OPT2025: 17th Annual Workshop on Optimization for Machine Learning

A Simplified Analysis of SGD for Linear Regression with Weight Averaging

Alexandru Meterez* AMETEREZ@G.HARVARD.EDU
Harvard University, Kempner Institute
Depen Morwani* DMORWANI@G.HARVARD.EDU
Harvard University, Kempner Institute
Costin-Andrei Oncescu
Harvard University
Jingfeng Wu
University of California, Berkeley
Cengiz Pehlevan
Harvard University, Kempner Institute
Sham Kakade
Harvard University, Kempner Institute
**: Equal contribution*

Abstract

Theoretically understanding stochastic gradient descent (SGD) in overparameterized models has led to the development of several optimization algorithms that are widely used in practice today. Recent work by Zou et al. [7] provides sharp rates for SGD optimization in linear regression using constant learning rate, both with and without tail iterate averaging, based on a bias-variance decomposition of the risk. In our work, we provide a simplified analysis recovering the same bias and variance bounds provided in [7] based on simple linear algebra tools, bypassing the requirement to manipulate operators on positive semi-definite (PSD) matrices. We believe our work makes the analysis of SGD on linear regression very accessible and will be helpful in further analyzing mini-batching and learning rate scheduling, leading to improvements in the training of realistic models.

1. Problem setup and notation

We use bolded small letters for vectors and bolded capital letters for matrices, and we use $\mathbf{1}$ for a vector of ones. Moreover, for any vector $\mathbf{v}$, we denote as $\mathbf{v}^k$ the operation of raising each entry of $\mathbf{v}$ to the power k, and we use $\preceq$ to denote PSD ordering. Moreover for inequalities of the form $\mathbf{v}_1 \leqslant \mathbf{v}_2$, we refer to an elementwise inequality between the two vectors. For a vector $\mathbf{v}$, we also denote the norm induced by matrix $\mathbf{A}$ as $\|\mathbf{v}\|_{\mathbf{A}}^2 = \mathbf{v}^\top \mathbf{A}\mathbf{v}$. We consider the eigenvalues ordered as $\lambda_{\max} = \lambda_1$.

We are going to consider the linear regression setting with Gaussian inputs and homoscedastic noise additive and well specified setting. Let $\mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ be the independent noise, where $\mathbf{x} \in \mathbb{R}^d$ and $\epsilon \in \mathbb{R}$. Denote $y = \langle \mathbf{w}^\star, \mathbf{x} \rangle + \epsilon$ the ground truth label of $\mathbf{x}$. The iterate at time t for SGD will be denoted by $\mathbf{w}_t$. In the case of tail averaging, we use the bar notation for the weights i.e. $\bar{\mathbf{w}}_t = \frac{1}{t-s}\sum_{i=s}^{t} \mathbf{w}_i$ where each $\mathbf{w} \in \mathbb{R}^d$, with s denoting the iterations that we are not averaging over. We will consider $\mathbf{w}_0$ to be fixed. We will denote the covariance of the iterates at time t as $\mathbf{\Sigma}_t = \mathbb{E}[(\mathbf{w}_t - \mathbf{w}^\star)(\mathbf{w}_t - \mathbf{w}^\star)^\top]$, where the

© A. Meterez*, D. Morwani*, C.-A. Oncescu, J. Wu, C. Pehlevan & S. Kakade.

expectation is taken over the noise induced by the SGD algorithm. We will denote the eigendecomposition of $\mathbf{H} := \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}$. Note that $\mathbf{H}$ is PSD, as it is the covariance matrix of the input samples $\mathbf{x}$. Finally, The mean squared error of a given weight vector $\mathbf{w}$ is denoted by $L(\mathbf{w}) = \frac{1}{2}\mathbb{E}[(\langle \mathbf{w}, \mathbf{x}\rangle - y)^2]$.

The following document is structured as:

- Section 2 contains the gist for the simplified analysis to go through for computing the bias and variance bounds
- Section 3 contains the bias variance bounds for the case with tail averaging for the last N iterates

2. Main Technique

First, we will simplify the risk expression for the iterate at time t. If we denote as $\lambda = \text{diag}(\mathbf{\Lambda})$ and $\mathbf{m}_t = \text{diag}(\mathbf{M}_t)$, then the risk only depends on the *diagonal elements of* $\mathbf{M}_t$ where $\mathbf{M}_t := \mathbf{Q}^\top \mathbf{\Sigma}_t \mathbf{Q}$, denotes the covariance matrix of the iterates, rotated in the Hessian eigenbasis:

$$L(\mathbf{w}_t) = \frac{1}{2}\mathbb{E}\left[((\mathbf{w}_t - \mathbf{w}^\star)^\top \mathbf{x} + \varepsilon)^2\right] = \frac{1}{2}\text{Tr}(\mathbf{\Lambda}\mathbf{M}_t) + \frac{1}{2}\sigma^2 = \frac{1}{2}\langle \lambda, \mathbf{m}_t\rangle + \frac{1}{2}\sigma^2$$

The equation above shows that, instead of tracking the entire covariance matrix ($\mathbf{\Sigma}_t$) as done in the previous works [3, 7], we can track $\mathbf{m}_t$ for estimating the risk at time t. Note that a similar technique can be used in the tail averaging case - and we will expand on this in Section 3, since the risk of the averaged iterates also depends only on the vectors $\mathbf{m}_t$, for the times that are included in the tail averaging. Below, we will further show that for SGD, the evolution of $\mathbf{m}_t$ is only dependent on $\mathbf{m}_{t-1}$, and not on the other elements of $\mathbf{M}_t$. This diagonalization idea has also been used in previous literature [2, 5, 6]. We extend more the discussion regarding related work in Appendix A.

2.1. SGD with batch size 1

We begin with the SGD update rule with learning rate η, and we use this relationship to obtain a recurrence on $\mathbf{\Sigma}_t$:

$$\begin{aligned}
&\mathbf{w}_{t+1} - \mathbf{w}^\star = (\mathbf{I} - \eta \mathbf{x}_t \mathbf{x}_t^\top)(\mathbf{w}_t - \mathbf{w}^\star) - \eta \epsilon_t \mathbf{x}_t \\
&\implies \mathbf{\Sigma}_{t+1} = \mathbf{\Sigma}_t - \eta \mathbf{H}\mathbf{\Sigma}_t - \eta \mathbf{\Sigma}_t \mathbf{H} + \eta^2 \mathbb{E}_{\mathbf{x}_t}[(\mathbf{x}_t^\top \mathbf{\Sigma}_t \mathbf{x}_t)\mathbf{x}_t \mathbf{x}_t^\top] + \eta^2 \sigma^2 \mathbf{H}
\end{aligned}$$

The next step in order to derive the final recursion stems from calculating the fourth moment term. We compute it separately:

Lemma 1 *For* $\mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$*, we have that:*

$$\mathbb{E}_{\mathbf{x}}[(\mathbf{x}^\top \mathbf{\Sigma}_t \mathbf{x})\mathbf{x}\mathbf{x}^\top] = 2\mathbf{H}\mathbf{\Sigma}_t\mathbf{H} + \mathit{Tr}(\mathbf{H}\mathbf{\Sigma}_t)\mathbf{H}$$

We defer the proof of Lemma 1 to Appendix D.

Therefore by applying Lemma 1, we get the recurrence:

$$\mathbf{\Sigma}_{t+1} = \mathbf{\Sigma}_t - \eta \mathbf{\Sigma}_t \mathbf{H} - \eta \mathbf{H}\mathbf{\Sigma}_t + 2\eta^2 \mathbf{H}\mathbf{\Sigma}_t\mathbf{H} + \eta^2 \text{Tr}(\mathbf{H}\mathbf{\Sigma}_t)\mathbf{H} + \eta^2\sigma^2\mathbf{H}$$

Note that the above recurrence was also derived in various previous works including Bordelon and Pehlevan [2], Jain et al. [3], Zou et al. [7]. Similar ideas for SGD dynamics have also been studied from a random matrix theory point of view by modelling the risk as a Volterra equation [1, 4] in the high dimensional case. The crucial thing that we will show below is that the recurrence for $\mathbf{m}_t := \text{diag}(\mathbf{Q}^\top \mathbf{\Sigma}_t \mathbf{Q})$ is dependent only on $\mathbf{m}_{t-1}$.

Rotating both the sides of the equation above in the basis of $\mathbf{H}$, i.e, multiplying on the left by $\mathbf{Q}^\top$ and on the right by $\mathbf{Q}$ we get the following recurrence:

$$\mathbf{M}_{t+1} = \mathbf{M}_t - \eta \mathbf{M}_t \mathbf{\Lambda} - \eta \mathbf{\Lambda} \mathbf{M}_t + 2\eta^2 \mathbf{\Lambda} \mathbf{M}_t \mathbf{\Lambda} + \eta^2 \text{Tr}(\mathbf{\Lambda} \mathbf{M}_t) \mathbf{\Lambda} + \eta^2 \sigma^2 \mathbf{\Lambda}$$

Pushing a diag operator on both sides and rearranging the terms we get:

$$\mathbf{m}_{t+1} = \left(\mathbf{I} - 2\eta \mathbf{\Lambda} + 2\eta^2 \mathbf{\Lambda}^2 + \eta^2 \lambda \lambda^\top\right) \mathbf{m}_t + \eta^2 \sigma^2 \lambda \tag{1}$$

Thus, we can see that the recurrence for $\mathbf{m}_t$ only depends on $\mathbf{m}_{t-1}$. Note that the technique above also holds for general distributions with an assumption on the fourth moment operator (Assumption 2.2 in Zou et al. [7]), but in this note, we will restrict ourselves to Gaussian input distribution. In the next section, we will derive the precise bounds of Zou et al. [7] by using the recursion above, combined with the fact that the risk only depends on $\mathbf{m}_t$.

3. Tail averaging bounds

Before proceeding with the bounds for the tail-averaged iterate, we will simplify the equations above and split the recursion in a 'bias' and 'variance' iterate.

3.1. Bias-Variance decomposition

Denoting: $\mathbf{A} = \mathbf{I} - 2\eta \mathbf{\Lambda} + 2\eta^2 \mathbf{\Lambda}^2 + \eta^2 \lambda \lambda^\top$, we can unroll the recursion in equation 1 and get:

$$\mathbf{m}_{t+1} = \mathbf{A}^{t+1} \mathbf{m}_0 + \eta^2 \sigma^2 \left(\sum_{i=0}^{t} \mathbf{A}^i \right) \lambda := \tilde{\mathbf{m}}_{t+1} + \bar{\mathbf{m}}_{t+1}$$

where bias $:= \tilde{\mathbf{m}}_{t+1}$ and variance $:= \bar{\mathbf{m}}_{t+1}$. The nomenclature comes from the fact that the bias term can be recovered by starting the SGD algorithm from $\mathbf{w}_0$ in the noiseless setting (i.e. $\sigma^2 = 0$), whereas the variance term can be recovered by starting the process from $\mathbf{w}^\star$ (in which case $\tilde{\mathbf{m}}_0 = 0$), and we are stuck in a ball where the radius is controlled by the variance of the additive noise.

Lemma 2 (A bound) *For* $\mathbf{A} = \mathbf{I} - 2\eta \mathbf{\Lambda} + 2\eta^2 \mathbf{\Lambda}^2 + \eta^2 \lambda \lambda^\top$ *and* $\mathbf{B} = (\mathbf{I} - \eta \mathbf{\Lambda})^2 + 2\eta^2 \lambda \lambda^\top$*, we have that* $\mathbf{A} \leqslant \mathbf{B}$ *elementwise, hence,* $\mathbf{A}\mathbf{v} \leqslant \mathbf{B}\mathbf{v}$ *for any vector* $\mathbf{v} \geqslant 0$*.*

The proof is deferred to Lemma 2 in Appendix D. In general, we will work with a general $\alpha > 0$ and a matrix $\mathbf{A}$ given by

$$\mathbf{A} := (\mathbf{I} - \eta\mathbf{\Lambda})^2 + \alpha\eta^2\lambda\lambda^\top$$

In Appendix E, we show that, for a general batch size b, the α defined above varies as $2/b$. Moreover, we will assume that the learning rate $\eta \leqslant \frac{1}{\lambda_{max}(\mathbf{H})+\alpha\mathrm{Tr}(\mathbf{H})}$. This matches the previous works [3], where at batch size 1, the learning rate is close to $\frac{1}{\mathrm{Tr}(\mathbf{H})}$, but as batch size goes to ∞, learning rate becomes close to $\frac{1}{\lambda_{max}}$.

3.2. Simplification of tail averaged loss

In this section we reuse some of the above calculation, but we instead compute the bias and variance bounds for the case of tail averaging for the last N iterates. We begin with the following preliminary analysis, before specializing towards bounding the 2 separate terms:

$$\begin{aligned}
L(\bar{\mathbf{w}}_{s:s+N}) &= \frac{1}{2}\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=s}^{s+N-1}\langle \mathbf{w}_i - \mathbf{w}^\star, \mathbf{x}_t\rangle - \epsilon\right)^2\right] \\
&\leqslant \frac{1}{\eta N^2}\left[\sum_{i=s}^{s+N-1}\mathrm{Tr}\left(\mathbf{\Sigma}_i(\mathbf{I} - (\mathbf{I}-\eta\mathbf{H})^{s+N-i})\right)\right] + \frac{\sigma^2}{2} \\
&\leqslant \frac{1}{\eta N^2}\langle \sum_{i=s}^{s+N-1}\mathbf{m}_i, \mathbf{1} - (\mathbf{1}-\eta\lambda)^N\rangle + \frac{\sigma^2}{2}
\end{aligned}$$

Note that we provide a full derivation of the above inequality in Appendix B. Similar to before, now we need to separately bound the bias and variance terms, which make up the $\mathbf{m}_t$ term. We use the same notation as in the previous section.

3.3. Bounding the bias term

We begin by stating the main result of this section, which is the bound for the bias term. The proof is provided in the remainder of the section:

Theorem 3 *For threshold $k^\star = \max\{j : \lambda_j \geqslant \frac{1}{\eta N}\}$, the bias contribution to the risk is bounded by:*

$$\begin{aligned}
\textit{bias} \leqslant \frac{1}{\eta^2 N^2}\|(\mathbf{I}-\eta\mathbf{\Lambda})^s(\mathbf{w}_0 - \mathbf{w}^\star)\|^2_{\mathbf{\Lambda}^{-1}_{0:k^\star}} + 4\|(\mathbf{I}-\eta\mathbf{\Lambda})^s(\mathbf{w}_0 - \mathbf{w}^\star)\|^2_{\mathbf{\Lambda}_{k^\star:\infty}} + \\
+ \frac{\alpha(\|\mathbf{w}_0 - \mathbf{w}^\star\|^2_{\mathbf{I}_{0:k^\star}} + 2(s+N)\eta\|\mathbf{w}_0 - \mathbf{w}^\star\|^2_{\mathbf{\Lambda}_{k^\star:\infty}})}{\eta N(1-\eta\alpha \mathit{Tr}(\mathbf{H}))} \cdot \frac{k^\star + 4\eta^2 N^2 \sum_{j>k^\star}\lambda_j^2}{N}
\end{aligned}$$

We know $\tilde{\mathbf{m}}_i = \mathbf{A}\tilde{\mathbf{m}}_{i-1}$. Applying Lemma 2, we have:

$$\tilde{\mathbf{m}}_t \leqslant (\mathbf{I}-\eta\mathbf{\Lambda})^2\tilde{\mathbf{m}}_{t-1} + \alpha\eta^2\lambda\lambda^\top\tilde{\mathbf{m}}_{t-1}$$

If we denote by $\mathbf{D} = (\mathbf{I}-\eta\mathbf{\Lambda})^2$ and by $c_t = \alpha\eta^2\lambda^\top\tilde{\mathbf{m}}_{t-1}$ we have the recurrence:

$$\tilde{\mathbf{m}}_t = \mathbf{D}^t \tilde{\mathbf{m}}_0 + \sum_{i=0}^{t-1} c_{t-i} \mathbf{D}^i \lambda$$

First, we provide a bound on c_t in the recurrence above.

Lemma 4 **(c_t bound)** *Let $s_t = \langle \tilde{\mathbf{m}}_t, \mathbf{1} \rangle$ and $r_t = \langle \tilde{\mathbf{m}}_t, \lambda \rangle$. Then, for $c_t = \alpha\eta^2 \lambda^\top \tilde{\mathbf{m}}_{t-1}$, we have that:*

$$\sum_{t=0}^{k-1} c_t \leqslant \frac{\alpha\eta(s_0 - s_t)}{1 - \eta\alpha Tr(\mathbf{H})}$$

Finally, using Lemma 4, we can derive an iterate bound for the bias term:

Theorem 5 (Bias iterate bound) *For all quantities defined as before, we can upper bound the bias iterate as:*

$$\tilde{\mathbf{m}}_t \leqslant \mathbf{D}^t \tilde{\mathbf{m}}_0 + \frac{\alpha\eta(s_0 - s_t)}{1 - \eta\alpha Tr(\mathbf{H})} \lambda$$

Tail averaging Finally, applying Lemma 5, we can write down the bound for the tail averaging case:

$$\sum_{i=s}^{s+N-1} \tilde{\mathbf{m}}_i \leqslant \mathbf{D} \sum_{i=s}^{s+N-2} \tilde{\mathbf{m}}_i + \alpha\eta\lambda \frac{s_s - s_{s+N}}{1 - \eta\alpha \mathrm{Tr}(\mathbf{H})} + \tilde{\mathbf{m}}_s$$

Denote $S_{s:s+N} = \sum_{i=s}^{s+N-1} \tilde{\mathbf{m}}_i$, we have:

$$\begin{aligned}
S_{s:s+N} &\leqslant \mathbf{D} S_{s:s+N-1} + \alpha\eta\lambda \frac{s_s - s_{s+N}}{1 - \eta\alpha \mathrm{Tr}(\mathbf{H})} + \tilde{\mathbf{m}}_s \\
&\leqslant \sum_{k=0}^{N-1} \mathbf{D}^k \left(\alpha\eta \frac{s_s - s_{s+N}}{1 - \eta\alpha \mathrm{Tr}(\mathbf{H})} \lambda + \tilde{\mathbf{m}}_s \right) \\
&= \sum_{k=0}^{N-1} \mathbf{D}^k \tilde{\mathbf{m}}_s + \sum_{k=0}^{N-1} \alpha\eta \frac{s_s - s_{s+N}}{1 - \eta\alpha \mathrm{Tr}(\mathbf{H})} \mathbf{D}^k \lambda
\end{aligned}$$

In order to bound the first term, we invoke Lemma 5 and get:

$$\begin{aligned}
\sum_{k=0}^{N-1} \mathbf{D}^k \tilde{\mathbf{m}}_s &\leqslant \sum_{k=0}^{N-1} \mathbf{D}^k \left(\mathbf{D}^s \tilde{\mathbf{m}}_0 + \frac{\alpha\eta(s_0 - s_s)}{1 - \eta\alpha \mathrm{Tr}(\mathbf{H})} \lambda \right) \\
&= \sum_{k=0}^{N-1} \mathbf{D}^{k+s} \tilde{\mathbf{m}}_0 + \sum_{k=0}^{N-1} \frac{\alpha\eta(s_0 - s_s)}{1 - \eta\alpha \mathrm{Tr}(\mathbf{H})} \mathbf{D}^k \lambda
\end{aligned}$$

Lemma 6 ($s_0 - s_{s+N}$ **bound)** *For s_i defined as before, we have that:*

$$s_0 - s_{s+N} \leqslant \langle \tilde{\mathbf{m}}_0, \mathbf{1} - (\mathbf{1} - \eta\lambda)^{2(s+N)} \rangle$$

We defer the proof of Lemma 6 to Appendix D.

Plugging back in the sum and assembling the terms we get the final bound:

$$\begin{aligned} S_{s:s+N} &\leqslant \sum_{k=0}^{N-1} \mathbf{D}^{k+s} \tilde{\mathbf{m}}_0 + \sum_{k=0}^{N-1} \frac{\alpha\eta(s_0 - s_{s+N})}{1 - \eta\alpha \mathrm{Tr}(\mathbf{H})} \mathbf{D}^k \lambda \\ &\leqslant \underbrace{\sum_{k=0}^{N-1} \mathbf{D}^{k+s} \tilde{\mathbf{m}}_0}_{I_1} + \underbrace{\frac{\alpha\eta \langle \tilde{\mathbf{m}}_0, \mathbf{1} - (\mathbf{1} - \eta\lambda)^{2(s+N)} \rangle}{1 - \eta\alpha \mathrm{Tr}(\mathbf{H})} \sum_{k=0}^{N-1} \mathbf{D}^k \lambda}_{I_2} \qquad \text{Lemma 6} \end{aligned}$$

Now plugging this into the contribution to the risk formula, we end up with the 2 terms, dotted into $\mathbf{1} - (\mathbf{1} - \eta\lambda)^N$. We defer this bound to Appendix C.

3.4. Bounding the variance term

As before, we begin this section by stating the main result, which we further prove in the next section.

Theorem 7 *For thresholds $k^\star = \max\{j : \lambda_j \geqslant \frac{1}{\eta N}\}$ and $k^\dagger = \max\{j : \lambda_j \geqslant \frac{1}{\eta(s+N)}\}$, the variance contribution to the risk is bounded by:*

$$\textit{variance} \leqslant \frac{\sigma^2}{1 - \eta\alpha \mathit{Tr}(\mathbf{H})} \left(\frac{k^\star}{N} + 4\eta \sum_{k^\star < j \leqslant k^\dagger} \lambda_j + 16\eta^2 (s+N) \sum_{j \geqslant k^\dagger} \lambda_j \right)$$

Remark 8 *For general batch size b, the bound stated above further scales down as $1/b$ as explained in Appendix E.*

We defer the proof of Theorem 7 to Appendix D.

4. Discussion

In this work we have provided a mathematically simplified technique of deriving existing SGD rates in the optimization literature [3, 7]. While previous works make use of operators on PSD matrices tracking the whole iterate covariance, our derivation shows that it suffices to only track the diagonal of this matrix (rotated in the eigenbasis of the data), thus reducing the complexity of the analysis to simple linear algebra. However, while we recover the upper bound from Zou et al. [7], this upper bound is not sharp: namely, Zou et al. [7] prove a tighter lower bound for tailed weight averaging (provided in Appendix G for completeness) and conjecture that the upper bound should also be improvable. We hope that our simplified framework will inspire future work to tackle this improvement. Furthermore, our analysis can be extended to various learning rate scheduler schemes, thus possibly leading to improvements in the optimization of realistic models.

References

[1] Alexander Atanasov, Blake Bordelon, Jacob A Zavatone-Veth, Courtney Paquette, and Cengiz Pehlevan. Two-point deterministic equivalence for stochastic gradient dynamics in linear models. *arXiv preprint arXiv:2502.05074*, 2025.

[2] Blake Bordelon and Cengiz Pehlevan. Learning curves for sgd on structured features. *arXiv preprint arXiv:2106.02713*, 2021.

[3] 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.

[4] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+ 3 phases of compute-optimal neural scaling laws. *arXiv preprint arXiv:2405.15074*, 2024.

[5] Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter L Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? *arXiv preprint arXiv:2310.08391*, 2023.

[6] Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Finite-sample analysis of learning high-dimensional single relu neuron. In *International Conference on Machine Learning*, pages 37919–37951. PMLR, 2023.

[7] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on Learning Theory*, pages 4633–4635. PMLR, 2021.

Appendix A. Related Work

[6] considered (misspecified) ReLU regression problems, in which the PSD operator methods failed to be applied directly due to certain non-symmetric matrices (as a result of the non-linearity of ReLU coupled with misspecified noise). They addressed this issue by applying the PSD operator method only to the diagonals of the relevant matrices [6, Section 7]. This idea is further extended by Wu et al. [5] in a matrix linear regression problem (motivated by in-context learning of linear regression). In their problem, the trainable parameters are a matrix (second-order tensor), so their covariance is a fourth-order tensor that is updated by a linear operator given by an eighth-order tensor. Such high-order tensors are hard to analyze directly; instead, Wu et al. [5] considered the evolution of the "diagonals" of the fourth-order tensors (corresponding to second-order tensors), reducing the problem to bounding a sequence of matrices updated by PSD operators [6, Section 6].

Appendix B. Extended derivation of the loss for tail averaging

In this section, we provide an extended derivation of the risk for the tail averaging case. First, we begin by expanding the expected value of the loss:

$$
\begin{aligned}
L(\bar{\mathbf{w}}_{s:s+N}) &= \frac{1}{2}\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=s}^{s+N-1}\langle \mathbf{w}_i - \mathbf{w}^\star, \mathbf{x}_t\rangle - \epsilon\right)^2\right] \\
&= \frac{1}{2N^2}\mathbb{E}\left[\sum_{i=s}^{s+N-1}\mathbf{x}_t^\top(\mathbf{w}_i-\mathbf{w}^\star)(\mathbf{w}_i-\mathbf{w}^\star)^\top\mathbf{x}_t\right] + \frac{1}{N^2}\mathbb{E}\left[\sum_{i=s}^{s+N-1}\sum_{j=i+1}^{s+N-1}\mathbf{x}_t^\top(\mathbf{w}_i-\mathbf{w}^\star)(\mathbf{w}_j-\mathbf{w}^\star)^\top\mathbf{x}_t\right] + \frac{\sigma^2}{2} \\
&\leqslant \frac{1}{N^2}\mathbb{E}\left[\sum_{i=s}^{s+N-1}\mathbf{x}_t^\top(\mathbf{w}_i-\mathbf{w}^\star)(\mathbf{w}_i-\mathbf{w}^\star)^\top\mathbf{x}_t\right] + \frac{1}{N^2}\mathbb{E}\left[\sum_{i=s}^{s+N-1}\sum_{j=i+1}^{s+N-1}\mathbf{x}_t^\top(\mathbf{w}_i-\mathbf{w}^\star)(\mathbf{w}_j-\mathbf{w}^\star)^\top\mathbf{x}_t\right] + \frac{\sigma^2}{2} \\
&= \frac{1}{N^2}\mathbb{E}\left[\sum_{i=s}^{s+N-1}\sum_{j=i}^{s+N-1}\mathbf{x}_t^\top(\mathbf{w}_i-\mathbf{w}^\star)(\mathbf{w}_j-\mathbf{w}^\star)^\top\mathbf{x}_t\right] + \frac{\sigma^2}{2} \\
&= \frac{1}{\eta N^2}\left[\sum_{i=s}^{s+N-1}\operatorname{Tr}\left(\mathbf{\Sigma}_i\mathbf{H}(\mathbf{I}-(\mathbf{I}-\eta\mathbf{H})^{s+N-i})\mathbf{H}^{-1}\right)\right] + \frac{\sigma^2}{2} \\
&= \frac{1}{\eta N^2}\left[\sum_{i=s}^{s+N-1}\operatorname{Tr}\left(\mathbf{\Sigma}_i(\mathbf{I}-(\mathbf{I}-\eta\mathbf{H})^{s+N-i})\right)\right] + \frac{\sigma^2}{2}
\end{aligned}
$$

So if we do again the decomposition $\mathbf{H} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, we get:

$$
\begin{aligned}
(\mathbf{I}-(\mathbf{I}-\eta\mathbf{H})^{s+N-i}) &= \mathbf{Q}\mathbf{I}\mathbf{Q}^\top - (\mathbf{Q}\mathbf{I}\mathbf{Q}^\top - \mathbf{Q}\eta\mathbf{\Lambda}\mathbf{Q}^\top)^{s+N-i} \\
&= \mathbf{Q}\mathbf{I}\mathbf{Q}^\top - \mathbf{Q}(\mathbf{I}-\eta\mathbf{\Lambda})^{s+N-i}\mathbf{Q}^\top \\
&= \mathbf{Q}\left(\mathbf{I}-(\mathbf{I}-\eta\mathbf{\Lambda})^{s+N-i}\right)\mathbf{Q}^\top
\end{aligned}
$$

Plugging this identity in the previous inequality and using the cyclic property of the trace, followed by the fact that we only need to consider the diagonal we obtain:

$$
\begin{aligned}
L(\bar{\mathbf{w}}_{s:s+N}) &\leqslant \frac{1}{\eta N^2}\left[\sum_{i=s}^{s+N-1} \mathrm{Tr}(\mathbf{M}_i(\mathbf{I}-(\mathbf{I}-\eta\mathbf{\Lambda})^{s+N-i}))\right] + \frac{\sigma^2}{2} \\
&= \frac{1}{\eta N^2}\sum_{i=s}^{s+N-1} \langle \mathbf{m}_i, \mathbf{1}-(\mathbf{1}-\eta\lambda)^{s+N-i}\rangle + \frac{\sigma^2}{2} \\
&\leqslant \frac{1}{\eta N^2}\langle \sum_{i=s}^{s+N-1} \mathbf{m}_i, \mathbf{1}-(\mathbf{1}-\eta\lambda)^{N}\rangle + \frac{\sigma^2}{2}
\end{aligned}
$$

Appendix C. Bounding I_1 and I_2 in the bias term

In this section, we provided the extended bound for the I_1 and I_2 terms from the bias, dotted into $\mathbf{1}-(\mathbf{1}-\eta\lambda)^N$ in order to obtain the bias contribution to the excess risk.

$$
\begin{aligned}
\frac{1}{\eta N^2}\langle I_1, \mathbf{1}-(\mathbf{1}-\eta\lambda)^N\rangle &= \frac{1}{\eta N^2}\langle \sum_{k=0}^{N-1} \mathbf{D}^{k+s}\tilde{\mathbf{m}}_0, \mathbf{1}-(\mathbf{1}-\eta\lambda)^N\rangle \\
&= \frac{1}{\eta N^2}\sum_{k=0}^{N-1}\langle \mathbf{D}^s\tilde{\mathbf{m}}_0, \mathbf{D}^k(\mathbf{1}-(\mathbf{1}-\eta\lambda)^N)\rangle \\
&= \frac{1}{\eta N^2}\sum_{k=0}^{N-1}\langle \omega, \mathbf{D}^k(\mathbf{1}-(\mathbf{1}-\eta\lambda)^N)\rangle \\
&= \frac{1}{\eta N^2}\langle \omega, (\mathbf{I}-\mathbf{D}^N)(\mathbf{I}-\mathbf{D})^{-1}(\mathbf{1}-(\mathbf{1}-\eta\lambda)^N)\rangle \\
&= \frac{1}{\eta N^2}\sum_j \omega_j(1-(1-\eta\lambda_j)^{2N})(1-(1-\eta\lambda_j)^N)(1-(1-\eta\lambda_j)^2)^{-1} \\
&\leqslant \frac{1}{\eta^2 N^2}\sum_j \frac{\omega_j}{\lambda_j}\left[1-(1-\eta\lambda_j)^{2N}\right]^2 \\
&\leqslant \frac{1}{\eta^2 N^2}\sum_{j\leqslant k^\star}\frac{\omega_j}{\lambda_j} + 4\sum_{j>k^\star}\omega_j\lambda_j \\
&= \frac{1}{\eta^2 N^2}\|(\mathbf{I}-\eta\mathbf{\Lambda})^s(\mathbf{w}_0-\mathbf{w}^\star)\|^2_{\mathbf{\Lambda}^{-1}_{0:k^\star}} + 4\|(\mathbf{I}-\eta\mathbf{\Lambda})^s(\mathbf{w}_0-\mathbf{w}^\star)\|^2_{\mathbf{\Lambda}_{k^\star:\infty}}
\end{aligned}
$$

where $\omega = \mathbf{D}^s\tilde{\mathbf{m}}_0$ and we have defined $k^\star = \max\{k : \lambda_k \geqslant \frac{1}{2\eta N}\}$.

For the contribution of the second term to the risk, we proceed in a similar fashion.

$$
\begin{aligned}
\frac{1}{\eta N^2}\langle I_2, \mathbf{1}-(\mathbf{1}-\eta\lambda)^N\rangle &= \frac{\alpha\langle \tilde{\mathbf{m}}_0, \mathbf{1}-(\mathbf{1}-\eta\lambda)^{2(s+N)}\rangle}{N^2(1-\eta\alpha\mathrm{Tr}(\mathbf{H}))}\sum_{k=0}^{N-1}\langle \mathbf{1}-(\mathbf{1}-\eta\lambda)^N, \mathbf{D}^k\lambda\rangle \\
&= \frac{\alpha\langle \tilde{\mathbf{m}}_0, \mathbf{1}-(\mathbf{1}-\eta\lambda)^{2(s+N)}\rangle}{N^2(1-\eta\alpha\mathrm{Tr}(\mathbf{H}))}\langle \mathbf{1}-(\mathbf{1}-\eta\lambda)^N, (\mathbf{I}-\mathbf{D}^N)(\mathbf{I}-\mathbf{D})^{-1}\lambda\rangle \\
&\leqslant \frac{\alpha\langle \tilde{\mathbf{m}}_0, \mathbf{1}-(\mathbf{1}-\eta\lambda)^{2(s+N)}\rangle}{\eta N^2(1-\eta\alpha\mathrm{Tr}(\mathbf{H}))}\langle \mathbf{1}-(\mathbf{1}-\eta\lambda)^{2N}, \mathbf{1}-(\mathbf{1}-\eta\lambda)^{2N}\rangle \\
&\leqslant \frac{\alpha\langle \tilde{\mathbf{m}}_0, \mathbf{1}-(\mathbf{1}-\eta\lambda)^{2(s+N)}\rangle}{\eta(1-\eta\alpha\mathrm{Tr}(\mathbf{H}))}\left(\frac{k^\star}{N^2}+4\eta^2\sum_{j>k^\star}\lambda_j^2\right) \\
&\leqslant \frac{\alpha(\|\mathbf{w}_0-\mathbf{w}^\star\|^2_{\mathbf{I}_{0:k^\star}}+2(s+N)\eta\|\mathbf{w}_0-\mathbf{w}^\star\|^2_{\mathbf{\Lambda}_{k^\star:\infty}})}{\eta(1-\eta\alpha\mathrm{Tr}(\mathbf{H}))}\left(\frac{k^\star}{N^2}+4\eta^2\sum_{j>k^\star}\lambda_j^2\right)
\end{aligned}
$$

where again we have split at the same $k^\star$ as before, and bounded as:

$$
\begin{aligned}
\langle \mathbf{1}-(\mathbf{1}-\eta\lambda)^{2N}, \mathbf{1}-(\mathbf{1}-\eta\lambda)^{2N}\rangle &\leqslant k^\star + 4N^2\eta^2\sum_{j>k^\star}\lambda_j^2 \\
\langle \tilde{\mathbf{m}}_0, \mathbf{1}-(\mathbf{1}-\eta\lambda)^{2(s+N)}\rangle &\leqslant \sum_{j\leqslant k^\star}\tilde{\mathbf{m}}_0^j + \sum_{j>k^\star}\tilde{\mathbf{m}}_0^j 2(s+N)\eta\lambda_j \\
&= \|\mathbf{w}_0-\mathbf{w}^\star\|^2_{\mathbf{I}_{0:k^\star}} + 2(s+N)\eta\|\mathbf{w}_0-\mathbf{w}^\star\|^2_{\mathbf{\Lambda}_{k^\star:\infty}}
\end{aligned}
$$

Appendix D. Additional proofs

Proof [Proof of Lemma 1] Since $\mathbf{P} = (\mathbf{w}_t - \mathbf{w}^\star)(\mathbf{w}_t - \mathbf{w}^\star)^\top$ is independent of $\mathbf{x}_t$, we can take the expectation conditioned on $\mathbf{P}$, then apply tower rule. For any centered Gaussian vector, Iserlis theorem gives us that:

$$
\mathbb{E}[\mathbf{x}_i\mathbf{x}_k\mathbf{x}_l\mathbf{x}_j] = \mathbf{H}_{ik}\mathbf{H}_{lj} + \mathbf{H}_{il}\mathbf{H}_{kj} + \mathbf{H}_{ij}\mathbf{H}_{kl}
$$

Thus we get that (skipping the t index for $\mathbf{x}_t$ to simplify notation):

$$
\begin{aligned}
\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{P}\mathbf{x}\mathbf{x}^\top]_{ij} &= \sum_{k,l}\mathbf{P}_{kl}(\mathbf{H}_{ik}\mathbf{H}_{lj} + \mathbf{H}_{il}\mathbf{H}_{kj} + \mathbf{H}_{ij}\mathbf{H}_{kl}) \\
&= 2(\mathbf{H}\mathbf{P}\mathbf{H})_{ij} + \mathrm{Tr}(\mathbf{H}\mathbf{P})\mathbf{H}_{ij}
\end{aligned}
$$

Reassembling and applying tower rule we get:

$$
\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top(\mathbf{w}_t-\mathbf{w}^\star)(\mathbf{w}_t-\mathbf{w}^\star)^\top\mathbf{x}_t\mathbf{x}_t^\top] = 2\mathbf{H}\mathbf{\Sigma}_t\mathbf{H} + \mathrm{Tr}(\mathbf{H}\mathbf{\Sigma}_t)\mathbf{H}
$$

■

Proof [Proof of Lemma 2] We can upper bound $\mathbf{A}$ as:

$$\begin{aligned}
\mathbf{A} &= (\mathbf{I} - \eta\mathbf{\Lambda})^2 + \eta^2(\mathbf{\Lambda}^2 + \lambda\lambda^\top) \\
&\leqslant (\mathbf{I} - \eta\mathbf{\Lambda})^2 + \eta^2(\lambda\lambda^\top + \lambda\lambda^\top) && \mathbf{\Lambda}^2 \leqslant \lambda\lambda^\top \\
&= \mathbf{B}
\end{aligned}$$

where in the first inequality we have upper bounded $\mathbf{\Lambda}^2$ elementwise since each $\lambda_i \geqslant 0$, being the eigenvalues of the data covariance matrix which is PSD. ■

Proof [Proof of Lemma 4.] We start from:

$$\begin{aligned}
\tilde{\mathbf{m}}_t \leqslant (\mathbf{I} - \eta\mathbf{\Lambda})^2\tilde{\mathbf{m}}_{t-1} + \alpha\eta^2\lambda\lambda^\top\tilde{\mathbf{m}}_{t-1} &\leqslant (\mathbf{I} - \eta\mathbf{\Lambda})\tilde{\mathbf{m}}_{t-1} + \alpha\eta^2\lambda\lambda^\top\tilde{\mathbf{m}}_{t-1} \\
&= \tilde{\mathbf{m}}_{t-1} - \eta\mathbf{\Lambda}\tilde{\mathbf{m}}_{t-1} + \alpha\eta^2\lambda\lambda^\top\tilde{\mathbf{m}}_{t-1}
\end{aligned}$$

Now doing a dot product with $\mathbf{1}$ we get:

$$s_t \leqslant s_{t-1} - \eta r_{t-1} + \alpha\eta^2\mathrm{Tr}(\mathbf{H})r_{t-1} \implies \sum_{t=0}^{k-1} r_t \leqslant \frac{s_0 - s_k}{\eta - \eta^2\alpha\mathrm{Tr}(\mathbf{H})} \implies \sum_{t=0}^{k-1} c_t \leqslant \frac{\alpha\eta(s_0 - s_t)}{1 - \eta\alpha\mathrm{Tr}(\mathbf{H})}$$

■

Proof [Proof of Theorem 5.] Based on the recursion on $\tilde{\mathbf{m}}_t$ we have:

$$\begin{aligned}
\tilde{\mathbf{m}}_t &= \mathbf{D}^t\tilde{\mathbf{m}}_0 + \sum_{i=0}^{t-1} c_{t-i}\mathbf{D}^i\lambda \\
&\leqslant \mathbf{D}^t\tilde{\mathbf{m}}_0 + \sum_{i=0}^{t-1} c_i\lambda && \mathbf{D} \leqslant \mathbf{I} \\
&\leqslant \mathbf{D}^t\tilde{\mathbf{m}}_0 + \frac{\alpha\eta(s_0 - s_t)}{1 - \eta\alpha\mathrm{Tr}(\mathbf{H})}\lambda && \text{Lemma 4}
\end{aligned}$$

■

Proof [Proof of Lemma 6.] First, note that $\mathbf{A} \geqslant \mathbf{D}$ elementwise. Thus, we can bound the following quantity:

$$\tilde{\mathbf{m}}_0 - \tilde{\mathbf{m}}_t = \tilde{\mathbf{m}}_0 - \mathbf{A}^t\tilde{\mathbf{m}}_0 \leqslant \tilde{\mathbf{m}}_0 - \mathbf{D}^t\tilde{\mathbf{m}}_0 = (\mathbf{I} - \mathbf{D}^t)\tilde{\mathbf{m}}_0$$

Using this, we get:

$$s_0 - s_{s+N} \leqslant \langle(\mathbf{I} - \mathbf{D}^t)\tilde{\mathbf{m}}_0, \mathbf{1}\rangle = \langle\tilde{\mathbf{m}}_0, \mathbf{1} - (\mathbf{1} - \eta\lambda)^{2(s+N)}\rangle$$

■

Proof [Proof of Theorem 7.] First note that the recursion for the variance term is:

$$\bar{\mathbf{m}}_t = \eta^2\sigma^2\sum_{i=0}^{t-1}\mathbf{A}^i\lambda$$

Remark 9 *Note that while an explicit variance iterate bound is not necessary to conclude this proof, we do provide such a bound in Lemma 11.*

In order to compute the variance contribution to the risk, we need the following intermediate result obtained via the Sherman–Morrison formula:

$$(\mathbf{I}-\mathbf{A})^{-1}\lambda \leqslant (\mathbf{I}-\mathbf{B})^{-1}\lambda \leqslant \frac{1}{\eta}(\mathbf{\Lambda}-\alpha\eta\lambda\lambda^\top)^{-1}\lambda = \frac{1}{\eta}\frac{1}{1-\eta\alpha\mathrm{Tr}(\mathbf{H})}\mathbf{1}$$

Thus, the variance contribution to the risk becomes:

$$\begin{aligned}
\frac{1}{\eta N^2}\langle \sum_{i=s}^{s+N-1} \bar{\mathbf{m}}_i, \mathbf{1}-(\mathbf{1}-\eta\lambda)^N\rangle &= \frac{\eta\sigma^2}{N^2}\langle \sum_{i=s}^{s+N-1}\sum_{j=0}^{i}\mathbf{A}^j\lambda, \mathbf{1}-(\mathbf{1}-\eta\lambda)^N\rangle \\
&= \frac{\eta\sigma^2}{N^2}\langle \sum_{i=s}^{s+N-1}(\mathbf{I}-\mathbf{A}^{i+1})(\mathbf{I}-\mathbf{A})^{-1}\lambda, \mathbf{1}-(\mathbf{1}-\eta\lambda)^N\rangle \\
&\leqslant \frac{\sigma^2}{N^2(1-\eta\alpha\mathrm{Tr}(\mathbf{H}))}\langle \sum_{i=s}^{s+N-1}(\mathbf{I}-\mathbf{A}^{i+1})\mathbf{1}, \mathbf{1}-(\mathbf{1}-\eta\lambda)^N\rangle \\
&\leqslant \frac{\sigma^2}{N^2(1-\eta\alpha\mathrm{Tr}(\mathbf{H}))}\langle \sum_{i=s}^{s+N-1}\mathbf{1}-(\mathbf{1}-2\eta\lambda)^{2(i+1)}, \mathbf{1}-(\mathbf{1}-\eta\lambda)^N\rangle \\
&\leqslant \frac{\sigma^2}{N(1-\eta\alpha\mathrm{Tr}(\mathbf{H}))}\langle \mathbf{1}-(\mathbf{1}-2\eta\lambda)^{2(s+N)}, \mathbf{1}-(\mathbf{1}-2\eta\lambda)^{2N}\rangle
\end{aligned}$$

Similar to [7], we split into 3 eigenvalue bands, at thresholds $k^\star = \max\{j : \lambda_j \geqslant \frac{1}{\eta N}\}$ and $k^\dagger = \max\{j : \lambda_j \geqslant \frac{1}{\eta(s+N)}\}$. Based on this, we can bound the dot product above as:

$$\begin{aligned}
&(\mathbf{1}-(\mathbf{1}-2\eta\lambda)^{2(s+N)})(\mathbf{1}-(\mathbf{1}-2\eta\lambda)^{2N}) \leqslant 1 && \lambda_i \geqslant \frac{1}{\eta N} \\
&(\mathbf{1}-(\mathbf{1}-2\eta\lambda)^{2(s+N)})(\mathbf{1}-(\mathbf{1}-2\eta\lambda)^{2N}) \leqslant 4\cdot\eta N\lambda_i && \frac{1}{\eta(s+N)} \leqslant \lambda_i < \frac{1}{\eta N} \\
&(\mathbf{1}-(\mathbf{1}-2\eta\lambda)^{2(s+N)})(\mathbf{1}-(\mathbf{1}-2\eta\lambda)^{2N}) \leqslant 16\eta(s+N)\lambda_i\cdot\eta N\lambda_i && \lambda_i < \frac{1}{\eta(s+N)}
\end{aligned}$$

Using these bounds we get

$$\frac{1}{\eta N^2}\langle \sum_{i=s}^{s+N-1}\bar{\mathbf{m}}_i, \mathbf{1}-(\mathbf{1}-\eta\lambda)^N\rangle \leqslant \frac{\sigma^2}{1-\eta\alpha\mathrm{Tr}(\mathbf{H})}\left(\frac{k^\star}{N} + 4\eta\sum_{k^\star<j\leqslant k^\dagger}\lambda_j + 16\eta^2(s+N)\sum_{j\geqslant k^\dagger}\lambda_j\right)$$

■

Appendix E. General batch size result

For a general batch size b, we can show that the following lemma holds:

Lemma 10 *For SGD with a general batch size b, $\mathbf{m}_t$ satisfies the following recurrence*

$$\mathbf{m}_{t+1} = \left[\mathbf{I} - 2\eta\mathbf{\Lambda} + \eta^2\left(1+\frac{1}{b}\right)\mathbf{\Lambda}^2 + \frac{\eta^2}{b}\lambda\lambda^\top\right]\mathbf{m}_t + \frac{\eta^2}{b}\sigma^2\lambda \tag{2}$$

The matrix $\mathbf{A}_b := \left[\mathbf{I} - 2\eta\mathbf{\Lambda} + \eta^2\left(1+\frac{1}{b}\right)\mathbf{\Lambda}^2 + \frac{\eta^2}{b}\lambda\lambda^\top\right]$ is upper bounded by $(\mathbf{I}-\eta\mathbf{\Lambda})^2 + \frac{2}{b}\eta^2\lambda\lambda^\top$, thus falling into our general framework with $\alpha = \frac{2}{b}$. Moreover, we can also see that the variance term recurrence scales down as $\frac{1}{b}$, leading to reduced variance with increasing batch size.

Proof [Proof of Lemma 10.] In the case of minibatching, the update equation is:

$$\mathbf{w}_{t+1} - \mathbf{w}^\star = (\mathbf{I} - \frac{\eta}{b}\sum_{i=1}^{b}\mathbf{x}_i\mathbf{x}_i^\top)(\mathbf{w}_t - \mathbf{w}^\star) - \frac{\eta}{b}\epsilon\sum_{i=1}^{b}\mathbf{x}_i$$

Since the risk depends on the covariance matrix $\mathbf{\Sigma}_t$, we will write down the recursion for $\mathbf{\Sigma}_t$. Note that the cross terms that have an ϵ factor will simplify in expectation, since ϵ is independent from the other terms. Thus, we have:

$$\begin{aligned}
\mathbb{E}[(\mathbf{w}_{t+1} - \mathbf{w}^\star)(\mathbf{w}_{t+1} - \mathbf{w}^\star)^\top] &= \mathbb{E}\left[(\mathbf{I} - \frac{\eta}{b}\sum_{i=1}^{b}\mathbf{x}_i\mathbf{x}_i^\top)(\mathbf{w}_t - \mathbf{w}^\star)(\mathbf{w}_t - \mathbf{w}^\star)^\top(\mathbf{I} - \frac{\eta}{b}\sum_{i=1}^{b}\mathbf{x}_i\mathbf{x}_i^\top)^\top\right] \\
&\quad + \frac{\eta^2}{b^2}\sigma^2\sum_{i,j=1}^{b}\mathbb{E}[\mathbf{x}_i\mathbf{x}_j] \\
\implies \mathbf{\Sigma}_{t+1} &= \mathbf{\Sigma}_t - \eta\mathbf{\Sigma}_t\mathbf{H} - \eta\mathbf{H}\mathbf{\Sigma}_t + \frac{\eta^2}{b^2}\sum_{i,j=1}^{b}\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^\top(\mathbf{w}_t - \mathbf{w}^\star)(\mathbf{w}_t - \mathbf{w}^\star)^\top\mathbf{x}_j\mathbf{x}_j^\top\right]
\end{aligned}$$

Similar to the previous case, we need to compute a 4th moment term in the update of $\mathbf{\Sigma}_t$. We compute it below:

$$\begin{aligned}
\frac{\eta^2}{b^2}\sum_{i,j=1}^{b}\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^\top(\mathbf{w}_t - \mathbf{w}^\star)(\mathbf{w}_t - \mathbf{w}^\star)^\top\mathbf{x}_j\mathbf{x}_j^\top\right] &= \frac{\eta^2}{b}\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^\top(\mathbf{w}_t - \mathbf{w}^\star)(\mathbf{w}_t - \mathbf{w}^\star)^\top\mathbf{x}_i\mathbf{x}_i^\top\right] \\
&\quad + \frac{\eta^2 b(b-1)}{b^2}\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^\top(\mathbf{w}_t - \mathbf{w}^\star)(\mathbf{w}_t - \mathbf{w}^\star)^\top\mathbf{x}_j\mathbf{x}_j^\top\right] \\
&= \frac{\eta^2}{b}(2\mathbf{H}\mathbf{\Sigma}_t\mathbf{H} + \mathrm{Tr}(\mathbf{H}\mathbf{\Sigma}_t)\mathbf{H}) + \frac{\eta^2(b-1)}{b}\mathbf{H}\mathbf{\Sigma}_t\mathbf{H} \\
&= \frac{\eta^2}{b}\left[(b+1)\mathbf{H}\mathbf{\Sigma}_t\mathbf{H} + \mathrm{Tr}(\mathbf{H}\mathbf{\Sigma}_t)\mathbf{H}\right]
\end{aligned}$$

where the first term was computed in the proof of Lemma 1, and the second term comes from factorizing the expectation across the factors since they are independent.

Plugging back in the previous formula for the update of $\mathbf{\Sigma}_t$ we thus get the final formula:

$$\mathbf{\Sigma}_{t+1} = \mathbf{\Sigma}_t - \eta\mathbf{\Sigma}_t\mathbf{H} - \eta\mathbf{H}\mathbf{\Sigma}_t + \eta^2\left(1+\frac{1}{b}\right)\mathbf{H}\mathbf{\Sigma}_t\mathbf{H} + \frac{\eta^2}{b}\mathrm{Tr}(\mathbf{H}\mathbf{\Sigma}_t)\mathbf{H} + \frac{\eta^2}{b}\sigma^2\mathbf{I}$$

Multiplying on the left by $\mathbf{Q}$ and on the right by $\mathbf{Q}^\top$ we obtain the recurrence for $\mathbf{M}_t$:

$$\mathbf{M}_{t+1} = \mathbf{M}_t - \eta\mathbf{M}_t\mathbf{\Lambda} - \eta\mathbf{\Lambda}\mathbf{M}_t + \eta^2\left(1+\frac{1}{b}\right)\mathbf{\Lambda}\mathbf{M}_t\mathbf{\Lambda} + \frac{\eta^2}{b}\mathrm{Tr}(\mathbf{\Lambda}\mathbf{M}_t)\mathbf{\Lambda} + \frac{\eta^2}{b}\sigma^2\mathbf{I}$$

Finally, pushing a diag operator through the whole equation yields the final $\mathbf{A}_b$ expression. ■

Appendix F. Variance iterate bound

In this section we provide a short lemma, giving us a variance iterate bound:

Theorem 11 *For all quantities defined as before, we can bound the variance iterate as:*

$$\bar{\mathbf{m}}_t \leqslant \frac{\eta\sigma^2}{1-\eta\alpha Tr(\mathbf{H})}(\mathbf{1}-(\mathbf{1}-\eta\lambda)^{2t})$$

The proof is provided in the remainder of this section. We start with:

$$\bar{\mathbf{m}}_t = \eta^2\sigma^2\sum_{i=0}^{t-1}\mathbf{A}^i\lambda$$

Now note that since $\mathbf{A} = (\mathbf{I}-\eta\mathbf{\Lambda})^2 + \alpha\eta^2\lambda\lambda^\top \geqslant 0$, we have that $0 \leqslant \bar{\mathbf{m}}_t \leqslant \bar{\mathbf{m}}_{t+1}$ for all t. Taking $t \to \infty$ we have that:

$$\bar{\mathbf{m}}_\infty = \eta^2\sigma^2(\mathbf{I}-\mathbf{A})^{-1}\lambda \leqslant \frac{\eta\sigma^2}{1-\eta\alpha\mathrm{Tr}(\mathbf{H})}\mathbf{1}$$

Using this, we have that:

$$
\begin{aligned}
\bar{\mathbf{m}}_t &= \eta^2\sigma^2 \sum_{i=0}^{t-1} \mathbf{A}^i \lambda \\
&= \mathbf{A}\eta^2\sigma^2 \sum_{i=0}^{t-1} \mathbf{A}^{i-1} \lambda \\
&= \mathbf{A}\eta^2\sigma^2 \sum_{i=-1}^{t-2} \mathbf{A}^{i} \lambda \\
&= \mathbf{A}(\bar{\mathbf{m}}_{t-1} + \eta^2\sigma^2\mathbf{A}^{-1}\lambda) \\
&= \mathbf{A}\bar{\mathbf{m}}_{t-1} + \eta^2\sigma^2\lambda \\
&= \mathbf{D}\bar{\mathbf{m}}_{t-1} + \alpha\eta^2\lambda\lambda^\top\bar{\mathbf{m}}_{t-1} + \eta^2\sigma^2\lambda \\
&\leqslant \mathbf{D}\bar{\mathbf{m}}_{t-1} + \alpha\eta^2\mathrm{Tr}(\mathbf{H})\frac{\eta\sigma^2}{1-\eta\alpha\mathrm{Tr}(\mathbf{H})}\lambda + \eta^2\sigma^2\lambda \\
&= \mathbf{D}\bar{\mathbf{m}}_{t-1} + \eta^2\sigma^2\left(\frac{\alpha\mathrm{Tr}(\mathbf{H})\eta}{1-\eta\alpha\mathrm{Tr}(\mathbf{H})} + 1\right)\lambda \\
&= \mathbf{D}\bar{\mathbf{m}}_{t-1} + \frac{\eta^2\sigma^2}{1-\eta\alpha\mathrm{Tr}(\mathbf{H})}\lambda \\
&= \mathbf{D}^t\bar{\mathbf{m}}_0 + \frac{\eta^2\sigma^2}{1-\eta\alpha\mathrm{Tr}(\mathbf{H})}\sum_{i=0}^{t-1}\mathbf{D}^i\lambda
\end{aligned}
$$

Now unrolling this recursion and using the fact that $\bar{\mathbf{m}}_0 = 0$, we get:

$$
\begin{aligned}
\bar{\mathbf{m}}_t &\leqslant \mathbf{D}^t\bar{\mathbf{m}}_0 + \frac{\eta^2\sigma^2}{1-\eta\alpha\mathrm{Tr}(\mathbf{H})}\sum_{i=0}^{t-1}\mathbf{D}^i\lambda \\
&= \mathbf{D}^t\bar{\mathbf{m}}_0 + \frac{\eta^2\sigma^2}{1-\eta\alpha\mathrm{Tr}(\mathbf{H})}(\mathbf{I}-\mathbf{D}^t)(\mathbf{I}-\mathbf{D})^{-1}\lambda \\
&\leqslant \mathbf{D}^t\bar{\mathbf{m}}_0 + \frac{\eta\sigma^2}{1-\eta\alpha\mathrm{Tr}(\mathbf{H})}(\mathbf{1}-(\mathbf{1}-\eta\lambda)^{2t}) \\
&= \frac{\eta\sigma^2}{1-\eta\alpha\mathrm{Tr}(\mathbf{H})}(\mathbf{1}-(\mathbf{1}-\eta\lambda)^{2t})
\end{aligned}
$$

Appendix G. Lower bounds

While we do not cover lower bounds in this note, the work by Zou et al. [7] provides a lower bound that is sharper than the current upper bound. For sake of completeness, we will first restate (without proof) the lower bound from Zou et al. [7] in our mathematical framework:

Theorem 12 (Theorem 5.2 from Zou et al. [7]) *For SGD with tail averaging and thresholds $k^\star = \max\{j : \lambda_j \geqslant \frac{1}{\eta N}\}$ and $k^\dagger = \max\{j : \lambda_j \geqslant \frac{1}{\eta(s+N)}\}$, the bias and variance are lower bounded by (under mild as-*

sumptions):

$$
\begin{aligned}
\textit{bias} &\gtrsim \frac{1}{\eta^2 N^2} \|(\mathbf{I} - \eta \mathbf{\Lambda})^s (\mathbf{w}_0 - \mathbf{w}^\star)\|^2_{\mathbf{\Lambda}^{-1}_{0:k^\star}} + \|(\mathbf{I} - \eta \mathbf{\Lambda})^s (\mathbf{w}_0 - \mathbf{w}^\star)\|^2_{\mathbf{\Lambda}_{k^\star:\infty}} \\
&\quad + \|\mathbf{w}_0 - \mathbf{w}^\star\|^2_{\mathbf{\Lambda}_{k^\dagger:\infty}} \left(\frac{k^\star}{N} + N\eta^2 \sum_{j > k^\star} \lambda_j^2 \right) \\
\textit{variance} &\gtrsim \sigma^2 \left(\frac{k^\star}{N} + \eta \sum_{k^\star < j \leqslant k^\dagger} \lambda_j + (s + N)\eta^2 \sum_{j > k^\dagger} \lambda_j^2 \right)
\end{aligned}
$$

Note that the theorem is stated in a simplified way, ignoring the constants and focusing only on the leading order terms. The extra term $\|\mathbf{w}_0 - \mathbf{w}^\star\|^2_{\mathbf{I}_{0:k^\dagger}}$ that appears in the upper bound is not present in the lower bound, thus showing that the upper bound is not sharp.