S}_Z\|_{\mathbf{H}}^2 + \mathbb{E}\|\mathbf{S}_X\|_{\mathbf{H}}^2 + \mathbb{E}\|\mathbf{R}^\circ\|_{\mathbf{H}}^2 \geq \mathbb{E}\|\mathbf{S}_Z\|_{\mathbf{H}}^2 + \mathbb{E}\|\mathbf{S}_X\|_{\mathbf{H}}^2. \tag{43}$$

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 (43).

Define $\mathbf{f}_t := \mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\mathrm{res}} \mid \mathcal{X}]$.Taking conditional expectation in (40) and using $\mathbb{E}[z_t - p \mid \mathcal{X}] = 0$, we obtain

$$\mathbf{f}_t = (\mathbf{I} - \gamma_t \mathbf{x}_t\mathbf{x}_t^\top)\mathbf{f}_{t-1} + \gamma_t(\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{H})\mathbf{c}_{t-1}, \qquad \mathbf{f}_0 = \mathbf{0}. \tag{44}$$

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 (44) gives

$$\mathbf{S}_X = \sum_{u=1}^{N} \mathbf{U}_u, \qquad \mathbf{U}_u := \mathbf{K}_{u,N}\boldsymbol{\zeta}_u. \tag{45}$$

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

$$\mathbb{E}\|\mathbf{S}_X\|_{\mathbf{H}}^2 = \sum_{u=1}^{N} \mathbb{E}\|\mathbf{U}_u\|_{\mathbf{H}}^2. \tag{46}$$

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

$$\mathbb{E}\|\mathbf{U}_u\|_{\mathbf{H}}^2 = \mathbb{E}\big[\boldsymbol{\zeta}_u^\top \mathbf{K}_{u,N}^\top \mathbf{H}\mathbf{K}_{u,N}\boldsymbol{\zeta}_u\big] = \big\langle \mathbb{E}[\mathbf{K}_{u,N}^\top \mathbf{H}\mathbf{K}_{u,N}],\, \mathbf{G}_u\big\rangle, \tag{47}$$

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 (47) yields

$$\mathbb{E}\|\mathbf{U}_u\|_{\mathbf{H}}^2 \geq \beta \, \|\mathbf{c}_{u-1}\|_{\mathbf{H}}^2 \, \mathbb{E}\big[\, \mathrm{tr}(\mathbf{H}\mathbf{K}_{u,N}\mathbf{H}\mathbf{K}_{u,N}^\top)\big]. \tag{48}$$

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

$$\kappa_u := \sum_{i\geq 1} \gamma_u^2 \lambda_i^2 \prod_{j=u+1}^{N} (1 - \gamma_j \lambda_i)^2. \tag{49}$$

Then (48) gives

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

Combining this with (46),

$$\mathbb{E}\|\mathbf{S}_X\|_{\mathbf{H}}^2 \geq \beta \sum_{u=1}^{N} \|\mathbf{c}_{u-1}\|_{\mathbf{H}}^2 \, \kappa_u. \tag{50}$$

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

$$\|\mathbf{c}_{u-1}\|^2_{\mathbf{H}_{k^*:\infty}} = \sum_{i>k^*} \lambda_i \prod_{j=1}^{u-1} (1-\gamma_j\lambda_i)^2 (\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta})_i^2. \tag{51}$$

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

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

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

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

Substituting this into (51) gives

$$\|\mathbf{c}_{u-1}\|^2_{\mathbf{H}} \ge \|\mathbf{c}_{u-1}\|^2_{\mathbf{H}_{k^*:\infty}} \ge e^{-8}\|\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}\|^2_{\mathbf{H}_{k^*:\infty}}, \qquad u = 1, \dots, N. \tag{52}$$

Therefore (50) yields

$$\mathbb{E}\|\mathbf{S}_X\|^2_{\mathbf{H}} \ge \beta e^{-8}\|\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}\|^2_{\mathbf{H}_{k^*:\infty}} \sum_{u=1}^{N} \kappa_u. \tag{53}$$

Finally, by applying Lemma 14,

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

Substituting this into (53), we conclude that

$$\mathbb{E}\|\mathbf{S}_X\|^2_{\mathbf{H}} \ge \frac{\beta e^{-8}}{400}\|\mathbf{w}_0 - \mathbf{w}_1^* - p\boldsymbol{\delta}\|^2_{\mathbf{H}_{k^*:\infty}} \left(\frac{k^*}{N_{\texttt{eff}}} + \gamma^2 N_{\texttt{eff}} \sum_{i>k^*} \lambda_i^2\right). \tag{54}$$

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 (40) and using independence of $\mathbf{x}_t$ from $(\mathcal{Z}, \mathbf{x}_1, \dots, \mathbf{x}_{t-1})$, 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

$$\mathbf{S}_Z = \sum_{u=1}^{N} (z_u - p)\mathbf{v}_u, \qquad \mathbf{v}_u := \gamma_u\Big(\prod_{j=u+1}^{N} (\mathbf{I} - \gamma_j\mathbf{H})\Big)\mathbf{H}\boldsymbol{\delta}. \tag{55}$$

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,

$$\mathbb{E}\|\mathbf{S}_Z\|_{\mathbf{H}}^2 = \frac{p(1-p)}{2(N-1)} \sum_{u=1}^{N} \sum_{r=1}^{N} \|\mathbf{v}_u - \mathbf{v}_r\|_{\mathbf{H}}^2. \tag{56}$$

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

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

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

$$\|\mathbf{v}_u - \mathbf{v}_r\|_{\mathbf{H}}^2 \geq (e^{-4} - 2^{-7})^2\gamma^2 \sum_{i>k^*} \lambda_i^3\delta_i^2. \tag{58}$$

Substituting (58) into (57), 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,

$$\mathbb{E}\|\mathbf{S}_Z\|_{\mathbf{H}}^2 \geq \frac{(e^{-4} - 2^{-7})^2}{40} p(1-p)\gamma^2 N_{\mathtt{eff}} \sum_{i>k^*} \lambda_i^3\delta_i^2. \tag{59}$$

Finally, combining (43), (54), and (59), 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 13

Proof Combining Lemma 15 with Eq. 39 yields the desired result. ■

D.4. Proof of Corollary 1

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 6 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 $\widetilde{D}_{\texttt{eff}} = o(\widetilde{N}_{\texttt{eff}})$, 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 13:

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

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

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

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. ■

Appendix E. Two-stage Training Analysis

Theorem 16 (Two-stage training) *Given M synthetic samples and N real samples. Consider last iterate SGD with geometrically decaying stepsizes. 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*

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

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

First, consider the upper bound for training on pure one-source data from [63], which can also be recovered by Theorem 6.

Theorem 17 (Learning with only real data) *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*

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

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

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 6 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 6 conditionally, we obtain

$$\begin{aligned}\mathbb{E}\big[\mathcal{R}(\mathbf{w}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*) \mid \mathcal{F}_M\big] \lesssim &\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I} - \gamma_t\mathbf{H})\Big)(\mathbf{w}_M - \mathbf{w}_1^*)\Big\|_{\mathbf{H}}^2 \\ &+ \left(\alpha\|\mathbf{w}_M - \mathbf{w}_1^*\|^2_{\mathbf{I}_{0:k^*}/\gamma K + \mathbf{H}_{k^*:\infty}} + \sigma_1^2\right)\frac{D_{\texttt{eff}}}{N_{\texttt{eff}}},\end{aligned} \tag{60}$$

Taking expectation with respect to $\mathcal{F}_M$ on both sides of (60), we get

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

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

$$\begin{aligned}&\mathbb{E}\left[\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I} - \gamma_t\mathbf{H})\Big)(\mathbf{w}_M - \mathbf{w}_1^*)\Big\|_{\mathbf{H}}^2\right] \\ &= \mathbb{E}\left[\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I} - \gamma_t\mathbf{H})\Big)(\mathbf{w}_M - \mathbf{w}_2^*) + \Big(\prod_{t=1}^{N}(\mathbf{I} - \gamma_t\mathbf{H})\Big)\boldsymbol{\delta}\Big\|_{\mathbf{H}}^2\right] \\ &\leq 2\mathbb{E}\left[\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I} - \gamma_t\mathbf{H})\Big)(\mathbf{w}_M - \mathbf{w}_2^*)\Big\|_{\mathbf{H}}^2\right] + 2\Big\|\Big(\prod_{t=1}^{N}(\mathbf{I} - \gamma_t\mathbf{H})\Big)\boldsymbol{\delta}\Big\|_{\mathbf{H}}^2.\end{aligned} \tag{62}$$

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

$$\begin{aligned} \mathbb{E}\Big[\|\mathbf{P}_N(\mathbf{w}_M - \mathbf{w}_2^*)\|_{\mathbf{H}}^2\Big] &\le \mathbb{E}\Big[\|\mathbf{w}_M - \mathbf{w}_2^*\|_{\mathbf{H}}^2\Big] \\ &= 2\,\mathbb{E}[L_2(\mathbf{w}_M) - L_2(\mathbf{w}_2^*)]. \end{aligned} \tag{63}$$

Next, let

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

By the definition of k^*,

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

Therefore

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

and hence

$$\mathbf{B} = \frac{\mathbf{I}_{0:k^*}}{\gamma K} + \mathbf{H}_{k^*:\infty} \preceq \mathbf{H}_{0:k^*} + \mathbf{H}_{k^*:\infty} = \mathbf{H}. \tag{64}$$

It follows that

$$\begin{aligned} \mathbb{E}\Big[\|\mathbf{w}_M - \mathbf{w}_2^*\|_{\mathbf{B}}^2\Big] &\le \mathbb{E}\Big[\|\mathbf{w}_M - \mathbf{w}_2^*\|_{\mathbf{H}}^2\Big] \\ &= 2\,\mathbb{E}[L_2(\mathbf{w}_M) - L_2(\mathbf{w}_2^*)]. \end{aligned} \tag{65}$$

On the other hand,

$$\begin{aligned} \mathbb{E}\Big[\|\mathbf{w}_M - \mathbf{w}_1^*\|_{\mathbf{B}}^2\Big] &= \mathbb{E}\Big[\|(\mathbf{w}_M - \mathbf{w}_2^*) + \boldsymbol{\delta}\|_{\mathbf{B}}^2\Big] \\ &\le 2\mathbb{E}\Big[\|\mathbf{w}_M - \mathbf{w}_2^*\|_{\mathbf{B}}^2\Big] + 2\|\boldsymbol{\delta}\|_{\mathbf{B}}^2 \\ &\le 4\,\mathbb{E}[L_2(\mathbf{w}_M) - L_2(\mathbf{w}_2^*)] + 2\|\boldsymbol{\delta}\|_{\mathbf{B}}^2, \end{aligned} \tag{66}$$

where in the last step we used (65).

Substituting (62), (63), and (66) into (61), we obtain

$$\begin{aligned} \mathbb{E}\big[\mathcal{R}(\mathbf{w}_{M+N}) - \mathcal{R}(\mathbf{w}_1^*)\big] \lesssim\ & \mathbb{E}[\mathcal{R}_2(\mathbf{w}_M) - \mathcal{R}_2(\mathbf{w}_2^*)] + \big\|\mathbf{P}_N \boldsymbol{\delta}\big\|_{\mathbf{H}}^2 \\ & + \big(\alpha\,\mathbb{E}[\mathcal{R}_2(\mathbf{w}_M) - \mathcal{R}_2(\mathbf{w}_2^*)] + \alpha\|\boldsymbol{\delta}\|_{\mathbf{B}}^2 + \sigma_1^2\big)\,\frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}. \end{aligned} \tag{67}$$

Finally, by (64),

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

Therefore

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

Equivalently, writing

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

we have

$$\mathtt{Excess}(\mathbf{w}_{M+N}) \lesssim \mathtt{Excess}_2(\mathbf{w}_M) + \big\|\mathbf{P}_N\boldsymbol{\delta}\big\|_{\mathbf{H}}^2 + \Big(\alpha\,\mathtt{Excess}_2(\mathbf{w}_M) + \alpha\|\boldsymbol{\delta}\|_{\mathbf{H}}^2 + \sigma_1^2\Big)\frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}}. \tag{69}$$

This is the desired two-stage upper bound.

Appendix F. Random Sketch Model

Following [42], we begin by decomposing the population risk into three components: irreducible risk, approximation error, and excess risk:

$$\mathcal{R}_D(\mathbf{v}_{M+N}) = \underbrace{\min\mathcal{R}(\cdot)}_{\text{Irreducible}} + \underbrace{\min\mathcal{R}_D(\cdot) - \min\mathcal{R}(\cdot)}_{\text{Approx}} + \underbrace{\mathcal{R}_D(\mathbf{v}_{M+N}) - \min\mathcal{R}_D(\cdot)}_{\text{Excess}}. \tag{70}$$

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

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

For Approximation Error, as established in Lemma C.4 in [42], under Assumption 4, with probability at least $1 - e^{\Omega(D)}$,

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

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

F.1. Mixed Training Analysis

Recall the upper bound in Theorem 6, 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 $\widetilde{\lambda}_j$ be the eigenvalue of $\mathbf{SHS}^\top$. Suppose $\gamma < 1/(4\alpha\,\mathrm{tr}(\mathbf{SHS}^\top))$. Then we have

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

with

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

Note that when $\sigma_1^2, \sigma_2^2 \eqsim 1$, Theorem A.4 in [42] shows that

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

Future assume that Assumption 5 holds and zero initialiazation where $\mathbf{v}_0 = \mathbf{0}$, [42](Appendix D and E) shows that with probability at least $1 - e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$

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

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

and

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

For the Fluct term, we have

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

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

$$\mathbb{E}_{\boldsymbol{\delta}}\|\mathbf{v}_2^* - \mathbf{v}_1^*\|_{\mathbf{SHS}^\top}^2 \lesssim \mathbb{E}\|\boldsymbol{\delta}\|_{\mathbf{H}}^2 \eqsim \sum_i i^{-b} \eqsim 1 \tag{71}$$

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

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

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

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

For the second minus term, [42](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 $\mathbf{S}$

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

So that under Assumption 5,

$$\mathbb{E}_{\boldsymbol{\delta}}\left\|\left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}^\top\mathbf{H}^{\frac{1}{2}}\right)\mathbf{H}^{\frac{1}{2}}\boldsymbol{\delta}\right\|^2 \eqsim D^{1-b} \tag{72}$$

Combine Eq.71 with 72, we get

$$\mathbb{E}_{\boldsymbol{\delta}}\mathrm{Drift} \eqsim p^2(1 - D^{1-b})$$

Finally, put $\mathsf{Irreducible}, \mathsf{Approx}$ 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 $\mathbf{S}$:

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

F.2. Two-stage Training Analysis

First, Recall the upper bound in Theorem 16. 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 $\gamma < 1/(4\alpha\,\mathrm{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top))$. Then we have

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

where

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

Suppose $\sigma_1^2, \sigma_2^2 \eqsim 1$. For the Var term, similar to the analysis in mixed training analysis and [42](Lemma E.1), with probability at least $1 - e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$

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

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

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

Using Theorem 6.1 in [42], for $\gamma < 1/(4\alpha\ \mathrm{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top))$ and zero initialization,

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

where

$$D_{\mathtt{eff}1} := \#\{\widetilde{\lambda}_j \geq 1/(M_{\mathtt{eff}}\gamma)\} + (M_{\mathtt{eff}}\gamma)^2 \sum_{\widetilde{\lambda}_j < 1/(M_{\mathtt{eff}}\gamma)} \widetilde{\lambda}_j^2,$$

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

By Assumption 5, let $(\lambda_i, \mathbf{v}_i)_{i\geq 1}$ be the eigenvalue-eigenvector pairs of $\mathbf{H}$, and $\mathbf{\Sigma}$ shares the eigenvectors $(\mathbf{v}_i)_{i\geq 1}$ 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 [42] with a cutoff $k \leq D/3$ yields

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

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

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

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

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

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 [42], we obtain

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

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 $(M_{\mathtt{eff}}\gamma)^{1/a} \le D/c_0$, this upper bound is tight up to constants.

Finally, by Lemma E.1 in [42],

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

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 $\|(\prod_{t=1}^N \mathbf{I} - \gamma_t \mathbf{SHS}^\top)(\mathbf{v}_2^* - \mathbf{v}_1^*)\|_{\mathbf{SHS}^\top}^2$, under Assumption 5, we can apply Lemma D.4 in [42] directly and get that probability at least $1 - e^{-\Omega(D)}$ over the randomness of the sketch matrix $\mathbf{S}$:

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

and

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

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

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

For the last term

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

let $\mathbf{P}_N := \prod_{t=1}^N (\mathbf{I} - \gamma_t \mathbf{SHS}^\top)$. Conditioning on the sketch matrix $\mathbf{S}$, we diagonalize $\mathbf{SHS}^\top = \mathbf{U}\mathrm{diag}(\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_D)\mathbf{U}^\top$ with $\widetilde{\lambda}_1 \ge \cdots \ge \widetilde{\lambda}_D > 0$. Then

$$\mathbf{P}_N^\top \mathbf{SHS}^\top \mathbf{P}_N = \mathbf{U}\mathrm{diag}\Big(\widetilde{\lambda}_i \prod_{t=1}^N (1 - \gamma_t \widetilde{\lambda}_i)^2\Big)\mathbf{U}^\top \preceq \Big(\prod_{t=1}^N (1 - \gamma_t \widetilde{\lambda}_D)\Big)^2 \mathbf{SHS}^\top.$$

Hence

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

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

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

for some constant $c > 0$. Since

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

it follows that

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

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

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

and let $\mathcal{E}_{\text{ex}}$ 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 [42] 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 $\mathcal{E}_{\text{spec}} \cap \mathcal{E}_{\text{ex}}$, we have

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

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

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

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 $\mathbf{S}$:

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

Appendix G. 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.

G.1. Upper bound analysis

We use the same definition in Appendix C.

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

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:

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

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

$$\mathbf{B}_t = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{B}_{t-1}, \quad \mathbf{B}_0 = \boldsymbol{\eta}_0 \otimes \boldsymbol{\eta}_0, \tag{73}$$

Note that this bias term $\langle \mathbf{H}, \bar{\mathbf{B}}_N \rangle$ is the same as that in [69], so we can apply Lemma B.11 in [69] to bound it directly:

Lemma 18 (A bias upper bound for Avg-SGD) *Suppose Assumptions 1 and 2 hold. Let $N_{\texttt{eff}} = N$. Consider (73). Suppose $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$. We have*

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

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 See proof of Lemma B.11 in [69]. ■

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

$$\mathbf{V}_t = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{V}_{t-1} + \gamma^2 \boldsymbol{\Sigma}_t, \qquad \mathbf{V}_0 = \mathbf{0}, \tag{74}$$

where

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

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

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

Define

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

Then (74) has exactly the same form as the variance recursion for constant-stepsize SGD with iterate averaging analyzed in [69], with noise level upper bounded by σ_{mix}^2. Therefore, by applying the averaged-SGD variance bound in [69], 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 [69].

Lemma 19 (A variance upper bound for Avg-SGD) *Suppose Assumptions 1, 2 and 3 hold. Consider (74). Let $N_{\texttt{eff}} = N$ and $\bar{\sigma}^2 = (1-p)\sigma_1^2 + p\sigma_2^2$. Suppose $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$. Then*

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

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

Proof The recursion (74) 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 [69] applies verbatim with σ^2 replaced by σ^2_{mix}. Hence

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

This completes the proof. ■

Drift upper bound We now analyze the drift contribution

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

Lemma 20 (A drift upper bound for Avg-SGD) *Suppose Assumptions 1, 2 and 3 hold. Let $N_{\mathtt{eff}} = N$. Suppose $\gamma < 1/(\alpha\,\mathrm{tr}(\mathbf{H}))$. Then*

$$\frac{1}{2}\langle \mathbf{H}, \bar{\mathbf{D}}_N \rangle \le \frac{p^2}{2}\left\| \left(\mathbf{I} - \frac{1}{\gamma N_{\mathtt{eff}}}\mathbf{H}^{-1}\left(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^{N_{\mathtt{eff}}}\right)\right)\boldsymbol{\delta}\right\|_{\mathbf{H}}^2 + \frac{\alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}},$$

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

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

We first separate the drift process into its mean part and centered fluctuation part. Define

$$\mathbf{m}_t := \mathbb{E}[\boldsymbol{\eta}_t^{\text{drift}}].$$

Since $\mathbf{x}_t$ is independent of $\boldsymbol{\eta}_{t-1}^{\text{drift}}$ and $\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top] = \mathbf{H}$, taking expectation on both sides yields

$$\mathbf{m}_t = (\mathbf{I} - \gamma\mathbf{H})\mathbf{m}_{t-1} + \gamma p \mathbf{H}\boldsymbol{\delta}, \qquad \mathbf{m}_0 = \mathbf{0}. \tag{75}$$

Let

$$\widetilde{\boldsymbol{\eta}}_t^{\text{drift}} := \boldsymbol{\eta}_t^{\text{drift}} - \mathbf{m}_t.$$

Subtracting (75) from the recursion of $\boldsymbol{\eta}_t^{\text{drift}}$ gives

$$\widetilde{\boldsymbol{\eta}}_t^{\text{drift}} = (\mathbf{I} - \gamma\mathbf{x}_t\mathbf{x}_t^\top)\widetilde{\boldsymbol{\eta}}_{t-1}^{\text{drift}} + \gamma(\mathbf{H} - \mathbf{x}_t\mathbf{x}_t^\top)\mathbf{m}_{t-1}, \qquad \widetilde{\boldsymbol{\eta}}_0^{\text{drift}} = \mathbf{0}. \tag{76}$$

Moreover, by construction,

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

Now define the averaged quantities

$$\bar{\mathbf{m}}_N := \frac{1}{N}\sum_{t=0}^{N-1} \mathbf{m}_t, \qquad \bar{\widetilde{\boldsymbol{\eta}}}_N^{\text{drift}} := \frac{1}{N}\sum_{t=0}^{N-1} \widetilde{\boldsymbol{\eta}}_t^{\text{drift}}.$$

Then

$$\bar{\boldsymbol{\eta}}_N^{\text{drift}} = \bar{\mathbf{m}}_N + \bar{\widetilde{\boldsymbol{\eta}}}_N^{\text{drift}}.$$

Since $\mathbb{E}[\bar{\widetilde{\boldsymbol{\eta}}}_N^{\text{drift}}] = \mathbf{0}$, the cross term vanishes and

$$\bar{\mathbf{D}}_N = \bar{\mathbf{m}}_N \otimes \bar{\mathbf{m}}_N + \bar{\widetilde{\mathbf{D}}}_N, \qquad \bar{\widetilde{\mathbf{D}}}_N := \mathbb{E}\Big[\bar{\widetilde{\boldsymbol{\eta}}}_N^{\text{drift}} \otimes \bar{\widetilde{\boldsymbol{\eta}}}_N^{\text{drift}}\Big]. \tag{77}$$

Therefore,

$$\frac{1}{2}\big\langle \mathbf{H}, \bar{\mathbf{D}}_N \big\rangle = \frac{1}{2}\|\bar{\mathbf{m}}_N\|_{\mathbf{H}}^2 + \frac{1}{2}\big\langle \mathbf{H}, \bar{\widetilde{\mathbf{D}}}_N \big\rangle. \tag{78}$$

Unrolling (75), 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

$$\bar{\mathbf{m}}_N = p\left(\mathbf{I} - \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^N\big)\right)\boldsymbol{\delta}. \tag{79}$$

Therefore,

$$\frac{1}{2}\|\bar{\mathbf{m}}_N\|_{\mathbf{H}}^2 = \frac{p^2}{2}\left\|\left(\mathbf{I} - \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^N\big)\right)\boldsymbol{\delta}\right\|_{\mathbf{H}}^2. \tag{80}$$

We next bound

$$\frac{1}{2}\big\langle \mathbf{H}, \bar{\widetilde{\mathbf{D}}}_N \big\rangle.$$

Define

$$\widetilde{\mathbf{D}}_t := \mathbb{E}[\widetilde{\boldsymbol{\eta}}_t^{\text{drift}} \otimes \widetilde{\boldsymbol{\eta}}_t^{\text{drift}}].$$

From (76), we obtain the recursion

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

where

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

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

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

Since $\mathcal{M} - \widetilde{\mathcal{M}}$ is a PSD mapping, we have $\widetilde{\mathbf{\Sigma}}_t^{\text{drift}} \succeq \mathbf{0}$. On the other hand, by Assumption 22,

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

Therefore,

$$\widetilde{\mathbf{\Sigma}}_t^{\text{drift}} \preceq \alpha\,\|\mathbf{m}_{t-1}\|_{\mathbf{H}}^2\,\mathbf{H}. \tag{82}$$

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 (82) gives the uniform bound

$$\widetilde{\mathbf{\Sigma}}_t^{\text{drift}} \preceq \alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2\,\mathbf{H}, \qquad \forall t. \tag{83}$$

Now (81) has exactly the same form as the variance recursion for constant-stepsize SGD with iterate averaging, with noise level upper bounded by

$$\sigma_{\text{drift}}^2 := \alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2.$$

Therefore, applying Lemma 19 with $\bar{\sigma}^2$ replaced by σ_{drift}^2, we obtain

$$\frac{1}{2}\langle \mathbf{H}, \bar{\widetilde{\mathbf{D}}}_N \rangle \le \frac{\alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1 - \gamma\alpha\operatorname{tr}(\mathbf{H})} \cdot \frac{D_{\texttt{eff}}}{N}. \tag{84}$$

Finally, combining (78), (80) and (84), we conclude that

$$\frac{1}{2}\langle \mathbf{H}, \bar{\mathbf{D}}_N \rangle \le \frac{p^2}{2} \left\| \left(\mathbf{I} - \frac{1}{\gamma N} \mathbf{H}^{-1} \big(\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^N \big) \right) \boldsymbol{\delta} \right\|_{\mathbf{H}}^2 + \frac{\alpha p^2 \|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1 - \gamma \alpha \operatorname{tr}(\mathbf{H})} \cdot \frac{D_{\texttt{eff}}}{N}.$$

This completes the proof. ■

Fluctuation upper bound Recall that the fluctuation process satisfies

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

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

$$\boldsymbol{\eta}_t^{\text{fluct}} = \boldsymbol{\eta}_t^{\text{cov}} + \boldsymbol{\eta}_t^{\text{sch}},$$

where

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

Thus, for the averaged iterates,

$$\bar{\boldsymbol{\eta}}_N^{\text{fluct}} = \bar{\boldsymbol{\eta}}_N^{\text{cov}} + \bar{\boldsymbol{\eta}}_N^{\text{sch}}.$$

Hence, by $(a+b)^2 \le 2a^2 + 2b^2$,

$$\frac{1}{2}\langle \mathbf{H}, \bar{\mathbf{F}}_N \rangle = \frac{1}{2} \mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_N^{\text{fluct}}\|_{\mathbf{H}}^2 \big] \le \mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_N^{\text{cov}}\|_{\mathbf{H}}^2 \big] + \mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_N^{\text{sch}}\|_{\mathbf{H}}^2 \big]. \tag{85}$$

Define $\mathbf{F}_t^{\text{cov}} := \mathbb{E}[\boldsymbol{\eta}_t^{\text{cov}} \otimes \boldsymbol{\eta}_t^{\text{cov}}]$. Then $\mathbf{F}_t^{\text{cov}}$ satisfies the recursion

$$\mathbf{F}_t^{\text{cov}} = (\mathcal{I} - \gamma \mathcal{T}) \circ \mathbf{F}_{t-1}^{\text{cov}} + \gamma^2 \boldsymbol{\Sigma}_t^{\text{cov}}, \qquad \mathbf{F}_0^{\text{cov}} = \mathbf{0}, \tag{86}$$

where

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

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

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

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

$$\boldsymbol{\Sigma}_t^{\mathrm{cov}} \preceq p\alpha\|\boldsymbol{\delta}\|_{\mathbf{H}}^2\,\mathbf{H}. \tag{87}$$

Therefore, the recursion (86) 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 19 with $\bar{\sigma}^2$ replaced by σ_{cov}^2, we get

$$\frac{1}{2}\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_N^{\mathrm{cov}}\|_{\mathbf{H}}^2\big] = \frac{1}{2}\big\langle \mathbf{H}, \bar{\mathbf{F}}_N^{\mathrm{cov}}\big\rangle \le \frac{p\alpha\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \frac{D_{\mathtt{eff}}}{N}. \tag{88}$$

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 21 (A schedule-fluctuation upper bound for Avg-SGD) *Suppose Assumptions 1 and 2 hold. Assume the source indicators $(z_1, \ldots, z_N)$ 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 $\gamma < 1/(\alpha\,\mathrm{tr}(\mathbf{H}))$. Let

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

Then

$$\frac{1}{2}\mathbb{E}\Big[\|\bar{\boldsymbol{\eta}}_N^{\mathrm{sch}}\|_{\mathbf{H}}^2\Big] \le \frac{p(1-p)\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{N-1}\left(1 + \frac{2\alpha\gamma\,\mathrm{tr}(\mathbf{H})}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})} D_{\mathtt{eff}}\right). \tag{89}$$

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 $t \in \{1, \dots, N\}$,

$$\boldsymbol{\eta}_t^{\mathrm{sch}} = \gamma \sum_{s=1}^{t} \left(\prod_{j=s+1}^{t} \mathbf{B}_j \right) (z_s - p)\mathbf{H}\boldsymbol{\delta}, \tag{90}$$

where the empty product is interpreted as the identity operator.

Therefore,

$$\begin{aligned} \bar{\boldsymbol{\eta}}_N^{\mathrm{sch}} &= \frac{1}{N} \sum_{t=1}^{N} \boldsymbol{\eta}_t^{\mathrm{sch}} \\ &= \frac{\gamma}{N} \sum_{t=1}^{N} \sum_{s=1}^{t} \left(\prod_{j=s+1}^{t} \mathbf{B}_j \right) (z_s - p)\mathbf{H}\boldsymbol{\delta} \\ &= \sum_{s=1}^{N} (z_s - p)\mathbf{v}_s, \end{aligned} \tag{91}$$

where

$$\mathbf{v}_s := \frac{\gamma}{N} \sum_{t=s}^{N} \left(\prod_{j=s+1}^{t} \mathbf{B}_j \right) \mathbf{H}\boldsymbol{\delta}. \tag{92}$$

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 $(\mathbf{x}_1, \dots, \mathbf{x}_N)$,

$$\begin{aligned} \mathbb{E}\left[\left\| \sum_{s=1}^{N} (z_s - p)\mathbf{v}_s \right\|_{\mathbf{H}}^2 \,\middle|\, \mathbf{x}_1, \dots, \mathbf{x}_N \right] &= \sum_{s=1}^{N} \sum_{r=1}^{N} \mathbb{E}[(z_s - p)(z_r - p)] \langle \mathbf{v}_s, \mathbf{v}_r \rangle_{\mathbf{H}} \\ &= p(1-p) \sum_{s=1}^{N} \|\mathbf{v}_s\|_{\mathbf{H}}^2 - \frac{p(1-p)}{N-1} \sum_{s \neq r} \langle \mathbf{v}_s, \mathbf{v}_r \rangle_{\mathbf{H}}. \end{aligned} \tag{93}$$

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 (93) 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 (91), and taking expectation over the samples, we obtain

$$\mathbb{E}\Big[\|\bar{\boldsymbol{\eta}}_N^{\mathrm{sch}}\|_{\mathbf{H}}^2\Big] \le \frac{Np(1-p)}{N-1}\sum_{s=1}^{N}\mathbb{E}\big[\|\mathbf{v}_s\|_{\mathbf{H}}^2\big]. \tag{94}$$

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,

$$\begin{aligned}\mathbf{v}_s &= \frac{\gamma}{N}\sum_{t=s}^{N}\left(\prod_{j=s+1}^{t}\mathbf{B}_j\right)\mathbf{H}\boldsymbol{\delta}\\ &= \frac{1}{N}\sum_{r=0}^{n_s-1}\mathbf{u}_r^{(s)} = \frac{n_s}{N}\,\bar{\mathbf{u}}_{n_s}^{(s)},\end{aligned} \tag{95}$$

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

$$\mathbb{E}[\|\mathbf{v}_s\|_{\mathbf{H}}^2] = \left(\frac{n_s}{N}\right)^2\mathbb{E}\big[\|\bar{\mathbf{u}}_{n_s}(\gamma\mathbf{H}\boldsymbol{\delta})\|_{\mathbf{H}}^2\big], \tag{96}$$

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 (96) into (94), and changing variables from s to $n = n_s$, yields

$$\mathbb{E}\Big[\|\bar{\boldsymbol{\eta}}_N^{\mathrm{sch}}\|_{\mathbf{H}}^2\Big] \le \frac{Np(1-p)}{N-1}\sum_{n=1}^{N}\left(\frac{n}{N}\right)^2 \mathbb{E}\big[\|\bar{\mathbf{u}}_n(\gamma\mathbf{H}\boldsymbol{\delta})\|_{\mathbf{H}}^2\big]\,. \tag{97}$$

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 18 with horizon n and initialization $\mathbf{u}_0 = \gamma\mathbf{H}\boldsymbol{\delta}$, we obtain

$$\begin{aligned}\frac{1}{2}\mathbb{E}\big[\|\bar{\mathbf{u}}_n(\gamma\mathbf{H}\boldsymbol{\delta})\|_{\mathbf{H}}^2\big] \le{}& \frac{1}{\gamma^2 n^2}\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{0:k_n}^{-1}}^2 + \|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{k_n:\infty}}^2 \\ &+ \frac{2\alpha\big(\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{I}_{0:k_n}}^2 + n\gamma\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{k_n:\infty}}^2\big)}{n\gamma(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\cdot\frac{D_n}{n}.\end{aligned} \tag{98}$$

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

$$\frac{1}{\gamma^2 n^2}\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{0:k_n}^{-1}}^2 = \frac{1}{n^2}\sum_{i\le k_n}\lambda_i\delta_i^2, \tag{99}$$

$$\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{k_n:\infty}}^2 = \gamma^2\sum_{i>k_n}\lambda_i^3\delta_i^2, \tag{100}$$

$$\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{I}_{0:k_n}}^2 = \gamma^2\sum_{i\le k_n}\lambda_i^2\delta_i^2. \tag{101}$$

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 (99) and (100),

$$\begin{aligned}\frac{1}{\gamma^2 n^2}\|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{0:k_n}^{-1}}^2 + \|\gamma\mathbf{H}\boldsymbol{\delta}\|_{\mathbf{H}_{k_n:\infty}}^2 &\le \frac{1}{n^2}\sum_{i\le k_n}\lambda_i\delta_i^2 + \frac{1}{n^2}\sum_{i>k_n}\lambda_i\delta_i^2 \\ &= \frac{\|\boldsymbol{\delta}\|_{\mathbf{H}}^2}{n^2}.\end{aligned} \tag{102}$$

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

$$\begin{aligned}\|\gamma\mathbf{H}\boldsymbol{\delta}\|^2_{\mathbf{I}_{0:k_n}} + n\gamma\|\gamma\mathbf{H}\boldsymbol{\delta}\|^2_{\mathbf{H}_{k_n:\infty}} &= \gamma^2 \sum_{i\le k_n} \lambda_i^2\delta_i^2 + n\gamma^3 \sum_{i>k_n} \lambda_i^3\delta_i^2 \\ &\le \gamma^2 \sum_i \lambda_i^2\delta_i^2 \\ &\le \gamma^2 \operatorname{tr}(\mathbf{H}) \sum_i \lambda_i\delta_i^2 = \gamma^2 \operatorname{tr}(\mathbf{H})\|\boldsymbol{\delta}\|^2_{\mathbf{H}}. \end{aligned} \tag{103}$$

Substituting (102) and (103) into (98), we obtain

$$\frac{1}{2}\mathbb{E}\big[\|\bar{\mathbf{u}}_n(\gamma\mathbf{H}\boldsymbol{\delta})\|^2_{\mathbf{H}}\big] \le \frac{\|\boldsymbol{\delta}\|^2_{\mathbf{H}}}{n^2}\left(1 + \frac{2\alpha\gamma\operatorname{tr}(\mathbf{H})}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})} D_n\right). \tag{104}$$

Substituting (104) into (97) yields

$$\begin{aligned}\frac{1}{2}\mathbb{E}\Big[\|\bar{\boldsymbol{\eta}}^{\mathrm{sch}}_N\|^2_{\mathbf{H}}\Big] &\le \frac{Np(1-p)}{N-1}\sum_{n=1}^{N}\left(\frac{n}{N}\right)^2 \cdot \frac{\|\boldsymbol{\delta}\|^2_{\mathbf{H}}}{n^2}\left(1 + \frac{2\alpha\gamma\operatorname{tr}(\mathbf{H})}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})} D_n\right) \\ &= \frac{p(1-p)\|\boldsymbol{\delta}\|^2_{\mathbf{H}}}{N(N-1)}\sum_{n=1}^{N}\left(1 + \frac{2\alpha\gamma\operatorname{tr}(\mathbf{H})}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})} D_n\right). \end{aligned} \tag{105}$$

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 (105) gives

$$\begin{aligned}\frac{1}{2}\mathbb{E}\Big[\|\bar{\boldsymbol{\eta}}^{\mathrm{sch}}_N\|^2_{\mathbf{H}}\Big] &\le \frac{p(1-p)\|\boldsymbol{\delta}\|^2_{\mathbf{H}}}{N(N-1)}\sum_{n=1}^{N}\left(1 + \frac{2\alpha\gamma\operatorname{tr}(\mathbf{H})}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})} D_{\mathtt{eff}}\right) \\ &= \frac{p(1-p)\|\boldsymbol{\delta}\|^2_{\mathbf{H}}}{N-1}\left(1 + \frac{2\alpha\gamma\operatorname{tr}(\mathbf{H})}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})} D_{\mathtt{eff}}\right), \end{aligned}$$

which proves (89). ■

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 22 (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 $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$. Then*

$$\begin{aligned}\mathbb{E}[\mathcal{R}(\bar{\mathbf{w}}_N) - \mathcal{R}(\mathbf{w}_1^*)] \lesssim \frac{1}{\gamma^2 N^2}\|\mathbf{w}_0 - \mathbf{w}_1^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}_0 - \mathbf{w}_1^*\|^2_{\mathbf{H}_{k^*:\infty}} \\ + \left(\alpha\|\mathbf{w}_0 - \mathbf{w}_1^*\|^2_{\mathbf{H}} + \bar{\sigma}^2 + \alpha p\|\boldsymbol{\delta}\|^2_{\mathbf{H}}\right) \cdot \frac{D_{\texttt{eff}}}{N} + \frac{p(1-p)\|\boldsymbol{\delta}\|^2_{\mathbf{H}}}{N}, \\ + p^2 \left\| \left(\mathbf{I} - \frac{1}{\gamma N}\mathbf{H}^{-1}\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^N\right)\right)\boldsymbol{\delta}\right\|^2_{\mathbf{H}}\end{aligned}$$

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

G.2. Lower bound analysis

The starting point is again the error recursion

$$\boldsymbol{\eta}_t = (\mathbf{I} - \gamma\mathbf{x}_t\mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1} + \gamma z_t\mathbf{x}_t\mathbf{x}_t^\top\boldsymbol{\delta} + \gamma\xi_t\mathbf{x}_t, \qquad \boldsymbol{\eta}_0 = \mathbf{w}_0 - \mathbf{w}_1^*, \tag{106}$$

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

$$\mathbb{E}[\mathcal{R}(\bar{\mathbf{w}}_N) - \mathcal{R}(\mathbf{w}_1^*)] = \frac{1}{2}\|\bar{\boldsymbol{\mu}}_N\|_{\mathbf{H}}^2 + \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\widetilde{\bar{\boldsymbol{\eta}}}_N \otimes \widetilde{\bar{\boldsymbol{\eta}}}_N]\rangle. \tag{107}$$

Since the second term is nonnegative, we immediately obtain

$$\mathbb{E}[\mathcal{R}(\bar{\mathbf{w}}_N) - \mathcal{R}(\mathbf{w}_1^*)] \geq \frac{1}{2}\|\bar{\boldsymbol{\mu}}_N\|_{\mathbf{H}}^2. \tag{108}$$

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 (106), 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

$$\boldsymbol{\mu}_t = (\mathbf{I} - \gamma\mathbf{H})^t(\mathbf{w}_0 - \mathbf{w}_1^*) + p\sum_{s=1}^{t}\gamma(\mathbf{I} - \gamma\mathbf{H})^{t-s}\mathbf{H}\boldsymbol{\delta}. \tag{109}$$

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 (109), 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,

$$\|\bar{\boldsymbol{\mu}}_N\|_{\mathbf{H}}^2 = \left\| \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N\big)(\mathbf{w}_0 - \mathbf{w}_1^*) + p\left(\mathbf{I} - \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N\big)\right)\boldsymbol{\delta}\right\|_{\mathbf{H}}^2. \tag{110}$$

Combining (108) and (110), we arrive at the lower bound

$$\mathbb{E}[\mathcal{R}(\bar{\mathbf{w}}_N) - \mathcal{R}(\mathbf{w}_1^*)] \gtrsim \left\| \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N\big)(\mathbf{w}_0 - \mathbf{w}_1^*) + p\left(\mathbf{I} - \frac{1}{\gamma N}\mathbf{H}^{-1}\big(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N\big)\right)\boldsymbol{\delta}\right\|_{\mathbf{H}}^2.$$

G.3. Strong model collapse behavior

Corollary 23 (Strong model collapse in Avg-SGD) *Consider averaged SGD with constant step-size 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,

$$\begin{aligned}\mathbb{E}[\mathcal{E}_1(\bar{\mathbf{w}}_T)] \lesssim & \underbrace{\frac{1}{\gamma^2T^2}\|\mathbf{w}_0-\mathbf{w}_1^*\|^2_{\mathbf{H}^{-1}_{0:k^*}}}_{(I)} + \underbrace{\|\mathbf{w}_0-\mathbf{w}_1^*\|^2_{\mathbf{H}_{k^*:\infty}}}_{(II)} \\ & + \underbrace{\left(\alpha\|\mathbf{w}_0-\mathbf{w}_1^*\|^2_{\mathbf{H}} + \bar{\sigma}^2 + \alpha p\|\boldsymbol{\delta}\|^2_{\mathbf{H}}\right)\cdot\frac{D_{\texttt{eff}}}{T}}_{(III)} + \underbrace{\frac{p(1-p)\|\boldsymbol{\delta}\|^2_{\mathbf{H}}}{T}}_{(IV)} \\ & + \underbrace{p^2\|(\mathbf{I}-\mathbf{A}_T)\boldsymbol{\delta}\|^2_{\mathbf{H}}}_{(V)},\end{aligned}$$

where

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

By the assumption $D_{\texttt{eff}} = o(T)$, terms (III) and (IV) vanish as $T\to\infty$. For term (I), note that

$$\frac{1}{\gamma^2T^2}\|\mathbf{w}_0-\mathbf{w}_1^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} \leq \|\mathbf{w}_0-\mathbf{w}_1^*\|_2^2\cdot\frac{1}{\gamma T},$$

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

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

Hence

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

Combining the upper and lower bounds yields

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

This proves the corollary. ■

Appendix H. Experiments

Simulation setup. 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 mismatch satisfies the source condition $\lambda_i\mathbb{E}[\delta_i^2] \propto i^{-2.0}$. We consider two models: the full high-dimensional linear model and a random sketch model with sketch width D. We train with last-iterate SGD under the two protocols in Section 2, varying both the synthetic proportion $p \in (0, 1)$ and the number of samples $T \in [50, 50000]$. We evaluate excess risk on the real distribution and, for the sketch model, further vary D to study how the synthetic-induced floor depends on model size. To isolate the synthetic-data contribution, we consider the risk increase over $p = 0$, removes the intrinsic advantage of larger models and highlights the synthetic-data-induced degradation. All experiments are repeated with 5 random seeds. The results are shown in Figure 1.

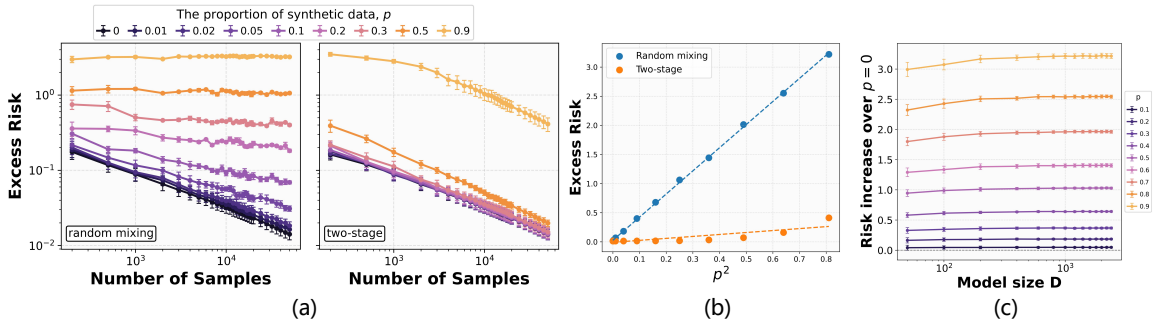

Figure 1: **Simulation results.** (a) Excess risk of two protocols in full high-dimensional linear model: mixed training exhibits a clear p-dependent error floor as T scales, while two-stage training continues to improve without saturation. (b) Excess risk at fixed $T = 50000$ versus p^2: under mixing, 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.

Real-data experiment setup We conduct experiments on the `CIFAR10` [35] 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 $\widetilde{y} = q_{\text{teacher}}(y \mid x)$, , 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.

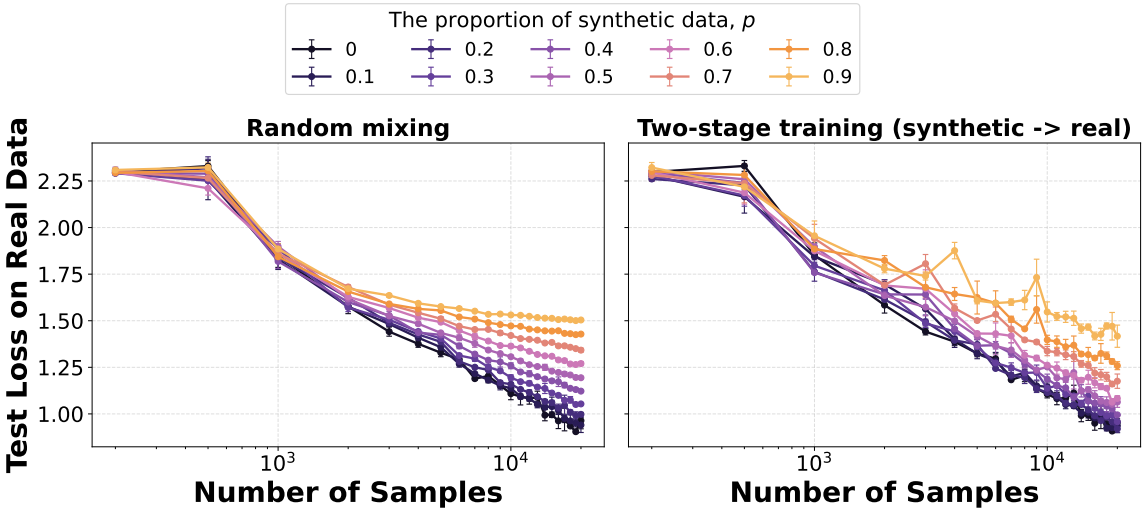

Figure 2: **Real-data experiments of different training protocols.** Test cross-entropy loss on the real distribution as a function of the number of training samples T, under different synthetic proportions p. **Left:** random mixing of real and synthetic data. **Right:** two-stage training (synthetic-data training followed by real-data training).

From Figure 2, 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.

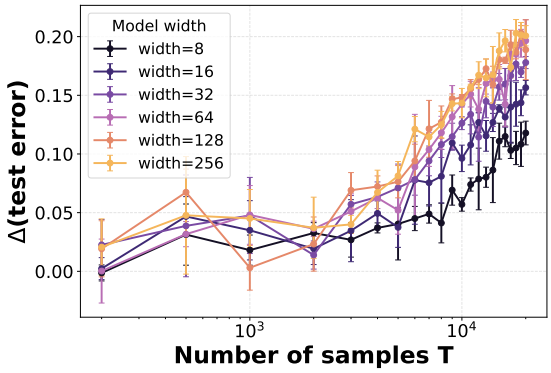

Figure 3: Test error difference

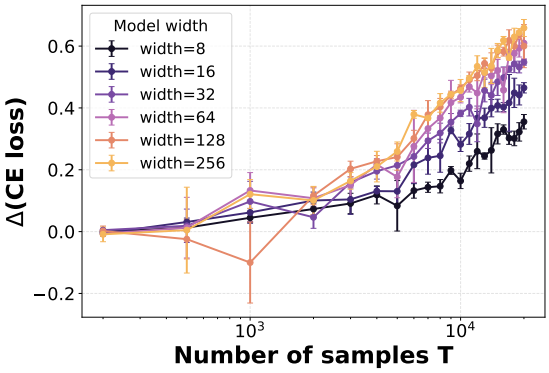

Figure 4: Test cross-entropy difference

Figure 5: **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 5 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.