Proceedings of Machine Learning Research vol 75:1–21, 2018 31st Annual Conference on Learning Theory

Iterate Averaging as Regularization for Stochastic Gradient Descent

Gergely Neu GERGELY.NEU@GMAIL.COM
Universitat Pompeu Fabra, Barcelona, Spain

Lorenzo Rosasco LROSASCO@MIT.EDU
LCSL, Massachusetts Institute of Technology, Cambridge MA, USA
Istituto Italiano di Tecnologia, Genova, Italy
Università degli Studi di Genova, Genova, Italy

Editors: Sébastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We propose and analyze a variant of the classic Polyak–Ruppert averaging scheme, broadly used in stochastic gradient methods. Rather than a uniform average of the iterates, we consider a weighted average, with weights decaying in a geometric fashion. In the context of linear least-squares regression, we show that this averaging scheme has the same regularizing effect, and indeed is asymptotically equivalent, to ridge regression. In particular, we derive finite-sample bounds for the proposed approach that match the best known results for regularized stochastic gradient methods.

Keywords: stochastic gradient descent, least squares, regularization, Tikhonov regularization

1. Introduction

Stochastic gradient methods are ubiquitous in machine learning, where they are typically referred to as SGD (stochastic gradient descent[1]). Since these incremental methods use little computation per data point, they are naturally adapted to processing very large data sets or streams of data. Stochastic gradient methods have a long history, starting from the pioneering paper by Robbins and Monro (Robbins and Monro, 1951). For a more detailed discussion, we refer to the excellent review given by Nemirovski et al. (2009). In the present paper, we propose a variant of SGD based on a weighted average of the iterates. The idea of averaging iterates goes back to Polyak (1990) and Ruppert (1988), and it is often referred to as Polyak–Ruppert averaging (see also Polyak and Juditsky, 1992). In this paper, we study SGD in the context of the linear least-squares regression problem, considering both finite- and infinite-dimensional settings. This latter case allows to derive results for nonparametric learning with kernel methods—we refer to the appendix in Rosasco and Villa (2015) for a detailed discussion on this subject. The study of SGD for least squares is classical in stochastic approximation (Kushner and Yin, 2003), where it is commonly known as the *least-mean-squares* (LMS) algorithm.

In the context of machine learning theory, prediction-error guarantees for online algorithms can be derived through a regret analysis in a sequential-prediction setting and a so called online-to-batch conversion (Shalev-Shwartz, 2012; Hazan, 2016). This technique has been particularly successful for optimization of strongly convex (and potentially non-smooth) loss functions (Hazan and Kale,

1. Albeit in general they might not be descent methods.

2014; Rakhlin et al., 2012; Shamir and Zhang, 2013), where it has lead to essentially optimal performance guarantees. While the same approach can be used to obtain fairly strong results for the specific problem of least squares through converting the bounds of Vovk (2001); Azoury and Warmuth (2001) and Hazan et al. (2007), the strongest known results are achieved by a direct analysis of SGD for this particular setting. This path was taken by Smale and Yao (2006); Ying and Pontil (2008); Tarres and Yao (2014), where the last iterate and decaying step-size are considered, and more recently by Rosasco and Villa (2015); Lin and Rosasco (2017) where multiple passes and mini-batching are considered. A recently popular approach is combining *constant* step-sizes with Polyak–Ruppert averaging, which was first shown to lead to minimax optimal finite-time prediction guarantees after a single pass on the data by Bach and Moulines (2013). This approach was first studied by Györfi and Walk (1996) and subsequent progress was made by Défossez and Bach (2015); Dieuleveut and Bach (2016); Dieuleveut et al. (2017); Jain et al. (2016, 2017); Lakshminarayanan and Szepesvári (2018).

In this paper we propose and analyze a novel form of weighted averaging, given by a sequence of weights decaying geometrically, so that the first iterates have more weight. Our main technical contribution is a characterization of the properties of this particular weighting scheme that we call *geometric Polyak–Ruppert averaging*. Our first result shows that SGD with geometric Polyak–Ruppert averaging is in expectation equivalent to considering SGD with a regularized loss function, and both sequences converge to the Tikhonov-regularized solution of the expected least-squares problem. The regularization parameter is a tuning parameter defining the sequence of geometric weights. This result strongly suggests that geometric Polyak–Ruppert averaging can be used to control the bias-variance properties of the corresponding SGD estimator. Indeed, our main result quantifies this intuition deriving a finite-sample bound, matching previous results for regularized SGD, and leading to optimal rates (Tsybakov, 2008). While averaging is widely considered to have a stabilizing effect, to the best of our knowledge this is the first result characterizing the stability of an averaging scheme in terms of its regularization properties and corresponding prediction guarantees. Our findings can be contrasted to recent results on tail averaging (Jain et al., 2016) and provide some guidance on when and how different averaging strategies can be useful. On a high level, our results suggest that geometric averaging should be used when the data is poorly-conditioned and/or relatively small, and tail averaging should be used in the opposite case. Further, from a practical point of view, geometric Polyak–Ruppert averaging provides an efficient approach to perform model selection, since *a regularization path* (Friedman et al., 2001) is computed efficiently. Indeed, it is possible to compute a full pass of SGD *once* and store all the iterates, to then rapidly compute offline the solutions corresponding to different geometric weights (or tail averages), hence different regularization levels. As the averaging operation is entirely non-serial, this method lends itself to trivially easy parallelization.

The rest of the paper is organized as follows. In Section 2, we introduce the necessary background and present the geometric Polyak–Ruppert averaging scheme. In Section 3, we show the asymptotic equivalence between ridge regression and constant-stepsize SGD with geometric iterate averaging. Section 4 presents and discusses our main results regarding the finite-time prediction error of the method. Section 5 describes the main steps in the proofs. The full proof is included in the Appendix. We conclude this section by introducing some basic notation used throughout the paper.

Notation. Let $\mathcal{H}$ be a separable Hilbert space with the inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. We let $v \otimes v$ be the outer product of vector $v \in \mathcal{H}$, which acts on $u \in \mathcal{H}$ as the rank-one operator $(v \otimes v)\, u = \langle u, v \rangle\, v$. For a linear operator A acting on $\mathcal{H}$, we let A^* be its adjoint and $\|A\| = \sqrt{\operatorname{tr}[A^* A]}$ the Frobenius norm. An operator A is positive semi-definite (PSD) if it satisfies $\langle x, Ax \rangle \geq 0$ for all $x \in \mathcal{H}$ and Hermitian if $A = A^*$. We use $A \succcurlyeq 0$ to denote that an operator A is Hermitian and PSD (in short, HPSD). For HPSD operators A and B, we use $A \succcurlyeq B$ to denote $A - B \succcurlyeq 0$. For a HPSD operator A, we use $A^{1/2}$ to denote the unique HPSD operator satisfying $A^{1/2} A^{1/2} = A$. The identity operator on $\mathcal{H}$ is denoted as I. Besides standard asymptotic notation like $\mathcal{O}(\cdot)$ or $o(\cdot)$, we will sometimes use the cleaner but less standard notation $a \lesssim b$ to denote $a = \mathcal{O}(b)$. We will consider algorithms that interact with stochastic data in a sequential fashion. The sequence of random variables observed during the interaction induce a filtration $(\mathcal{F}_t)_{t \geq 0}$. We will use the shorthand $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_t]$ to denote expectations conditional on the history.

2. Preliminaries

We study the problem of linear regression under the square loss, more commonly known as linear least-squares regression. The objective in this problem is to minimize the expected risk

$$R(w) = \frac{1}{2}\mathbb{E}\left[(\langle w, x \rangle - y)^2\right], \tag{1}$$

where $w \in \mathcal{H}$ is a parameter vector, $x \in \mathcal{H}$ is a covariate and $y \in \mathbb{R}$ is a label, with (x, y) drawn from a fixed (but unknown) distribution $\mathcal{D}$. Letting $\Sigma = \mathbb{E}[x \otimes x]$ denote the covariance operator, the minimizer of the risk is given by

$$w^* = \Sigma^{-1}\mathbb{E}[xy] \tag{2}$$

and satisfies $R(w^*) = \inf_{w \in \mathcal{H}} R(w)$. Without loss of generality, we assume Σ is positive definite, as this can be always satisfied by restricting $\mathcal{H}$ to the subspace in which all the covariates x_n lie almost surely. We also assume w^* to exist, even though in general this might not be true when $\mathcal{H}$ is infinite dimensional.

We study algorithms that take as input a set of data points $\{(x_t, y_t)\}_{t=1}^n$ drawn identically and independently from $\mathcal{D}$ and output a weight vector w to approximately minimize (1). The quality of an approximate solution is measured by the the excess risk

$$\Delta(w) = R(w) - R(w_*) = \left\|\Sigma^{1/2}(w - w^*)\right\|^2.$$

To compute a solution from data, we consider the stochastic gradient method, a.k.a. SGD, that for least squares takes the form

$$w_t = w_{t-1} - \eta_t \left(x_t \langle x_t, w_{t-1} \rangle - x_t y_t\right), \tag{3}$$

where $(\eta_t)_t > 0$ is a sequence of stepsizes (or learning rates), and $w_0 \in \mathcal{H}$ is an initial point. Typically, a decaying stepsize sequence is chosen to ensure convergence of the final iterate w_n, see Nemirovski et al. (2009) and references therein. A well-studied alternative approach is using a *constant* stepsize η and studying the properties of the average of the iterates

$$\overline{w}_n = \frac{1}{n+1}\sum_{t=0}^{n} w_t,$$

as first proposed by Polyak (1990) and Ruppert (1988). We will duly refer to this postprocessing step as *Polyak–Ruppert (PR) averaging*. As shown by Bach and Moulines (2013), this approach leads to minimax optimal performance guarantees for the finite dimensional setting without requiring significant prior knowledge about Σ (see also Györfi and Walk, 1996). This result was later strengthened in various respects by Défossez and Bach (2015) and Dieuleveut, Flammarion, and Bach (2017), notably by weakening the assumptions in Bach and Moulines (2013) and separating error terms related to "bias" and "variance". We highlight one result from Dieuleveut et al. (2017) which considers optimizing the objective (1) plus an additive regularization term of $\frac{1}{2}\left\|w - w_0\right\|^2$ by iteratively computing the updates

$$w_t = w_{t-1} - \eta\left(\Sigma w_{t-1} - x_t y_t + \lambda\left(w_{t-1} - w_0\right)\right). \tag{4}$$

Under technical assumptions discussed later, Dieuleveut et al. (2017) prove the excess-risk bound

$$\mathbb{E}\left[\Delta(\overline{w}_n)\right] \lesssim \frac{\sigma^2 \mathrm{tr}\left[\Sigma^2\left(\Sigma + \lambda I\right)^{-2}\right]}{n} + \left(\lambda + \frac{1}{\eta n}\right)^2 \left\|\Sigma^{1/2}\left(\Sigma + \lambda I\right)^{-1}\left(w_0 - w^*\right)\right\|^2, \tag{5}$$

where $\sigma^2 > 0$ is an upper bound on the variance of the label noise. The iteration (4) is only of theoretical interest since the covariance Σ is not known in practice. However, the obtained bound is simpler to present allows easier comparison with our result. A bound slightly more complex than (5) can be obtained when Σ is not known (Dieuleveut et al., 2017, Theorem 2).

In this paper, we propose a generalized version of Polyak–Ruppert averaging that we call *geometric Polyak–Ruppert averaging*. Specifically, the algorithm we study computes the standard SGD iterates with some constant stepsize η as given by Equation (3) and outputs

$$\widetilde{w}_n = \frac{1}{\sum_{k=0}^{n}\left(1 - \gamma\lambda\right)^k} \cdot \sum_{t=0}^{n}\left(1 - \gamma\lambda\right)^t w_t \tag{6}$$

after round n, where $\lambda \in [0, 1/\gamma)$ is a tuning parameter and γ satisfies $\eta = \frac{\gamma}{(1-\gamma\lambda)}$. That is, the output is a geometrically discounted (and appropriately normalized) average of the plain SGD iterates that puts a larger weight on initial iterates. It is easy to see that setting $\lambda = 0$ exactly recovers the standard form of Polyak–Ruppert averaging. Our main result essentially shows that the resulting estimate $\widetilde{w}_n$ satisfies[2]

$$\mathbb{E}\left[\Delta(\widetilde{w}_n)\right] \lesssim \left(\frac{\gamma\lambda}{2} + \frac{1}{n}\right)\sigma^2 \mathrm{tr}\left[\Sigma^2\left(\Sigma + \lambda I\right)^{-2}\right] + \left(\lambda + \frac{1}{\gamma n}\right)^2 \left\|\Sigma^{1/2}\left(\Sigma + \lambda I\right)^{-1}\left(w_0 - w^*\right)\right\|^2$$

under the same assumptions as the ones made by Dieuleveut et al. (2017). Notably, this guarantee matches the bound of Equation (5), with the key difference being that the factor $\frac{1}{n}$ in the first term is replaced by $\frac{1}{n} + \frac{\gamma\lambda}{2}$. This observation suggests the (perhaps surprising) conclusion that geometric Polyak–Ruppert averaging has a regularization effect qualitatively similar to Tikhonov regularization. Before providing the proof of the main result stated above, we first show that this similarity is more than a coincidence. Specifically, we begin by showing in Section 3 that the limit of the weighted iterates is *exactly* the ridge regression solution on expectation.

2. The bound shown here concerns an iteration similar to the one shown on Equation (4), and is proved in Appendix C. We refer to Theorem 3 for the precise statement of our main result.

3. Geometric iterate averaging realizes Tikhonov regularization on expectation

We begin by studying the relation between the averaged iterates of unregularized SGD and the iterates of regularized SGD on expectation. This setting will allow us to make minimal assumptions: We merely assume that $\mathbb{E}\|x\|^2 < \infty$ so that the covariance operator Σ satisfies $\Sigma \preccurlyeq B^2 I$ for some $B > 0$. For ease of exposition, we assume in this section that $w_0 = 0$. First, we notice the relation

$$\mathbb{E}_t [w_t] = w_{t-1} - \eta \left(\mathbb{E}_t [x_t \otimes x_t] w_{t-1} - \mathbb{E}_t [x_t y_t]\right) = (I - \eta\Sigma) w_{t-1} + \eta \mathbb{E}[xy]$$

between w_t and w_{t-1}, which can be iteratively applied to obtain

$$\mathbb{E}[w_t] = \eta \sum_{k=1}^{t} (I - \eta\Sigma)^{k-1} \mathbb{E}[xy] .$$

In contrast we also define the iterates of regularized SGD with stepsize $\gamma > 0$ and regularization parameter λ as

$$\widehat{w}_t = \widehat{w}_{t-1} - \gamma \left(x_t \langle x_t, \widehat{w}_{t-1}\rangle - x_t y_t + \lambda \widehat{w}_{t-1}\right), \tag{7}$$

which can be similarly shown to satisfy

$$\mathbb{E}[\widehat{w}_t] = \gamma \sum_{k=1}^{t} (I - \gamma\Sigma - \gamma\lambda I)^{k-1} \mathbb{E}[xy] .$$

This latter definition can be seen as an empirical version of the iteration in Eq. (4) with $w_0 = 0$.

The following proposition reveals a profound connection between the limits of $\mathbb{E}[\widehat{w}_t]$ and the geometrically discounted average $\mathbb{E}[\widetilde{w}_t]$ as $t \to \infty$, given that the stepsizes are carefully chosen and small enough for the limits to exist.

Proposition 1 *Let η, γ and λ be such that $\gamma\lambda < 1$ and $\eta = \frac{\gamma}{1-\gamma\lambda}$, and assume that $\gamma \leq \frac{1}{B^2}$. Then, $\widehat{w}_\infty = \lim_{t\to\infty} \mathbb{E}[\widehat{w}_t]$ and $\widetilde{w}_\infty = \lim_{t\to\infty} \mathbb{E}[\widetilde{w}_t]$ both exist and satisfy*

$$\widehat{w}_\infty = \widetilde{w}_\infty = (\Sigma + \lambda I)^{-1} \mathbb{E}[xy] .$$

Proof The analysis crucially relies on defining the geometric random variable G with law $\mathbb{P}[G = k] = \gamma\lambda (1 - \gamma\lambda)^{k-1}$ for all $k = 1, 2, \ldots$ and noticing that $\mathbb{P}[G \geq k] = (1 - \gamma\lambda)^{k-1}$. We let $\mathbb{E}_G[\cdot]$ stand for taking expectations with respect to the distribution of G. First we notice that the limit of $\mathbb{E}[w_t]$ can be written as

$$\lim_{t\to\infty} \mathbb{E}[\widehat{w}_t] = \gamma \sum_{k=1}^{\infty} (I - \gamma\Sigma - \gamma\lambda I)^{k-1} \mathbb{E}[xy] .$$

By our assumption on γ, we have $\gamma\Sigma \preccurlyeq I$, which implies that the series on the right-hand side converges and satisfies $\sum_{k=1}^{\infty} (I - \gamma\Sigma - \gamma\lambda I)^{k-1} = (\gamma\Sigma + \gamma\lambda I)^{-1}$. Having established the existence of the limit, we rewrite the regularized SGD iterates (7) as

$$\begin{aligned}\lim_{t\to\infty} \mathbb{E}[\widehat{w}_t] &= \gamma \sum_{k=1}^{\infty} (I - \gamma\Sigma - \gamma\lambda I)^{k-1} \mathbb{E}[xy] = \gamma \sum_{k=1}^{\infty} (I - \eta\Sigma)^{k-1} (1 - \gamma\lambda)^{k-1} \mathbb{E}[xy] \\ &= \gamma \sum_{k=1}^{\infty} (I - \eta\Sigma)^{k-1} \mathbb{P}[G \geq k] \mathbb{E}[xy] = \gamma \mathbb{E}_G \left[\sum_{k=1}^{\infty} (I - \eta\Sigma)^{k-1} \mathbb{I}_{\{k \leq G\}} \mathbb{E}[xy]\right]\end{aligned}$$

$$
\begin{aligned}
&= \gamma \mathbb{E}_G \left[\sum_{k=1}^{G} (I - \eta\Sigma)^{k-1} \mathbb{E}\left[xy\right] \right] = \gamma \mathbb{E}_G \left[\frac{1}{\eta} \mathbb{E}\left[w_G\right] \right] = \frac{\gamma}{\eta} \sum_{k=1}^{\infty} \mathbb{P}\left[G = k\right] \mathbb{E}\left[w_k\right] \\
&= \frac{\gamma}{\eta} \sum_{k=1}^{\infty} \gamma\lambda(1-\gamma\lambda)^{k-1} \mathbb{E}\left[w_k\right] = \frac{\gamma}{\eta(1-\gamma\lambda)} \sum_{k=0}^{\infty} \gamma\lambda(1-\gamma\lambda)^{k} \mathbb{E}\left[w_k\right] \\
&= \lim_{t\to\infty} \sum_{k=0}^{t} \gamma\lambda(1-\gamma\lambda)^{k} \mathbb{E}\left[w_k\right] = \lim_{t\to\infty} \sum_{k=0}^{t} \gamma\lambda(1-\gamma\lambda)^{k} \mathbb{E}\left[\widetilde{w}_t\right] = \lim_{t\to\infty} \mathbb{E}\left[\widetilde{w}_t\right].
\end{aligned}
$$

This concludes the proof. ∎

It is useful to recall that $(\Sigma + \lambda I)^{-1}\mathbb{E}\left[xy\right]$ is the solution of the problem

$$
\min_{w\in\mathcal{H}} \frac{1}{2}\mathbb{E}\left[(y - \langle w, x\rangle)^2\right] + \frac{\lambda}{2}\left\|w\right\|^2,
$$

that is Tikhonov regularization applied to the expected risk or, in other words, the population version of the ridge regression estimator. Then, the above result shows that SGD with geometric averaging (6) and the regularized iteration (7) both converge to this same solution. Besides this result, we can also show that the expected iterates $\mathbb{E}\left[\widehat{w}_t\right]$, $\mathbb{E}\left[w_t\right]$ and $\mathbb{E}\left[\widetilde{w}_t\right]$ are also closely connected for finite t, without any assumption on the learning rates.

Proposition 2 *Let η, γ and λ be such that $\gamma\lambda < 1$ and $\eta = \frac{\gamma}{1-\gamma\lambda}$. Then,*

$$
\mathbb{E}\left[\widehat{w}_t\right] = \left(1 - (1-\gamma\lambda)^t\right)\mathbb{E}\left[\widetilde{w}_t\right] + (1-\gamma\lambda)^t\mathbb{E}\left[w_t\right].
$$

This proposition is proved using the same ideas as Proposition 1; we include the proof in Appendix A for completeness.

4. Main result: Finite-time performance guarantees

While the previous section establishes a strong connection between the geometrically weighted SGD iterates with the iterates of regularized SGD on expectation, this connection is clearly not enough for the known performance guarantees to carry over to our algorithm. Specifically, the two iterative schemes propagate noise differently, thus the covariance of the resulting iterate sequences may be very different from each other. In this section, we prove our main result that shows that the prediction error of our algorithm also behaves similarly to that of SGD with Tikhonov regularization. For our analysis, we will make the same assumptions as Dieuleveut et al. (2017) and we will borrow several ideas from them, as well as most of their notation.

We now state the assumptions that we require for proving our main result. We first state an assumption on the fourth moment of the covariates.

Assumption 1 *There exists $R > 0$ such that*

$$
\mathbb{E}\left[\left\|x\right\|^2 x \otimes x\right] \preccurlyeq R^2\Sigma.
$$

This assumption implies that $\mathrm{tr}\left[\Sigma\right] \leq R^2$, and is satisfied, for example, when the covariates satisfy $\|x\| \leq R$ almost surely. We always assume a minimizer w_* of the expected risk to exist and also make an assumption on the residual ε defined as the random variable

$$\varepsilon = y - \langle w^*, x \rangle .$$

It is easy to show that $\mathbb{E}\left[\varepsilon x\right] = 0$, even though $\mathbb{E}\left[\varepsilon | \, x\right] = 0$ does not hold in general. We make the following assumption on the residual:

Assumption 2 *There exists $\sigma^2 > 0$ such that*

$$\mathbb{E}\left[\|\varepsilon\|^2 x \otimes x\right] \preccurlyeq \sigma^2 \Sigma.$$

This assumption is satisfied when $\|x\|$ and y are almost surely bounded, or when the model is well-specified and corrupted with bounded noise (i.e., when ε is independent of x and has variance bounded by σ^2). Under the above assumptions, we prove the following bound on the prediction error—our main result:

Theorem 3 *Suppose that Assumptions 1 and 2 hold and assume that $\eta \leq \frac{1}{2R^2}$ and $\lambda \in [0, 1/\eta)$. Then, the iterates computed by the recursion given by Equations (3) and (6) satisfy*

$$\begin{aligned}\mathbb{E}\left[\Delta(\widetilde{w}_n)\right] \leq & \frac{4}{1-\gamma\lambda}\left(\frac{\gamma\lambda}{(2-\gamma\lambda)} + \frac{2}{(2-\gamma\lambda)(n+1)}\right)\frac{\sigma^2 \mathrm{tr}\left[\Sigma^2 \left(\Sigma + \lambda I\right)^{-2}\right]}{2-\eta R^2} \\ & + 2\left(\lambda + \frac{1}{\gamma(n+1)}\right)^2 \left\|\Sigma^{1/2}\left(\Sigma + \tfrac{\lambda}{2} I\right)^{-1}(w_0 - w^*)\right\|^2 \\ & + \left(\lambda + \frac{1}{\gamma(n+1)}\right)^2 \mathrm{tr}\left[\Sigma\left(\Sigma + \lambda I\right)^{-1}\right]\left\|\left(\Sigma + \tfrac{\lambda}{2} I\right)^{-1/2}(w_0 - w^*)\right\|^2\end{aligned}$$

The (rather technical) proof of the theorem closely follows the proof of Theorem 2 of Dieuleveut et al. (2017). We describe the main components of the proof of our main result in Section 5. For didactic purposes, we also present a simplified version of our analysis where we assume full knowledge of Σ in Appendix C.

4.1. Discussion

We next discuss various aspects and implications of our results.

Comparison with Dieuleveut et al. (2017). Apart from constant factors[3], our bound above precisely matches that of Theorem 1 of Dieuleveut et al. (2017), except for an additional term of order $\gamma\lambda\sigma^2 \mathrm{tr}\left[\Sigma^2\left(\Sigma + \lambda I\right)^{-2}\right]$. This term, however, is *not* an artifact of our proof: in fact it captures a distinctive noise-propagating effect of our geometric PR averaging scheme. Indeed, the regularization effect of our iterate averaging scheme is different from that of Tikhonov-regularized SGD in one significant way: while Tikhonov regularization increases the bias and strictly decreases the variance, our scheme may actually increase the variance for certain choices of λ. To see this, observe that setting a large λ puts a large weight on the initial iterates, so that the initial noise is amplified

3. By enforcing $\gamma\lambda \leq 1/2$, the $1 - \gamma\lambda$ and $2 - \gamma\lambda$ terms in the denominator can be lower bounded by a constant.

compared to noise in the later stages, and the concentration of the total noise becomes worse. We note however that this extra term does not qualify as a serious limitation, since the commonly recommended setting $\lambda = \mathcal{O}\big(\frac{1}{\eta n}\big)$ still preserves the optimal rates for both the bias and the variance up to constant factors.

Optimal excess risk bounds. The bound in Theorem 3 is essentially the same as the one derived in Dieuleveut et al. (2017, Theorem 2). Following their same reasoning, the bound can be optimized with respect to λ, γ to derive the best parameters choice and explicit upper bounds on the corresponding excess risk. In the finite dimensional case, it is easy to derive a bound of order $\mathcal{O}(d/n)$, which is known to be optimal in a minimax sense (Tsybakov, 2008). In the infinite dimensional case, optimal minimax bounds can again be easily derived, and also refined under further assumptions on w_* and the covariance Σ (De Vito et al., 2005; Caponnetto and De Vito, 2007). We omit this derivation.

When should we set $\lambda > 0$? We have two answers depending on the dimensionality of the underlying Hilbert space $\mathcal{H}$. For infinite dimensional spaces, it is clearly necessary to set $\lambda > 0$. In the finite-dimensional case, the advantage of our regularization scheme is less clear at first sight: while Tikhonov regularization strictly decreases the variance, this is not necessarily true for our scheme (as discussed above). A closer look reveals that, under some (rather interpretable) conditions, we can reduce the variance, as quantified by the following proposition.

Proposition 4 *If* tr $\left[\Sigma^{-1}\right] > \frac{1}{2}\gamma d n$ *there exists a regularization parameter $\lambda^* > 0$ satisfying*

$$\left(\frac{\gamma\lambda^*}{2} + \frac{1}{n}\right) \operatorname{tr}\left[\Sigma^2 \left(\Sigma + \lambda^* I\right)^{-2}\right] < \frac{d}{n}.$$

Proof Letting $s_1, s_2, \ldots, s_d$ be the eigenvalues of Σ sorted in decreasing order, we have

$$f(\lambda) = \left(\frac{\gamma\lambda^*}{2} + \frac{1}{n}\right) \operatorname{tr}\left[\Sigma^2 \left(\Sigma + \lambda^* I\right)^{-2}\right] = \left(\frac{\gamma\lambda^*}{2} + \frac{1}{n}\right) \sum_{i=1}^{d} \frac{s_i^2}{(s_i + \lambda)^2}.$$

Taking derivative of f with respect to λ gives

$$f'(\lambda) = \frac{\gamma}{2} \sum_{i=1}^{d} \frac{s_i^2}{(s_i + \lambda)^2} - 2\left(\frac{\gamma\lambda^*}{2} + \frac{1}{n}\right) \sum_{i=1}^{d} \frac{s_i^2}{(s_i + \lambda)^3}.$$

In particular, we have

$$f'(0) = \frac{\gamma d}{2} - \frac{2}{n} \sum_{i=1}^{d} \frac{1}{s_i} = \frac{\gamma d}{2} - \frac{2 \operatorname{tr}\left[\Sigma^{-1}\right]}{n},$$

so $f'(0) < 0$ holds whenever tr $\left[\Sigma^{-1}\right] > \frac{1}{4}\gamma d n$. The proof is concluded by observing that $f'(0) < 0$ implies the existence of a λ^* with the claimed property. ∎

Intuitively, Proposition 4 suggests that the geometric PR averaging can definitely reduce the variance over standard PR averaging whenever the covariance matrix is poorly conditioned and/or the sample size is small. Notice however that the above argument only shows one example of a good choice of λ; many other good choices may exist, but these are harder to characterize.

What is the computational advantage? The main practical advantage of our averaging scheme over Tikhonov regularization is a computational one: validating regularization parameters becomes trivially easy to parallelize. Indeed, one can perform a *single* pass of unregularized SGD over the data, store the iterates and average them with various schedules to evaluate different choices of λ. Through parallelization, this approach can achieve huge computational speedups over running regularized SGD from scratch. To see this, observe that the averaging operation is *entirely non-serial*: one can cut the (stored) SGD iterates into K contiguous batches and let each individual worker perform a geometric averaging with the same discount factor $(1-\gamma\lambda)$. The resulting averages are then combined by the master with appropriate weights. In contrast, regularized SGD is entirely serial, so validation cannot be parallelized.

Geometric averaging vs. tail averaging. It is interesting to contrast our approach with the *tail averaging* scheme studied by Jain et al. (2016, 2017): instead of putting large weight on the initial iterates as our method does, Jain et al. suggest to average the *last* $n-\tau$ iterates of SGD for some τ. The effect of this operation is that the $\left\|\Sigma^{-1/2}\left(w^*-w_0\right)\right\|^2 n^{-2}$ term arising from Polyak–Ruppert averaging is replaced by a term of order $\exp(-\eta\mu\tau)\left\|w^*-w_0\right\|^2(n-\tau)^{-2}$, where $\mu>0$ is the smallest eigenvalue of Σ. Clearly, this yields a significant asymptotic speedup, but gives no advantage when $\gamma n \leq \mu^{-1}$ (noting that $\tau<n$). Contrasting this condition with our Proposition 4 leads to an interesting conclusion: for small values of n, geometric averaging has an edge over tail averaging and vice versa. Since the above method for post-processing the iterates can be also used for tuning τ, we conclude that choosing the right averaging can be done in a simple and highly parallelizable fashion.

Connections to early stopping. A close inspection of the proofs of our Propositions 1 and 2 reveals an interesting perspective on geometric iterate averaging: Thinking of the averaging operation as computing a probabilistic expectation, one can interpret the geometric average as a *probabilistic early stopping* method where the stopping time is geometrically distributed. Early stopping is a very well-studied regularization method for multipass stochastic gradient learning algorithms that has been observed and formally proved to have effects similar to Tikhonov regularization (Yao et al., 2007; Rosasco and Villa, 2015). So far, all[4] published results that we are aware of merely point out the qualitative similarities between the performance bounds obtained for these two regularization methods, showing that using the stopping time should be chosen as $t^*=1/\gamma\lambda$. In contrast, our Propositions 1 and 2 show a much deeper connection: geometric random stopping with expected stopping time $\mathbb{E}[G]=1/\gamma\lambda$ not only recovers the performance bounds, but *exactly* recovers the ridge-regression solution.

Open questions. It is natural to ask whether geometric iterate averaging has similar regularization effects in other stochastic approximation settings too. A possible direction for future work is studying the effects of our averaging scheme on accelerated and/or variance-reduced variants of SGD. Another promising direction is studying general linear stochastic approximation schemes (Lakshminarayanan and Szepesvári, 2018), and particularly Temporal-Difference learning algorithms for Reinforcement Learning that have so far resisted all attempts to regularize them (Sutton and Barto, 1998; Szepesvári, 2010; Farahmand, 2011).

4. With the notable exception of Fleming (1990), who shows an exact relation between the ridge-regression solution and a rather complicated early stopping rule involving a preconditioning step that requires computing the eigendecomposition of Σ.

5. The proof of Theorem 3

Our proof closely follows that of Dieuleveut et al. (2017, Theorem 1), with the key differences that

- we do not have to deal with an explicit "regularization-based" error term that gives rise to a term proportional to λ^2 in their bound, and
- the $\frac{1}{n}$ factors for iterate averaging are replaced by $c_n(1-\gamma\lambda)^t$ for each round, where

$$c_n = \frac{1}{\sum_{t=0}^{n}(1-\gamma\lambda)^t} = \frac{\gamma\lambda}{1-(1-\gamma\lambda)^n}.$$

As we will see, this change will propagate through the analysis and will eventually replace the $\frac{1}{\gamma^2 n}$ and $\left(2\lambda + \frac{1}{\eta n}\right)$ factors in the final bound by $c_n^2 \sum_{t=0}^{n}(1-\gamma\lambda)^{2t}$ and $\frac{c_n^2}{\gamma^2}$, respectively.

In the interest of space, we only provide an outline of the proof here and defer the proofs of the key lemmas to Appendix B. Throughout the proof, we will suppose that the conditions of Theorem 3 hold. The lemma below shows that the factors involving c_n^2 are of the order claimed in the Theorem.

Lemma 5 *For any $n \geq 1$, we have*

$$c_n^2 \leq \left(\gamma\lambda + \frac{1}{n+1}\right)^2$$

and

$$c_n^2 \sum_{t=0}^{n}(1-\gamma\lambda)^{2t} \leq \frac{\gamma\lambda}{(2-\gamma\lambda)} + \frac{2}{(2-\gamma\lambda)(n+1)}.$$

The straightforward proof is given in Appendix B.1.

Now we are ready to lay out the proof of Theorem 3. Let us start by introducing the notation

$$M_{i,j} = \left(\prod_{k=i+1}^{j}(I - \eta x_k \otimes x_k)\right)$$

for all $i < j$ and $M_{i,i} = I$ for all i, and recalling the definition $\varepsilon_t = y_t - \langle x_t, w^* \rangle$. A simple recursive argument shows that

$$\begin{aligned} w_t - w^* &= w_{t-1} - \eta\left(\langle x_t, w_{t-1}\rangle - y_t\right)x_t - w^* \\ &= (I - \eta x_t \otimes x_t)(w_{t-1} - w^*) + \eta x_t y_t - \eta x_t \langle x_t, w^*\rangle \\ &= (I - \eta x_t \otimes x_t)(w_{t-1} - w^*) + \eta \varepsilon_t x_t \\ &= M_{t-1,t}(w_{t-1} - w^*) + \eta \varepsilon_t x_t \\ &= M_{0,t}(w_0 - w^*) + \eta \sum_{j=1}^{t} M_{j,t}\varepsilon_j x_j. \end{aligned} \tag{8}$$

Thus, the averaged iterates satisfy

$$\widetilde{w}_n - w^* = c_n \cdot \sum_{t=0}^{n}(1-\gamma\lambda)^t (w_t - w^*)$$

$$= c_n \cdot \sum_{t=0}^{n} (1-\gamma\lambda)^t \left(M_{0,t}\,(w_0 - w^*) + \eta \sum_{j=1}^{t} M_{j,t}\varepsilon_j x_j \right).$$

We first show a simple upper bound on the excess risk $\Delta(\widetilde{w}_n) = \left\|\Sigma^{1/2}\left(\widetilde{w}_t - w^*\right)\right\|^2$:

Lemma 6

$$\mathbb{E}\left[\Delta(\widetilde{w}_n)\right] \le \frac{2c_n^2}{\gamma} \sum_{t=0}^{n} (1-\gamma\lambda)^{2t} \mathbb{E}\left[\left\|\Sigma^{1/2}\left(\Sigma+\lambda I\right)^{-1/2}\left(w_t - w^*\right)\right\|^2\right]$$

The proof is included in Appendix B.2. In order to further upper bound the right-hand side in the bound stated in Lemma 6, we can combine the decomposition of $w_t - w^*$ in Equation (8) with the Minkowski inequality to get

$$\begin{aligned}
&\sum_{t=0}^{n}(1-\gamma\lambda)^{2t}\mathbb{E}\left[\left\|\Sigma^{1/2}\left(\Sigma+\lambda I\right)^{-1/2}\left(w_t - w^*\right)\right\|^2\right] \\
&\le 2\underbrace{\sum_{t=0}^{n}(1-\gamma\lambda)^{2t}\mathbb{E}\left[\left\|\Sigma^{1/2}\left(\Sigma+\lambda I\right)^{-1/2} M_{0,t}\left(w_0 - w^*\right)\right\|^2\right]}_{\Delta_1} \\
&\qquad\qquad + 2\eta^2 \sum_{t=0}^{n}(1-\gamma\lambda)^{2t}\underbrace{\mathbb{E}\left[\left\|\Sigma^{1/2}\left(\Sigma+\lambda I\right)^{-1/2}\sum_{j=1}^{t} M_{j,t}\varepsilon_j x_j\right\|^2\right]}_{\Delta_{2,t}}
\end{aligned} \tag{9}$$

The first term in the above decomposition can be thought of as the excess risk of a "noiseless" process (where $\sigma = 0$) and the second term as that of a "pure noise" process (where $w_0 = w^*$). The rest of the analysis is devoted to bounding these two terms.

We begin with the conceptually simpler case of bounding $\Delta_{2,t}$, which can be done uniformly for all t. In particular, we have the following lemma:

Lemma 7 *For any t, we have*

$$\Delta_{2,t} \le \frac{\eta\sigma^2}{2-\eta R^2}\operatorname{tr}\left[\Sigma^2\left(\Sigma+\lambda I\right)^{-2}\right].$$

The rather technical proof is presented in Appendix B.3. We now turn to bounding the excess risk of the "noiseless" process, Δ_1:

$$\Delta_1 = \sum_{t=0}^{n}(1-\gamma\lambda)^{2t}\cdot\mathbb{E}\left[\operatorname{tr}\left[M_{0,t}^{*}\Sigma\left(\Sigma+\lambda I\right)^{-1} M_{0,t}\left(w_0 - w^*\right)\otimes\left(w_0 - w^*\right)\right]\right].$$

The following lemma states a bound on Δ_1 in terms of $E_0 = (w_0 - w^*)\otimes(w_0 - w^*)$.

Lemma 8

$$\Delta_1 \le \frac{1}{2\gamma}\operatorname{tr}\left[\Sigma\left(\Sigma+\lambda I\right)^{-1}\left(\Sigma+\tfrac{\lambda}{2}I\right)^{-1}E_0\right] + \frac{1}{4\gamma}\operatorname{tr}\left[\Sigma\left(\Sigma+\lambda I\right)^{-1}\right]\operatorname{tr}\left[\left(\Sigma+\tfrac{\lambda}{2}I\right)^{-1}E_0\right]$$

The extremely technical proof of this theorem is presented in Appendix B.4.

The proof of Theorem 3 is concluded by plugging the bounds of Lemmas 7 and 8 into Equation (9) and using Lemma 6 to obtain

$$\mathbb{E}\left[\Delta(\widetilde{w}_n)\right] \leq \frac{c_n^2}{\gamma^2}\left(2\mathrm{tr}\left[\Sigma\left(\Sigma+\lambda I\right)^{-1}\left(\Sigma+\tfrac{\lambda}{2}I\right)^{-1}E_0\right]+\mathrm{tr}\left[\Sigma\left(\Sigma+\lambda I\right)^{-1}\right]\mathrm{tr}\left[\left(\Sigma+\tfrac{\lambda}{2}I\right)^{-1}E_0\right]\right)$$
$$+\frac{4\eta c_n^2}{\gamma}\sum_{t=0}^{n}(1-\gamma\lambda)^{2t}\frac{\sigma^2\mathrm{tr}\left[\Sigma^2\left(\Sigma+\lambda I\right)^{-2}\right]}{2-\eta R^2}$$

Now we can finish by using the bounds on c_n^2 and $c_n^2\sum_{t=0}^{n}(1-\gamma\lambda)^{2t}$ given in Lemma 5.

Acknowledgments

GN is supported by the UPFellows Fellowship (Marie Curie COFUND program n° 600387). LR is funded by the AFOSR project FA9550-17-1-0390 (European Office of Aerospace Research and Development) and by the MSCA-RISE, 734564. The authors thank the three anonymous reviewers for their valuable feedback on a preliminary version of the paper, and Francesco Orabona, Csaba Szepesvári and Gábor Lugosi for interesting discussions.

References

Katy S. Azoury and Manfred K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning Journal*, 43(3):211–246, 2001.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 773–781, 2013.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Ernesto De Vito, Andrea Caponnetto, and Lorenzo Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5(1):59–85, 2005.

Alexandre Défossez and Francis Bach. Constant step size least-mean-square: Bias-variance trade-offs and optimal sampling distributions. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 205–213, 2015.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

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.

Amir-massoud Farahmand. *Regularization in reinforcement learning*. PhD thesis, University of Alberta, 2011.

Henry E. Fleming. Equivalence of regularization and truncated iteration in the solution of ill-posed image reconstruction problems. *Linear Algebra and its applications*, 130:133–150, 1990.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

László Györfi and Harro Walk. On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization*, 34(1):31–61, 1996.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1): 2489–2512, 2014.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69:169–192, 2007.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic approximation through mini-batching and tail-averaging. *arXiv preprint arXiv:1610.03774*, 2016.

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.

Harold J. Kushner and G. George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Chandrashekar Lakshminarayanan and Csaba Szepesvári. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1347–1355, 2018.

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

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Boris Polyak. New stochastic approximation type procedures. *Autom. i Telemekh., 7:98107, 1990*, 7(7):98–107, 1990.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1571–1578, 2012.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems Proceedings 28 (NIPS)*, pages 1630–1638. MIT Press, 2015.

David Ruppert. Efficient estimations from a slowly convergent Robbins–Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 71–79, 2013.

Steve Smale and Yuan Yao. Online learning algorithms. *Foundations of computational mathematics*, 6(2):145–170, 2006.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*, volume 1. 1998.

Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

Pierre Tarres and Yuan Yao. Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9):5716–5735, 2014.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

Vladimir Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

Appendix A. The proof of Proposition 2

The proof is similar to that of Proposition 1, although a little more cluttered due to the normalization constants involved in the definition of $\widetilde{w}_t$. The crucial difference is in defining the truncated sequence of SGD iterates (w'_k), where $w'_k = w_k$ for $k \le t$ and $w'_k = 0$ for $k > t$. Defining again the geometric random variable G with law $\mathbb{P}[G = k] = \gamma\lambda(1-\gamma\lambda)^{k-1}$ and letting $\mathbb{E}_G[\cdot]$ stand for taking expectations with respect to the distribution of G, we rewrite the regularized SGD iterates as

$$\mathbb{E}[\widehat{w}_t] = \gamma \sum_{k=1}^{t} (I - \gamma\Sigma - \gamma\lambda I)^{k-1}\, \mathbb{E}[xy] = \gamma \sum_{k=1}^{t} (I - \eta\Sigma)^{k-1} (1-\gamma\lambda)^{k-1} \mathbb{E}[xy]$$

$$
\begin{aligned}
&= \gamma \sum_{k=1}^{t} (I - \eta\Sigma)^{k-1} \mathbb{P}[G \geq k] \mathbb{E}[xy] = \gamma \mathbb{E}_G \left[\sum_{k=1}^{t} (I - \eta\Sigma)^{k-1} \mathbb{I}_{\{k \leq G\}} \mathbb{E}[xy] \right] \\
&= \gamma \mathbb{E}_G \left[\sum_{k=1}^{G \wedge t} (I - \eta\Sigma)^{k-1} \mathbb{E}[xy] \right] = \gamma \mathbb{E}_G \left[\frac{1}{\eta} \mathbb{E}[w_{G \wedge t}] \right] \\
&= \frac{\gamma}{\eta} \sum_{k=1}^{t-1} \mathbb{P}[G = k] \mathbb{E}[w_k] + \frac{\gamma}{\eta} \mathbb{P}[G \geq t] \mathbb{E}[w_t] \\
&= \frac{\gamma}{\eta} \sum_{k=1}^{t-1} \gamma\lambda (1 - \gamma\lambda)^{k-1} \mathbb{E}[w_k] + \frac{\gamma}{\eta} (1 - \gamma\lambda)^{t-1} \mathbb{E}[w_t] \\
&\overset{(*)}{=} \frac{\gamma}{\eta(1 - \gamma\lambda)} \sum_{k=0}^{t-1} \gamma\lambda (1 - \gamma\lambda)^{k} \mathbb{E}[w_k] + \frac{\gamma}{\eta(1 - \gamma\lambda)} (1 - \gamma\lambda)^{t} \mathbb{E}[w_t] \\
&= \sum_{k=0}^{t-1} \gamma\lambda (1 - \gamma\lambda)^{k} \mathbb{E}[\widetilde{w}_t] + (1 - \gamma\lambda)^{t} \mathbb{E}[w_t] \\
&= \left(1 - (1 - \gamma\lambda)^{t}\right) \mathbb{E}[\widetilde{w}_t] + (1 - \gamma\lambda)^{t} \mathbb{E}[w_t],
\end{aligned}
$$

where we used $w_0 = 0$ in the step marked by $(*)$. This concludes the proof.

Appendix B. Tools for proving Theorem 3

B.1. The proof of Lemma 5

Regarding the first statement, we have

$$
c_n = \frac{1}{\sum_{t=0}^{n} (1 - \gamma\lambda)^t} = \frac{\gamma\lambda}{1 - (1 - \gamma\lambda)^{n+1}} \leq \frac{\gamma\lambda}{1 - e^{-\gamma\lambda(n+1)}} \leq \gamma\lambda \left(1 + \frac{1}{\gamma\lambda\,(n+1)}\right),
$$

where the first inequality uses $1 - x \leq e^{-x}$ that holds for all $x \in \mathbb{R}$ and the second one uses $\frac{1}{1-e^{-x}} \leq \frac{1}{x} + 1$ that holds for all $x > 0$. The second statement is proven as

$$
\begin{aligned}
c_n^2 \sum_{t=0}^{n} (1 - \gamma\lambda)^{2t} &= \frac{\sum_{t=0}^{n} (1 - \gamma\lambda)^{2t}}{\left(\sum_{t=0}^{n} (1 - \gamma\lambda)^t\right)^2} = \frac{(\gamma\lambda)^2}{\left(1 - (1 - \gamma\lambda)^2\right)} \cdot \frac{\left(1 - (1 - \gamma\lambda)^{2(n+1)}\right)}{\left(1 - (1 - \gamma\lambda)^{n+1}\right)^2} \\
&= \frac{(\gamma\lambda)^2}{\gamma\lambda\,(2 - \gamma\lambda)} \cdot \frac{1 + (1 - \gamma\lambda)^{n+1}}{1 - (1 - \gamma\lambda)^{n+1}} \leq \frac{\gamma\lambda}{2 - \gamma\lambda} \cdot \frac{1 + e^{-\gamma\lambda(n+1)}}{1 - e^{-\gamma\lambda(n+1)}} \\
&= \frac{\gamma\lambda}{2 - \gamma\lambda} \cdot \left(1 + \frac{2e^{-\gamma\lambda(n+1)}}{1 - e^{-\gamma\lambda(n+1)}}\right) \leq \frac{\gamma\lambda}{2 - \gamma\lambda} \cdot \left(1 + \frac{2}{\gamma\lambda\,(n+1)}\right),
\end{aligned}
$$

where the first inequality again uses $1 - x \leq e^{-x}$ that holds for all $x \in \mathbb{R}$ and the second one uses $\frac{e^{-x}}{1-e^{-x}} \leq \frac{1}{x}$ that holds for all $x > 0$. ■

B.2. The proof of Lemma 6

We start by noticing that

$$\left\|\Sigma^{1/2}\left(\widetilde{w}_n - w^*\right)\right\|^2 = c_n^2 \sum_{t=0}^{n}\sum_{k=0}^{n}(1-\gamma\lambda)^{t+k}\big\langle w_t - w^*, \Sigma\left(w_k - w^*\right)\big\rangle \tag{10}$$

$$= c_n^2\sum_{t=0}^{n}(1-\gamma\lambda)^{2t}\big\langle w_t - w^*, \Sigma\left(w_t - w^*\right)\big\rangle + 2c_n^2\sum_{t=0}^{n}\sum_{k=t+1}^{n}(1-\gamma\lambda)^{t+k}\big\langle w_t - w^*, \Sigma\left(w_k - w^*\right)\big\rangle.$$

To handle the second term, we first notice that for any t and $k > t$, we have

$$\begin{aligned}\mathbb{E}_t\left[\big\langle w_t - w^*, \Sigma\left(w_k - w^*\right)\big\rangle\right] &= \mathbb{E}_t\left[\left\langle w_t - w^*, \Sigma\left(M_{t,k}\left(w_t - w^*\right) + \eta\sum_{j=t+1}^{k} M_{j,k}\varepsilon_j x_j\right)\right\rangle\right] \\ &= \left\langle w_t - w^*, \Sigma\left(I - \gamma\Sigma\right)^{k-t}\left(w_t - w^*\right)\right\rangle,\end{aligned}$$

where we used $\mathbb{E}_t\left[\varepsilon_j x_j\right] = 0$ that holds for $j > t$ and $\mathbb{E}_t\left[M_{t,k}\right] = (I - \gamma\Sigma)^{k-t}$. Using this insight, we obtain

$$\begin{aligned}
&c_n^2\mathbb{E}\left[\sum_{t=0}^{n}\sum_{k=t+1}^{n}(1-\gamma\lambda)^{t+k}\big\langle w_t - w^*, \Sigma\left(w_k - w^*\right)\big\rangle\right] \\
&\quad = c_n^2\mathbb{E}\left[\sum_{t=0}^{n}\sum_{k=t+1}^{n}(1-\gamma\lambda)^{t+k}\big\langle w_t - w^*, \Sigma\left(I - \gamma\Sigma\right)^{k-t}\left(w_t - w^*\right)\big\rangle\right] \\
&\quad = c_n^2\mathbb{E}\left[\sum_{t=0}^{n}(1-\gamma\lambda)^{t}\left\langle w_t - w^*, \Sigma\left(\sum_{k=t+1}^{n}(1-\gamma\lambda)^{k}\left(I - \eta\Sigma\right)^{k-t}\left(w_t - w^*\right)\right)\right\rangle\right] \\
&\quad = c_n^2\mathbb{E}\left[\sum_{t=0}^{n}(1-\gamma\lambda)^{t}\left\langle w_t - w^*, \Sigma\left(\sum_{k=t+1}^{n}(1-\gamma\lambda)^{t}\left(I - \gamma\Sigma - \gamma\lambda I\right)^{k-t}\left(w_t - w^*\right)\right)\right\rangle\right] \\
&\quad \le c_n^2\mathbb{E}\left[\sum_{t=0}^{n}(1-\gamma\lambda)^{2t}\left\langle w_t - w^*, \Sigma\left(\sum_{k=t+1}^{\infty}\left(I - \gamma\Sigma - \gamma\lambda I\right)^{k-t}\left(w_t - w^*\right)\right)\right\rangle\right] \\
&\qquad\qquad \text{(adding nonnegative terms to the sum)} \\
&\quad = \frac{c_n^2}{\gamma}\mathbb{E}\left[\sum_{t=0}^{n}(1-\gamma\lambda)^{2t}\left\langle w_t - w^*, \Sigma\left(\Sigma + \lambda I\right)^{-1}\left(I - \gamma\Sigma - \gamma\lambda I\right)\left(w_t - w^*\right)\right\rangle\right] \\
&\qquad\qquad \text{(using } \textstyle\sum_{j=1}^{\infty}(I - A)^j = A^{-1}\left(I - A\right) \text{ that holds for } A \preccurlyeq I\text{)} \\
&\quad \le \frac{c_n^2}{\gamma}\mathbb{E}\left[\sum_{t=0}^{n}(1-\gamma\lambda)^{2t}\left\langle w_t - w^*, \Sigma\left(\Sigma + \lambda I\right)^{-1}\left(w_t - w^*\right)\right\rangle\right] \\
&\qquad\qquad\qquad - c_n^2\mathbb{E}\left[\sum_{t=0}^{n}(1-\gamma\lambda)^{2t}\left\langle w_t - w^*, \Sigma\left(w_t - w^*\right)\right\rangle\right].
\end{aligned}$$

Noticing that the last term matches the first term on the right-hand side of Equation (10), the proof is concluded. ■

B.3. The proof of Lemma 7

The proof of this lemma crucially relies on the following inequality:

Lemma 9 *Assume that $\eta \leq \frac{1}{2R^2}$. Then, for all $k < t$, we have*

$$\mathbb{E}\left[M_{k+1,t}\Sigma M_{k+1,t}^*\right] \preccurlyeq \frac{1}{\eta\left(2-\eta R^2\right)}\left(\mathbb{E}\left[M_{k+1,t}\Sigma\left(\Sigma+\lambda I\right)^{-1}M_{k+1,t}^*\right]-\mathbb{E}\left[M_{k,t}\Sigma\left(\Sigma+\lambda I\right)^{-1}M_{k,t}^*\right]\right).$$

Proof We have

$$\begin{aligned}
&\mathbb{E}\left[M_{k,t}\Sigma\left(\Sigma+\lambda I\right)^{-1}M_{k,t}^*\right] \\
&= \mathbb{E}\left[M_{k+1,t}\left(I-\eta x_{k+1}\otimes x_{k+1}\right)\Sigma\left(\Sigma+\lambda I\right)^{-1}\left(I-\eta x_{k+1}\otimes x_{k+1}\right)M_{k+1,t}^*\right] \\
&= \mathbb{E}\left[M_{k+1,t}\Sigma\left(\Sigma+\lambda I\right)^{-1}\left(I-2\eta x_{k+1}\otimes x_{k+1}\right)M_{k+1,t}^*\right] \\
&\quad + \eta^2\mathbb{E}\left[M_{k+1,t}\left(x_{k+1}\otimes x_{k+1}\right)\Sigma\left(\Sigma+\lambda I\right)^{-1}\left(x_{k+1}\otimes x_{k+1}\right)M_{k+1,t}^*\right] \\
&\preccurlyeq \mathbb{E}\left[M_{k+1,t}\left(\Sigma\left(\Sigma+\lambda I\right)^{-1}-2\eta\Sigma^2\left(\Sigma+\lambda I\right)^{-1}+\eta^2R^2\Sigma^2\left(\Sigma+\lambda I\right)^{-1}\right)M_{k+1,t}^*\right] \\
&= \mathbb{E}\left[M_{k+1,t}\Sigma\left(\Sigma+\lambda I\right)^{-1}M_{k+1,t}^*\right]-\eta\left(2-\eta R^2\right)\mathbb{E}\left[M_{k+1,t}\Sigma^2\left(\Sigma+\lambda I\right)^{-1}M_{k+1,t}\right]
\end{aligned}$$

Reordering, we obtain

$$\mathbb{E}\left[M_{k+1,t}\Sigma^2\left(\Sigma+\lambda I\right)^{-1}M_{k+1,t}\right] \leq \frac{1}{\eta\left(2-\eta R^2\right)}\left(\mathbb{E}\left[M_{k+1,t}\Sigma\left(\Sigma+\lambda I\right)^{-1}M_{k+1,t}^*\right]-\mathbb{E}\left[M_{k,t}\Sigma\left(\Sigma+\lambda I\right)^{-1}M_{k,t}^*\right]\right).$$

The result follows from noticing that $\Sigma \preccurlyeq \Sigma^2\left(\Sigma+\lambda I\right)^{-1}$. ■

Now we are ready to prove Lemma 7. Let us fix t and observe that

$$\begin{aligned}
\Delta_{2,t} &= \mathbb{E}\left[\sum_{k=1}^{t}\sum_{j=1}^{t}\left\langle\varepsilon_j M_{j,t}x_j, \Sigma\left(\Sigma+\lambda I\right)^{-1}M_{k,t}x_k\varepsilon_k\right\rangle\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{t}\left\langle\varepsilon_k M_{k,t}x_k, \Sigma\left(\Sigma+\lambda I\right)^{-1}M_{k,t}x_k\varepsilon_k\right\rangle\right] \\
&\qquad + 2\mathbb{E}\left[\sum_{k=1}^{t}\sum_{j=k+1}^{t}\left\langle\varepsilon_j M_{j,t}x_j, \Sigma\left(\Sigma+\lambda I\right)^{-1}M_{k,t}x_k\varepsilon_k\right\rangle\right] \\
&= \operatorname{tr}\left[\mathbb{E}\left[\sum_{k=1}^{t}\varepsilon_k^2 M_{k,t}\left(x_k\otimes x_k\right)M_{k,t}^*\Sigma\left(\Sigma+\lambda I\right)^{-1}\right]\right] \qquad \text{(using that } \mathbb{E}_k\left[\varepsilon_j x_j\right]=0 \text{ for } j>k\text{)} \\
&\leq \sigma^2\operatorname{tr}\left[\sum_{k=1}^{t}M_{k,t}\Sigma M_{k,t}^*\Sigma\left(\Sigma+\lambda I\right)^{-1}\right],
\end{aligned}$$

where the last inequality uses our assumption on the noise that $\mathbb{E}\left[\varepsilon_k^2 \left(x_k \otimes x_k\right)\right] \preccurlyeq \sigma^2\Sigma$. Now, using Lemma 9, we obtain

$$
\begin{aligned}
&\operatorname{tr}\left[\sum_{k=1}^{t} \mathbb{E}\left[M_{k,t}\Sigma M_{k,t}\right]^* \Sigma\left(\Sigma+\lambda I\right)^{-1}\right] \\
&\leq \frac{1}{\eta\left(2-\eta R^2\right)}\operatorname{tr}\left[\sum_{k=1}^{t} \mathbb{E}\left[M_{k,t}\Sigma\left(\Sigma+\lambda I\right)^{-1} M_{k,t}^*\right] \Sigma\left(\Sigma+\lambda I\right)^{-1}\right] \\
&\qquad - \frac{1}{\eta\left(2-\eta R^2\right)}\operatorname{tr}\left[\sum_{k=1}^{t} \mathbb{E}\left[M_{k-1,t}\Sigma\left(\Sigma+\lambda I\right)^{-1} M_{k-1,t}^*\right] \Sigma\left(\Sigma+\lambda I\right)^{-1}\right] \\
&= \frac{1}{\eta\left(2-\eta R^2\right)}\operatorname{tr}\left[\left(\mathbb{E}\left[M_{t,t}\Sigma\left(\Sigma+\lambda I\right)^{-1} M_{t,t}^*\right] - \mathbb{E}\left[M_{0,t}\Sigma\left(\Sigma+\lambda I\right)^{-1} M_{0,t}^*\right]\right) \Sigma\left(\Sigma+\lambda I\right)^{-1}\right] \\
&\leq \frac{1}{\eta\left(2-\eta R^2\right)}\operatorname{tr}\left[\Sigma^2\left(\Sigma+\lambda I\right)^{-2}\right],
\end{aligned}
$$

where the last step uses the definition $M_{t,t} = I$. This concludes the proof. ■

B.4. The proof of Lemma 8

We begin by noticing that

$$
\Delta_1 \leq \sum_{t=0}^{\infty}(1-\gamma\lambda)^{2t} \cdot \mathbb{E}\left[\operatorname{tr}\left[M_{0,t}^*\Sigma\left(\Sigma+\lambda I\right)^{-1} M_{0,t}\left(w_0-w^*\right)\otimes\left(w_0-w^*\right)\right]\right]
$$

holds since the sum only has positive elements. Following Dieuleveut et al. (2017) again, we define the operator $\mathcal{T}$ acting on an arbitrary Hermitian operator A as

$$
\mathcal{T}A = \Sigma A + A\Sigma - \eta\mathbb{E}\left[\langle x_t, Ax_t\rangle\, x_t \otimes x_t\right],
$$

and also introduce the operator $\mathcal{S}$ defined as

$$
\mathcal{S}A = \mathbb{E}\left[\langle x_t, Ax_t\rangle\, x_t \otimes x_t\right],
$$

so that $\mathcal{T}A = \Sigma A + A\Sigma - \eta\mathcal{S}A$. We note that $\mathcal{S}$ and $\mathcal{T}$ are Hermitian and positive definite (the latter being true by our assumption about η). Finally, we define $\mathcal{I}$ as the identity operator acting on Hermitian matrices. With this notation, we can write

$$
\mathbb{E}\left[M_{0,t}^* A M_{0,t}\right] = \left(\mathcal{I}-\eta\mathcal{T}\right)^t A.
$$

Thus, defining $E_0 = \left(w_0-w^*\right)\otimes\left(w_0-w^*\right)$, we have

$$
\begin{aligned}
&\sum_{t=0}^{\infty}(1-\gamma\lambda)^{2t}\mathbb{E}\left[\operatorname{tr}\left[M_{0,t}^*\Sigma\left(\Sigma+\lambda I\right)^{-1} M_{0,t}E_0\right]\right] \\
&\qquad = \sum_{t=0}^{\infty}(1-\gamma\lambda)^{2t}\operatorname{tr}\left[\left(\mathcal{I}-\eta\mathcal{T}\right)^t\left[\Sigma\left(\Sigma+\lambda I\right)^{-1}\right] E_0\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{t=0}^{\infty} (1-\gamma\lambda)^t \text{tr}\left[(\mathcal{I} - \eta(1-\gamma\lambda)\mathcal{T} - \gamma\lambda\mathcal{I})^t \left[\Sigma(\Sigma+\lambda I)^{-1} \right] E_0 \right] \\
&\leq \sum_{t=0}^{\infty} \text{tr}\left[(\mathcal{I} - \gamma\mathcal{T} - \gamma\lambda\mathcal{I})^i \left[\Sigma(\Sigma+\lambda I)^{-1} \right] E_0 \right] \\
&= \frac{1}{\gamma} \text{tr}\left[(\mathcal{T} + \lambda\mathcal{I})^{-1} \left[\Sigma(\Sigma+\lambda I)^{-1} \right] E_0 \right],
\end{aligned}
$$

where the last step holds true if $\|\mathcal{I} - \eta\mathcal{T}\| < 1$. For a proof of this fact, we refer to Lemma 5 in Défossez and Bach (2015). Let us define $\mathcal{T}_\lambda = \mathcal{T} + \lambda\mathcal{I}$ and $W = \mathcal{T}_\lambda^{-1}\left[\Sigma(\Sigma+\lambda I)^{-1}\right]$, so that it remains to bound $\gamma^{-1}\text{tr}[WE_0]$. We notice that, by definition, W satisfies

$$
\Sigma(\Sigma+\lambda I)^{-1} = \Sigma W + W\Sigma + \lambda W - \eta\mathcal{S}W. \tag{11}
$$

Also introducing the operators $\mathcal{U}_L$ and $\mathcal{U}_R$ as the left- and right-multiplication operators with Σ, respectively, we get after reordering that

$$
\begin{aligned}
W &= (\mathcal{U}_L + \mathcal{U}_R + \lambda I)^{-1} \Sigma(\Sigma+\lambda I)^{-1} + \eta(\mathcal{U}_L + \mathcal{U}_R + \lambda I)^{-1}\mathcal{S}W \\
&= \frac{1}{2}\Sigma(\Sigma+\lambda I)^{-1}\left(\Sigma + \tfrac{\lambda}{2}I\right)^{-1} + \eta(\mathcal{U}_L + \mathcal{U}_R + \lambda I)^{-1}\mathcal{S}W.
\end{aligned}
$$

Using the fact that $\mathcal{U}_L + \mathcal{U}_R + \lambda I$ and its inverse are Hermitian, we can show

$$
\text{tr}[WE_0] = \frac{1}{2}\text{tr}\left[\Sigma(\Sigma+\lambda I)^{-1}\left(\Sigma+\tfrac{\lambda}{2}I\right)^{-1}E_0\right] + \eta\text{tr}\left[\mathcal{S}W(\mathcal{U}_L + \mathcal{U}_R + \lambda I)^{-1}E_0\right].
$$

Furthermore, by again following the arguments[5] of Dieuleveut et al. (2017, pp. 28), we can also show

$$
(\mathcal{U}_L + \mathcal{U}_R + \lambda I)^{-1}E_0 \preccurlyeq \frac{\left\langle w_0 - w^*, \left(\Sigma+\tfrac{\lambda}{2}I\right)^{-1}(w_0 - w^*)\right\rangle}{2} I.
$$

Since $\mathcal{S}W$ is positive, this leads to the bound

$$
\text{tr}[WE_0] \leq \frac{1}{2}\text{tr}\left[\Sigma(\Sigma+\lambda I)^{-1}\left(\Sigma+\tfrac{\lambda}{2}I\right)^{-1}E_0\right] + \frac{\eta\left\langle w_0 - w^*, \left(\Sigma+\tfrac{\lambda}{2}I\right)^{-1}(w_0-w^*)\right\rangle}{2}\text{tr}[\mathcal{S}W], \tag{12}
$$

so it remains to bound $\text{tr}[\mathcal{S}W]$. On this front, we have

$$
\text{tr}[\mathcal{S}W] = \text{tr}[\mathbb{E}[\langle x_t, Wx_t\rangle x_t \otimes x_t]] \leq R^2\text{tr}[W\Sigma]
$$

by our assumption on the covariates. Also, by Equation (11), we have

$$
\begin{aligned}
\text{tr}\left[\Sigma(\Sigma+\lambda I)^{-1}\right] &= 2\text{tr}[W\Sigma] + \lambda\text{tr}[W] - \eta\text{tr}[\mathcal{S}W] \\
&= 2\text{tr}\left[W\left(\Sigma+\tfrac{\lambda}{2}I\right)\right] - \eta\text{tr}[\mathcal{S}W]
\end{aligned}
$$

5. This result is proven for finite dimension by Dieuleveut et al. (2017), but can be easily generalized to infinite dimensions.

$$\begin{aligned}
&\geq 2\mathrm{tr}\left[W\Sigma\right] - \eta\mathrm{tr}\left[\mathcal{S}W\right] \\
&\geq \left(\frac{2}{R^2} - \eta\right)\mathrm{tr}\left[\mathcal{S}W\right] \\
&\geq \frac{\mathrm{tr}\left[\mathcal{S}W\right]}{R^2},
\end{aligned}$$

by crucially using the assumption $\eta \leq 1/2R^2$. Plugging into Equation (12) and using $\eta R^2 \leq 2$ again proves the lemma. ■

Appendix C. Analysis under additive noise

In this section, we consider the "additive-noise" model of Dieuleveut et al. (2017). This model assumes that the learner has access to the gradient estimator $\Sigma w - y_t x_t$, that is, the gradient is subject to the noise vector $\xi_t = y_t x_t - \mathbb{E}\left[y_t x_t\right]$ which doesn't depend on the parameter vector w. This assumption is admittedly very strong, and we mainly include our analysis for this case for didactic purposes. Indeed, the analysis for this case is significantly simpler than in the setting considered in the main body of our paper.

In this setting, stochastic gradient descent takes the form

$$w_t = w_{t-1} - \eta(\Sigma w_{t-1} - x_t y_t), \tag{13}$$

which is the unregularized counterpart of the iteration already introduced in Section 2 as Equation (4). Again, we will study the geometric average

$$\widetilde{w}_t = \frac{\sum_{k=0}^{t}\left(1-\eta\lambda\right)^k w_k}{\sum_{j=0}^{t}(1-\eta\lambda)^j}. \tag{14}$$

For the analysis, we recall the definition $\xi_t = x_t y_t - \mathbb{E}\left[x_t y_t\right]$ and study the evolution of $\widetilde{w}_t - w^*$:

$$\begin{aligned}
\widetilde{w}_t - w^* &= \left(I - \eta\Sigma\right)\left(\widetilde{w}_{t-1} - w^*\right) + \eta\xi_t \\
&= \left(I - \eta\Sigma\right)^t\left(\widetilde{w}_0 - w^*\right) + \eta\sum_{k=1}^{t}\left(I - \eta\Sigma\right)^{t-k}\xi_k.
\end{aligned}$$

We prove the following performance guarantee about this algorithm:

Proposition 10 *Suppose that $V = \mathbb{E}\left[\xi_t \otimes \xi_t\right] \leq \tau^2\Sigma$ for some $\sigma^2 > 0$ and assume that $\eta \leq \lambda_{\max}(\Sigma)$ and $\lambda \in [0, 1/\eta)$. Then, the iterates computed by the recursion given by Equations (13) and (14) satisfy*

$$\begin{aligned}
\mathbb{E}\left[\Delta(\widetilde{w}_n)\right] \leq &\left(\frac{\gamma\lambda}{(2-\gamma\lambda)} + \frac{2}{(2-\gamma\lambda)\left(n+1\right)}\right)\mathrm{tr}\left[\Sigma\left(\Sigma + \lambda I\right)^{-2} V\right] \\
&+ \left(\lambda + \frac{1}{\gamma\left(n+1\right)}\right)^2 \left\|\Sigma^{1/2}\left(\Sigma + \lambda I\right)^{-1}\left(w_0 - w^*\right)^2\right\|^2.
\end{aligned}$$

Proof Recalling the notation $c_n = \left(\sum_{t=0}^{n}(1-\gamma\lambda)^t\right)^{-1}$, the geometric average can be written as

$$\begin{aligned}
&\widetilde{w}_n - w^* = c_n \cdot \sum_{t=0}^{n} (1-\gamma\lambda)^t (w_t - w^*) \\
&= c_n \cdot \sum_{t=0}^{n} (1-\gamma\lambda)^t \left((I-\eta\Sigma)^t (w_0 - w^*) + \eta \sum_{j=1}^{t} (I-\eta\Sigma)^{t-j} \xi_j \right) \\
&= c_n \cdot \left(\sum_{t=0}^{n} (1-\gamma\lambda)^t (I-\eta\Sigma)^t (w_0 - w^*) + \eta \sum_{t=0}^{n} (1-\gamma\lambda)^t \sum_{j=1}^{t} (I-\eta\Sigma)^{t-j} \xi_j \right) \\
&= c_n \cdot \left(\sum_{t=0}^{n} (I-\gamma\Sigma-\gamma\lambda I)^t (w_0 - w^*) + \eta \sum_{j=1}^{n} (1-\gamma\lambda)^j \sum_{t=j}^{n} (1-\gamma\lambda)^{t-j} (I-\eta\Sigma)^{t-j} \xi_j \right) \\
&= c_n \cdot \left(\sum_{t=0}^{n} (I-\gamma\Sigma-\gamma\lambda I)^t (w_0 - w^*) + \eta \sum_{j=1}^{n} (1-\gamma\lambda)^j \sum_{t=j}^{n} (I-\gamma\Sigma-\gamma\lambda I)^{t-j} \xi_j \right) \\
&= c_n \cdot \left(\sum_{t=0}^{n} (I-\gamma\Sigma-\gamma\lambda I)^t (w_0 - w^*) + \eta \sum_{j=1}^{n} (1-\gamma\lambda)^j \sum_{t=0}^{n-j} (I-\gamma\Sigma-\gamma\lambda I)^{t} \xi_j \right) \\
&= c_n \cdot \Bigg(\frac{1}{\gamma} \left(I - (I-\gamma\Sigma-\gamma\lambda I)^{n+1} \right) (\Sigma+\lambda I)^{-1} (w_0 - w^*) \\
&\qquad + \frac{\eta}{\gamma} \sum_{j=1}^{n} (1-\gamma\lambda)^j \left(I - (I-\gamma\Sigma-\gamma\lambda I)^{n-j+1} \right) (\Sigma+\lambda I)^{-1} \xi_j \Bigg)
\end{aligned}$$

Now, exploiting the assumption that the noise is i.i.d. and zero-mean, we get

$$\begin{aligned}
\mathbb{E}\left[\Delta(\widetilde{w}_n)\right] &= \mathbb{E}\left[\left\| \Sigma^{1/2} (\widetilde{w}_n - w^*) \right\|^2 \right] \\
&= c_n^2 \cdot \left(\frac{\eta}{\gamma}\right)^2 \sum_{j=1}^{n} (1-\gamma\lambda)^{2j} \operatorname{tr}\left[\left(I - (I-\gamma\Sigma-\gamma\lambda I)^{n-j+1} \right)^2 \Sigma (\Sigma+\lambda I)^{-2} V \right] \\
&\quad + c_n^2 \cdot \frac{1}{\gamma^2} \left\| \left(I - (I-\gamma\Sigma-\gamma\lambda I)^{n+1} \right) \Sigma^{1/2} (\Sigma+\lambda)^{-1} (w_0 - w^*) \right\|^2 \\
&\le c_n^2 \left(\frac{\eta(1-\gamma\lambda)}{\gamma} \right)^2 \sum_{j=0}^{n-1} (1-\gamma\lambda)^{2j} \operatorname{tr}\left[\Sigma (\Sigma+\lambda I)^{-2} V \right] \\
&\quad + c_n^2 \cdot \left(\frac{1}{\gamma}\right)^2 \left\| \Sigma^{1/2} (\Sigma+\lambda I)^{-1} (w_0 - w^*) \right\|^2,
\end{aligned}$$

The proof is concluded by observing that $\gamma = \eta(1-\gamma\lambda)$ and appealing to Lemma 5. ∎