Journal of Machine Learning Research 18 (2017) 1-51 Submitted 7/16; Revised 4/17; Published 10/17

Harder, Better, Faster, Stronger[†] Convergence Rates for Least-Squares Regression

Aymeric Dieuleveut AYMERIC.DIEULEVEUT@ENS.FR
Nicolas Flammarion NICOLAS.FLAMMARION@ENS.FR
Francis Bach FRANCIS.BACH@ENS.FR

INRIA
Département d'Informatique de l'ENS, École normale supérieure, CNRS, PSL Research University Paris, France.

Editor: Alexander Rakhlin

Abstract

We consider the optimization of a quadratic objective function whose gradients are only accessible through a stochastic oracle that returns the gradient at any given point plus a zero-mean finite variance random error. We present the first algorithm that achieves jointly the optimal prediction error rates for least-squares regression, both in terms of forgetting the initial conditions in $O(1/n^2)$, and in terms of dependence on the noise and dimension d of the problem, as $O(d/n)$. Our new algorithm is based on averaged accelerated regularized gradient descent, and may also be analyzed through finer assumptions on initial conditions and the Hessian matrix, leading to dimension-free quantities that may still be small in some distances while the "optimal" terms above are large. In order to characterize the tightness of these new bounds, we consider an application to non-parametric regression and use the known lower bounds on the statistical performance (without computational limits), which happen to match our bounds obtained from a single pass on the data and thus show optimality of our algorithm in a wide variety of particular trade-offs between bias and variance.

Keywords: convex optimization, least-squares regression, stochastic gradient, accelerated gradient, non-parametric estimation

1. Introduction

Many supervised machine learning problems are naturally cast as the minimization of a smooth function defined on a Euclidean space. This includes least-squares regression, logistic regression (see, e.g., Hastie et al., 2009) or generalized linear models (McCullagh and Nelder, 1989). While small problems with few or low-dimensional input features may be solved precisely by many potential optimization algorithms (e.g., Newton method), large-scale problems with many high-dimensional features are typically solved with simple gradient-based iterative techniques whose per-iteration cost is small.

. [†] https://youtu.be/yydNF8tuVmU

In this paper, we consider a quadratic objective function f whose gradients are only accessible through a stochastic oracle that returns the gradient at any given point plus a zero-mean finite variance random error. In this stochastic approximation framework (Robbins and Monro, 1951), it is known that two quantities dictate the behavior of various algorithms, namely the covariance matrix V of the noise in the gradients, and the deviation $\theta_0 - \theta_*$ between the initial point of the algorithm θ_0 and any of the global minimizer θ_* of f. This leads to a "bias/variance" decomposition (Bach and Moulines, 2013; Hsu et al., 2014) of the performance of most algorithms as the sum of two terms: (a) the bias term characterizes how fast initial conditions are forgotten and thus is increasing in a well-chosen norm of $\theta_0 - \theta_*$; while (b) the variance term characterizes the effect of the noise in the gradients, independently of the starting point, and with a term that is increasing in the covariance of the noise.

For quadratic functions with (a) a noise covariance matrix V which is proportional (with constant σ^2) to the Hessian of f (a situation which corresponds to least-squares regression) and (b) an initial point characterized by the norm $\|\theta_0 - \theta_*\|^2$, the optimal bias and variance terms are known *separately* from the optimization and statistical theories. On the one hand, the optimal bias dependency after n iterations is proportional to $\frac{L\|\theta_0-\theta_*\|^2}{n^2}$, where L is the largest eigenvalue of the Hessian of f. This rate is achieved by accelerated gradient descent (Nesterov, 1983, 2004), and is known to be optimal if the number of iterations n is less than the dimension d of the underlying predictors, but the algorithm is not robust to random or deterministic noise in the gradients (d'Aspremont, 2008; Schmidt et al., 2011; Devolder et al., 2014). On the other hand, the optimal variance term is proportional to $\frac{\sigma^2 d}{n}$ (Tsybakov, 2008); it is known to be achieved by averaged gradient descent (Bach and Moulines, 2013), for which the bias term only achieves $\frac{L\|\theta_0-\theta_*\|^2}{n}$ instead of $\frac{L\|\theta_0-\theta_*\|^2}{n^2}$.

Our first contribution in this paper is to present a novel algorithm which attains optimal rates for *both the variance and the bias terms*. This algorithm analyzed in Section 4 is averaged accelerated gradient descent; beyond obtaining jointly optimal rates, our result shows that averaging is beneficial for accelerated techniques and provides a provable robustness to noise.

While optimal when measuring performance in terms of the dimension d and the initial distance to optimum $\|\theta_0 - \theta_*\|^2$, these rates are not adapted in many situations where either d is larger than the number of iterations n (i.e., the number of observations for regular stochastic gradient descent) or $L\|\theta_0 - \theta_*\|^2$ is much larger than n^2. Our second contribution is to provide in Section 5 an analysis of a new algorithm (based on some additional regularization) that can adapt our bounds to finer assumptions on $\theta_0 - \theta_*$ and the Hessian of the problem, leading in particular to dimension-free quantities that can thus be extended to the Hilbert space setting (in particular for non-parametric estimation).

In order to characterize the optimality of these new bounds, our third contribution is to consider an application to non-parametric regression in Section 6 and use the known lower bounds on the statistical performance (without computational limits), which happen to match our bounds obtained from a single pass on the data and thus show optimality of our algorithm in a wide variety of particular trade-offs between bias and variance.

Our paper is organized as follows: in Section 2, we present the main problem we tackle, namely least-squares regression, then introduce the two algorithms that we consider in

Section 2.2, as well as the two types of oracles on the gradient in Section 2.3. In Section 3, we present new results for averaged stochastic gradient descent that set the stage for Section 4, where we present our main novel result leading to an accelerated algorithm which is robust to noise. Our tighter analysis of convergence rates based on finer dimension-free quantities is presented in Section 5, and their optimality for kernel-based non-parametric regression is studied in Section 6. Organization of the main results is summarized in the Table 1 bellow.

	Averaged Algo.	**Averaged Accelerated Algo.**
Dimension dependent rates	*Section 3*	*Section 4*
Additive Noise	Lemma 1$^{\diamondsuit}$	Theorem 3
Multiplicative Noise	Theorem 2$^{\diamondsuit}$	$\natural$
Dimension independent rates	*Section 5*	*Section 5*
Additive Noise	$\sharp$	Theorem 5
Multiplicative Noise	4^{th} remark after Cor. 6$^{\flat}$	$\natural$
Kernel regression setting	*Section 6*	*Section 6*
Additive Noise	$\sharp$	Theorem 8
Multiplicative Noise	Theorem 7$^{\flat}$	$\natural$

Table 1: Organization of the paper. $\diamondsuit$: We extend results from (Bach and Moulines, 2013) to the setting in which extra regularization is added; $\sharp$: apart from Lemma 1 which is useful to develop intuition of the different terms in the upper bound, we do not state result for the averaged algorithm with additive noise, as the most powerful result is for the multiplicative noise; $\flat$: these results recover results from Dieuleveut and Bach (2015) (with the use of an extra regularization); $\natural$: it is still an open problem to get results in the accelerated setting for a multiplicative noise oracle.

2. Least-Squares Regression

In this section, we present our least-squares regression framework, which is risk minimization with the square loss, together with the main assumptions regarding our model and our algorithms. These algorithms will rely on stochastic gradient oracles, which will come in two kinds, an additive noise which does not depend on the current iterate, which will correspond in practice to the full knowledge of the covariance matrix, and a "multiplicative/additive" noise, which corresponds to the regular stochastic gradient obtained from a single pair of observations. This second oracle is much harder to analyze.

2.1 Statistical Assumptions

We consider the following general setting:

- $\mathcal{H}$ is a d-dimensional Euclidean space with $d \geq 1$. The (temporary) restriction to finite dimension will be relaxed in Section 6.

- The observations $(x_n, y_n) \in \mathcal{H} \times \mathbb{R}$, $n \geq 1$, are independent and identically distributed (i.i.d.), and such that $\mathbb{E}\|x_n\|^2$ and $\mathbb{E}y_n^2$ are finite.

- We consider the *least-squares regression* problem, namely the minimization of the expected loss $f(\theta) = \frac{1}{2}\mathbb{E}(\langle x_n, \theta\rangle - y_n)^2$ which is a quadratic function.

We first introduce an assumption on the distribution of x_n.

Covariance matrix. We denote by $\Sigma = \mathbb{E}(x_n \otimes x_n) \in \mathbb{R}^{d\times d}$ the population covariance matrix, which is the Hessian of f at all points. Without loss of generality, we can assume Σ is invertible by reducing $\mathcal{H}$ to the minimal subspace where all x_n, $n \geq 1$, lie almost surely. This implies that all eigenvalues of Σ are strictly positive (but they may be arbitrarily small). Following Bach and Moulines (2013), we assume there exists $R > 0$ such that

$$\mathbb{E}\|x_n\|^2 x_n \otimes x_n \preccurlyeq R^2\Sigma, \tag{$\mathcal{A}_1$}$$

where $A \preccurlyeq B$ means that $B - A$ is positive semi-definite. This assumption implies in particular that (a) $\mathbb{E}\|x_n\|^4$ is finite and (b) $\operatorname{tr}\Sigma = \mathbb{E}\|x_n\|^2 \leq R^2$ since taking the trace of the previous inequality we get $\mathbb{E}\|x_n\|^4 \leq R^2\mathbb{E}\|x_n\|^2$ and using Cauchy-Schwarz inequality we get $\mathbb{E}\|x_n\|^2 \leq \sqrt{\mathbb{E}\|x_n\|^4} \leq R\sqrt{\mathbb{E}\|x_n\|^2}$.

Assumption $(\mathcal{A}_1)$ is satisfied, for example, for least-square regression with almost surely bounded data, since $\|x_n\|^2 \leq R^2$ almost surely implies $\mathbb{E}\|x_n\|^2 x_n \otimes x_n \preccurlyeq \mathbb{E}\big[R^2 x_n \otimes x_n\big] = R^2\Sigma$. This assumption is also true for data with infinite support and a bounded *kurtosis* for the projection of the covariates x_n on any direction $z \in \mathcal{H}$, e.g, for which there exists $\kappa > 0$, such that:

$$\forall z \in \mathcal{H}, \quad \mathbb{E}\langle z, x_n\rangle^4 \leq \kappa\langle z, \Sigma z\rangle^2. \tag{1}$$

Indeed, by Cauchy-Schwarz inequality, Equation (1) implies for all $(z, t) \in \mathcal{H}^2$, the following bound $\mathbb{E}\langle z, x_n\rangle^2\langle t, x_n\rangle^2 \leq \kappa\langle z, \Sigma z\rangle\langle t, \Sigma t\rangle$, which in turn implies that for all positive semi-definite symmetric matrices M, N, we have $\mathbb{E}\langle x_n, Mx_n\rangle\langle x_n, Nx_n\rangle \leq \kappa \operatorname{tr}(M\Sigma)\operatorname{tr}(N\Sigma)$. Equation (1), which is true for Gaussian vectors with $\kappa = 3$, thus implies $(\mathcal{A}_1)$ for $R^2 = \kappa \operatorname{tr}\Sigma = \kappa\mathbb{E}\|x_n\|^2$.

In the next two paragraphs, we introduce some quantities that will be important in the analysis, in order to get tighter bounds.

Eigenvalue decay. Most convergence bounds depend on the dimension d of $\mathcal{H}$. However it is possible to derive dimension-free and often tighter convergence rates by considering bounds depending on the value $\operatorname{tr}\Sigma^b$ for $b \in [0, 1]$. Given b, if we consider the eigenvalues of Σ ordered in decreasing order, which we denote by s_i, then $\operatorname{tr}\Sigma^b = \sum_i s_i^b$, and the eigenvalues decay[1] at least as $\frac{(\operatorname{tr}\Sigma^b)^{1/b}}{i^{1/b}}$. Moreover, it is known that $(\operatorname{tr}\Sigma^b)^{1/b}$ is decreasing in b and thus, the smaller the b, the stronger the assumption. For b going to 0 then $\operatorname{tr}\Sigma^b$ tends to d and we are back in the classical low-dimensional case. When $b = 1$, we simply get $\operatorname{tr}\Sigma = \mathbb{E}\|x_n\|^2$, which will correspond to the weakest assumption in our context.

1. Indeed for any $i \geq 1$, we have $is_i^b \leq \sum_{t=1}^i s_t^b \leq \operatorname{tr}(\Sigma^b)$.

Optimal predictor. In finite dimension the regression function $f(\theta) = \frac{1}{2}\mathbb{E}(\langle x_n, \theta\rangle - y_n)^2$ always admits a global minimum $\theta_* = \Sigma^\dagger \mathbb{E}(y_n x_n)$. When initializing algorithms at $\theta_0 = 0$ or regularizing by the squared norm, rates of convergence generally depend on $\|\theta_*\|$, a quantity which could be arbitrarily large.

However there exists a systematic upper-bound[2] $\|\Sigma^{\frac{1}{2}}\theta_*\| \leq 2\sqrt{\mathbb{E}y_n^2}$. This leads naturally to the consideration of convergence bounds depending on $\|\Sigma^{r/2}\theta_*\|$ for $r \leq 1$. In infinite dimension this will correspond to assuming $\|\Sigma^{r/2}\theta_*\| < \infty$. This new assumption relates the optimal predictor with sources of ill-conditioning (since Σ is the Hessian of the objective function f), the smaller r, the stronger our assumption, with $r = 1$ corresponding to no assumption at all, $r = 0$ to θ_* in $\mathcal{H}$ and $r = -1$ to a convergence of the bias of least-squares regression with averaged stochastic gradient descent in $O\big(\frac{\|\Sigma^{-1/2}\theta_*\|^2}{n^2}\big)$(Dieuleveut and Bach, 2015; Défossez and Bach, 2015). In this paper, we will use arbitrary initial points θ_0 and thus our bounds will depend on $\|\Sigma^{r/2}(\theta_0 - \theta_*)\|$.

Finally , we make an assumption on the joint distribution of (x_n, y_n).

Noise. We denote by $\varepsilon_n = y_n - \langle \theta_*, x_n\rangle$ the residual for which we have $\mathbb{E}[\varepsilon_n x_n] = 0$. Although we do not have $\mathbb{E}[\varepsilon_n | x_n] = 0$ in general unless the model is well-specified, we assume the noise to be a structured process such that there exists $\sigma > 0$ with

$$\mathbb{E}[\varepsilon_n^2 x_n \otimes x_n] \preccurlyeq \sigma^2 \Sigma. \tag{$\mathcal{A}_2$}$$

Assumption $(\mathcal{A}_2)$ is satisfied for example for data almost surely bounded or when the model is well-specified, (e.g., $y_n = \langle \theta_*, x_n\rangle + \varepsilon_n$, with $(\varepsilon_n)_{n\in\mathbb{N}}$ i.i.d. of variance σ^2 and independent of x_n).

2.2 Averaged Gradient Methods and Acceleration

We focus in this paper on stochastic gradient methods with and without acceleration for the least-squares function regularized by $\frac{\lambda}{2}\|\theta - \theta_0\|^2$ for $\lambda \in \mathbb{R}^+$. The regularization will be useful when deriving tighter convergence rates in Section 5, and it has the additional benefit of making the problem λ-strongly-convex. Stochastic gradient descent (referred to from now on as "SGD"), applied to the regularized problem, can be described for $n \geq 1$ as

$$\theta_n = \theta_{n-1} - \gamma f_n'(\theta_{n-1}) - \gamma\lambda(\theta_{n-1} - \theta_0), \tag{2}$$

starting from $\theta_0 \in \mathcal{H}$, where $\gamma > 0$ is either called the step-size in optimization or the learning rate in machine learning, and $f_n'(\theta_{n-1})$ is an unbiased estimate of the gradient of f at θ_{n-1}, that is, its conditional expectation given all other sources of randomness is equal to $f'(\theta_{n-1})$.

Accelerated stochastic gradient descent is defined, for the regularized problem, by an iterative system with two parameters (θ_n, ν_n) satisfying for $n \geq 1$

$$\begin{aligned} \theta_n &= \nu_{n-1} - \gamma f_n'(\nu_{n-1}) - \gamma\lambda(\nu_{n-1} - \theta_0) \\ \nu_n &= \theta_n + \delta\big(\theta_n - \theta_{n-1}\big), \end{aligned} \tag{3}$$

2. Indeed for all $\theta \in \mathbb{R}^d$ and in particular $\theta = 0$, by Minkowski's inequality, $\|\Sigma^{\frac{1}{2}}\theta_*\| - \sqrt{\mathbb{E}y_n^2} = \sqrt{\mathbb{E}\langle\theta_*, x_n\rangle^2} - \sqrt{\mathbb{E}y_n^2} \leq \sqrt{\mathbb{E}(\langle\theta_*, x_n\rangle - y_n)^2} \leq \sqrt{\mathbb{E}(\langle\theta, x_n\rangle - y_n)^2} \leq \sqrt{\mathbb{E}(y_n)^2}$.

starting from $\theta_0 = \nu_0 \in \mathcal{H}$, with $\gamma, \delta \in \mathbb{R}^2$ and $f_n'(\theta_{n-1})$ described as before. It may be reformulated as the following second-order recursion

$$\theta_n = (1-\gamma\lambda)\big(\theta_{n-1} + \delta(\theta_{n-1} - \theta_{n-2})\big) - \gamma f_n'\big(\theta_{n-1} + \delta(\theta_{n-1} - \theta_{n-2})\big) + \gamma\lambda\theta_0.$$

The *momentum* coefficient $\delta \in \mathbb{R}$ is chosen to accelerate the convergence rate (Nesterov, 1983; Beck and Teboulle, 2009) and has its roots in the heavy-ball algorithm from Polyak (1964). We especially concentrate here, following Polyak and Juditsky (1992), on the average of the sequence

$$\bar{\theta}_n = \frac{1}{n+1}\sum_{i=0}^{n} \theta_n, \tag{4}$$

and we note that it can be computed online as $\bar{\theta}_n = \frac{n}{n+1}\bar{\theta}_{n-1} + \frac{1}{n+1}\theta_n$.

The key ingredient in the algorithms presented above is the unbiased estimate on the gradient $f_n'(\theta)$, which can take two forms that we now describe in our setting.

2.3 Additive versus Multiplicative Stochastic Oracles on the Gradient

We consider the standard stochastic approximation framework (Kushner and Yin, 2003). That is, we let $(\mathcal{F}_n)_{n\geq 0}$ be the increasing family of σ-fields that are generated by all variables (x_i, y_i) for $i \leq n$, and such that for each $\theta \in \mathcal{H}$ the random variable $f_n'(\theta)$ is square-integrable and $\mathcal{F}_n$-measurable with $\mathbb{E}[f_n'(\theta)|\mathcal{F}_{n-1}] = f'(\theta)$, for all $n \geq 0$. Consequently it is of the form

$$f_n'(\theta) = f'(\theta) - \xi_n, \tag{$\mathcal{A}_3$}$$

where the noise process ξ_n is $\mathcal{F}_n$-measurable with $\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0$ and $\mathbb{E}[\|\xi_n\|^2]$ is finite. We will consider two different gradient oracles.

Additive noise. The first oracle is the sum of the true gradient $f'(\theta)$ and an independent zero-mean noise that does not depend on θ. This oracle is equal to

$$f_n'(\theta) = \Sigma\theta - y_n x_n. \tag{5}$$

Since $f'(\theta) = \Sigma\theta - \mathbb{E}y_n x_n$, the oracle above has a noise vector $\xi_n = y_n x_n - \mathbb{E}y_n x_n$ independent of θ and therefore satisfies Assumption ($\mathcal{A}_3$). Furthermore we also assume that there exists $\tau \in \mathbb{R}$ such that

$$\mathbb{E}[\xi_n \otimes \xi_n] \preccurlyeq \tau^2\Sigma, \tag{$\mathcal{A}_4$}$$

that is, the noise has a particular structure adapted to least-squares regression. For optimal results for unstructured noise, with convergence rate for the noise part in $O(1/\sqrt{n})$, see Lan (2012). Our oracle above with an additive noise which is independent of the current iterate corresponds to the first setting studied in stochastic approximation (Robbins and Monro, 1951; Duflo, 1997; Polyak and Juditsky, 1992). While used by Bach and Moulines (2013) as an artifact of proof, for least-squares regression, such an additive noise corresponds to the situation where the distribution of x is known so that the population covariance matrix is computable, but the distribution of the outputs $(y_n)_{n\in\mathbb{N}}$ remains unknown. Thus it may be seen as an intermediate set-up between regression estimation with fixed and random design (see, e.g., Györfi et al., 2006, Section 1.9).

Assumption $(\mathcal{A}_4)$ will be satisfied, for example if the outputs are almost surely bounded because $\mathbb{E}[\xi_n \otimes \xi_n] \preccurlyeq \mathbb{E}[y_n^2 x_n \otimes x_n] \preccurlyeq \tau^2 \Sigma$ if $y_n^2 \leq \tau^2$ almost surely. But it will also be for data satisfying Equation (1) since we will have

$$\begin{aligned}\mathbb{E}[\xi_n \otimes \xi_n] &\preccurlyeq \mathbb{E}[y_n^2 x_n \otimes x_n] = \mathbb{E}[(\langle \theta_*, x_n \rangle + \varepsilon_n)^2 x_n \otimes x_n] \\ &\preccurlyeq 2\mathbb{E}[\langle \theta_*, x_n \rangle^2 x_n \otimes x_n] + 2\sigma^2 \Sigma \preccurlyeq 2(\kappa \|\Sigma^{1/2}\theta_*\|^2 + \sigma^2)\Sigma \preccurlyeq 2(4\kappa \mathbb{E}[y_n^2] + \sigma^2)\Sigma,\end{aligned}$$

and thus Assumption $(\mathcal{A}_4)$ is satisfied with $\tau^2 = 2(4\kappa \mathbb{E}[y_n^2] + \sigma^2)$.

Stochastic noise ("multiplicative/additive"). This corresponds to:

$$f_n'(\theta) = (\langle x_n, \theta \rangle - y_n) x_n = (\Sigma + \zeta_n)(\theta - \theta_*) - \Xi_n, \tag{6}$$

with $\zeta_n = x_n \otimes x_n - \Sigma$ and $\Xi_n = (y_n - \langle x_n, \theta_* \rangle) x_n = \varepsilon_n x_n$. This oracle corresponds to regular SGD, which is often referred to as the least-mean-square (LMS) algorithm for least-squares regression, where the noise comes from sampling a single pair of observations. While still satisfying Assumption $(\mathcal{A}_3)$, it combines an additive noise Ξ_n independent of θ as in Eq. (5) and a multiplicative noise ζ_n. This multiplicative noise makes this stochastic oracle harder to analyze which explains why it is often approximated by an additive noise oracle. However it is the most widely used and most practical one. Note that for the oracle in Eq. (6), from Equation $(\mathcal{A}_2)$, we have $\mathbb{E}[\Xi_n \otimes \Xi_n] \preccurlyeq \sigma^2 \Sigma$. It has a similar form to Assumption $(\mathcal{A}_4)$ which is valid for the additive noise oracle in Eq. (5): we use different constants σ^2 and τ^2 to highlight the difference between these two oracles.

3. Averaged Stochastic Gradient Descent

In this section, we provide convergence bounds for regularized averaged stochastic gradient descent. The main novelty compared to the work of Bach and Moulines (2013) is (a) the presence of regularization, which will be useful when deriving tighter convergence rates in Section 5 and (b) a much simpler proof. We first consider the additive noise in Section 3.1 before considering the multiplicative/additive noise in Section 3.2.

3.1 Additive Noise

We study here the convergence of the averaged SGD recursion defined by Eq. (2) under the simple oracle defined in Eq. (5). For least-squares regression, it takes the form:

$$\theta_n = \big[I - \gamma \Sigma - \gamma \lambda I\big] \theta_{n-1} + \gamma y_n x_n + \lambda \gamma \theta_0. \tag{7}$$

This is an easy adaptation of the work of Bach and Moulines (2013, Lemma 2) for the regularized case.

Lemma 1 *Assume $(\mathcal{A}_4)$. Consider the recursion in Eq. (7) with any regularization parameter $\lambda \in \mathbb{R}_+$ and any constant step-size γ such that $\gamma(\Sigma + \lambda I) \preccurlyeq I$. Then*

$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \leq \Big(\lambda + \frac{1}{\gamma n}\Big)^2 \|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2 + \frac{\tau^2 \operatorname{tr}\big[\Sigma^2 (\Sigma + \lambda I)^{-2}\big]}{n}. \tag{8}$$

We can make the following observations:

- The proof (see Appendix A) relies on the fact that $\theta_n - \theta_*$ is obtainable in closed form since the cost function is quadratic and thus the recursions are linear, and follows from Polyak and Juditsky (1992).

- The constraint on the step-size γ is equivalent to $\gamma(L+\lambda) \leqslant 1$ where L is the largest eigenvalue of Σ and we thus recover the usual step-size from deterministic gradient descent (Nesterov, 2004).

- When n tends to infinity, the algorithm converges to the minimum of $f(\theta) + \frac{\lambda}{2}\|\theta - \theta_0\|^2$ and our performance guarantee becomes $\lambda^2 \|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2$. This is the standard "bias term" from regularized ridge regression (Hsu et al., 2014) which we naturally recover here. The term $\frac{\tau^2}{n} \operatorname{tr}\left[\Sigma^2(\Sigma + \lambda I)^{-2}\right]$ is usually referred to as the "variance term" (Hsu et al., 2014), and is equal to $\frac{\tau^2}{n}$ times the quantity $\operatorname{tr}\left[\Sigma^2(\Sigma+\lambda I)^{-2}\right]$, which is often called the degrees of freedom of the ridge regression problem (Gu, 2013).

- For finite n, the first term in Eq. (8) is the usual bias term which depends on the distance from the initial point θ_0 to the objective point θ_* with an appropriate norm. It includes a regularization-based component which is proportional to λ^2 and optimization-based component which depends on $(\gamma n)^{-2}$. The regularization-based bias appears because the algorithm tends to minimize the regularized function instead of the true function f.

- Given Eq. (8), it is natural to set $\lambda\gamma = \frac{1}{n}$, and the two components of the bias term are exactly of the same order leading to $\frac{4}{\gamma^2 n^2}\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2$. It corresponds up to a constant factor to the bias term of regularized least-squares (Hsu et al., 2014), but it is achieved by an algorithm accessing only n stochastic gradients. Note that when λ or γ depend on n, this term is not necessarily of order $O(n^{-2})$, as the numerator might be arbitrarily large. Note also that here as in the rest of the paper, we only prove results in the finite horizon setting, meaning that the number of samples is known in advance and the parameters γ, λ may be chosen as functions of n, but remain constant along the iterations (when λ or γ depend on n, our bounds only hold for the last iterate).

- Note that the bias term can also be bounded by $\frac{1}{\gamma n}\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1/2}(\theta_0 - \theta_*)\|^2$ only when $\|\theta_0 - \theta_*\|$ is finite (note the difference in the powers of n and $(\Sigma + \lambda I)^{-1}$). See the proof in Appendix A.2 for details.

- The second term in Eq. (8) is the variance term. It depends on the noise in the gradient. When this one is not structured the variance turns to be also bounded by $\gamma \operatorname{tr}\left(\Sigma(\Sigma + \lambda I)^{-1}\mathbb{E}[\xi_n \otimes \xi_n]\right)$ (see Appendix A.3) and we recover for $\gamma = O(1/\sqrt{n})$, the usual rate of $\frac{1}{\sqrt{n}}$ for SGD in the smooth case (Shalev-Shwartz et al., 2009).

- Overall we get the same performance as the empirical risk minimizer with fixed design, but with an algorithm that performs a single pass over the data.

- When $\lambda = 0$ we recover Lemma 2 of Bach and Moulines (2013). In this case the variance term $\frac{\tau^2 d}{n}$ is optimal over all estimators in $\mathcal{H}$ (Tsybakov, 2008) even without

computational limits, in the sense that no estimator that uses the same information can improve upon this rate.

3.2 Multiplicative/Additive Noise

When the general stochastic oracle in Eq. (6) is considered, the regularized LMS algorithm defined by Eq. (2) takes the form:

$$\theta_n = \big[I - \gamma x_n \otimes x_n - \gamma\lambda I\big]\theta_{n-1} + \gamma y_n x_n + \lambda\gamma\theta_0. \tag{9}$$

We have a very similar result with an additional corrective term (second line below) compared to Lemma 1.

Theorem 2 *Assume $(\mathcal{A}_{1,2})$. Consider the recursion in Eq. (9). For any regularization parameter $\lambda \in \mathbb{R}^+$ and for any constant step-size γ such that $2\gamma(R^2+2\lambda) \leq 1$ we have:*

$$\begin{aligned}\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \;\leqslant\; & 3\Big(2\lambda + \frac{1}{\gamma n}\Big)^2 \|\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2 + \frac{6\sigma^2}{n+1}\operatorname{tr}\big[\Sigma^2(\Sigma+\lambda I)^{-2}\big] \\ & + 3\frac{\big\|(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\big\|^2 \operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-1})}{\gamma^2(n+1)^2}.\end{aligned}$$

We can make the following remarks:

- The proof (see Appendix B) relies on a bias-variance decomposition, each term being treated separately. We adapt a proof technique from Bach and Moulines (2013) which considers the difference between the recursions in Eq. (9) and in Eq. (7).
- As in Lemma 1, the bias term can also be bounded by $\frac{1}{\gamma n}\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\|^2$ and the variance term by $\gamma\operatorname{tr}[\Sigma(\Sigma+\lambda I)^{-1}\xi_n\otimes\xi_n]$ (see proof in Appendices B.4 and B.5). This is useful in particular when considering unstructured noise.
- The variance term is the same as in the previous case. However there is a residual term that now appears when we go to the fully stochastic oracle (second line). This term will go to zero when γ tends to zero and can be compared to the corrective term which also appears when Hsu et al. (2014) go from fixed to random design. Nevertheless our bounds are more concise than theirs, making significantly fewer assumptions and relying on an efficient single-pass algorithm.
- In this setting, the step-size may not exceed $1/(2(R^2+2\lambda))$, whereas with an additive noise in Lemma 1 the condition is $\gamma \leq 1/(L+\lambda)$, a quantity which can be much bigger than $1/(2(R^2+2\lambda))$, as L is the spectral radius of Σ whereas R^2 is of the order of $\operatorname{tr}(\Sigma)$. Note that in practice, computing L is as hard as computing θ_* so that the step-size $\gamma \propto 1/R^2$ is a good practical choice. See Défossez and Bach (2015) for larger allowed step-sizes that require more information.
- For $\lambda = 0$ the error is bounded by $\frac{3(1+d)}{(\gamma n)^2}\|\Sigma^{-1/2}(\theta_0-\theta_*)\|^2 + \frac{6\sigma^2 d}{n+1}$. We recover results from Défossez and Bach (2015) with a non-asymptotic bound but we lose the advantage of having an asymptotic equivalent (i.e., a limit rather than an upper-bound). We note that the assumption $(\mathcal{A}_{1,2})$ are close to the minimal assumptions required to obtain the optimal rate of convergence of $\sigma^2 d/n$ (Lecué and Mendelson, 2016; Oliveira, 2016)

4. Accelerated Stochastic Averaged Gradient Descent

We study the convergence under the stochastic oracle from Eq. (5) of averaged *accelerated* stochastic gradient descent defined by Eq. (3) which can be rewritten for the least-squares function f as a second-order iterative system with constant coefficients:

$$\theta_n = \big[I - \gamma\Sigma - \gamma\lambda I\big]\big[\theta_{n-1} + \delta(\theta_{n-1} - \theta_{n-2})\big] + \gamma y_n x_n + \gamma\lambda\theta_0. \tag{10}$$

When using averaging, we refer to this algorithm as "averaged-accelerated-SGD".

Theorem 3 *Assume ($\mathcal{A}_4$). For any regularization parameter $\lambda \in \mathbb{R}_+$ and for any constant step-size $\gamma(\Sigma + \lambda I) \preccurlyeq I$, we have for any $\delta \in \big[\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1\big]$, for the recursion in Eq. (10):*

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leq 2\Big(\lambda + \frac{36}{\gamma(n+1)^2}\Big)\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1/2}(\theta_0 - \theta_*)\|^2 + 8\tau^2\frac{\mathrm{tr}\,\big[\Sigma^2(\Sigma + \lambda I)^{-2}\big]}{n+1}.$$

The numerical constants are partially artifacts of the proof (see Appendices C and E). Thanks to a wise use of tight inequalities, the bound is independent of δ and valid for all $\lambda \in \mathbb{R}_+$. This results in the simple following corollary for $\lambda = 0$, which corresponds to the particularly simple recursion (with averaging to obtain $\bar{\theta}_n$):

$$\theta_n = \big[I - \gamma\Sigma\big](2\theta_{n-1} - \theta_{n-2}) + \gamma y_n x_n. \tag{11}$$

Corollary 4 *Assume ($\mathcal{A}_4$). For any constant step-size $\gamma\Sigma \preccurlyeq I$, we have for $\delta = 1$,*

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leq 36\frac{\|\theta_0 - \theta_*\|^2}{\gamma(n+1)^2} + 8\frac{\tau^2 d}{n+1}. \tag{12}$$

We can make the following observations:

- The proof technique relies on direct moment computations in each eigensubspace obtained by O'Donoghue and Candès (2013) in the deterministic case. Indeed as Σ is a symmetric matrix, the space can be decomposed on an orthonormal eigenbasis of Σ, and the iterations are decoupled in such an eigenbasis. Although we only provide an upper-bound, this is in fact an equality plus other exponentially small terms as shown in the proof which relies on linear algebra, with difficulties arising from the fact that this second-order system can be expressed as a linear stochastic dynamical system with non-symmetric matrices. We only provide a result for additive noise.

- The first bound $\frac{1}{\gamma n^2}\|\theta_0 - \theta_*\|^2$ in Eq. (12) corresponds to the usual accelerated rate. It has been shown by Nesterov (2004) to be the optimal rate of convergence for optimizing a quadratic function with a first-order method that can access only to sequences of gradients when $n \leq d$. We recover by averaging an algorithm dedicated to strongly-convex function the traditional convergence rate for non-strongly convex functions. Even if it seems surprising, the algorithm works also for $\lambda = 0$ and $\delta = 1$ (see also simulations in Section 7).

- The second bound in Eq. (12) also matches the optimal statistical performance $\frac{\tau^2 d}{n}$ described in the observations following Lemma 1. Accordingly this algorithm achieves joint bias/variance optimality (when measured in terms of τ^2 and $\|\theta_0 - \theta_*\|^2$).

- We have the same rate of convergence for the bias when compared to the regular Nesterov acceleration without averaging studied by Flammarion and Bach (2015), which corresponds to choosing $\delta_n = 1 - 2/n$ for all n. However if the problem is μ-strongly convex, this latter was shown to also converge at the linear rate $O\big((1-\gamma\mu)^n\big)$ and thus is adaptive to hidden strong-convexity (since the algorithm does not need to know μ to run). This explains that it ends up converging faster for quadratic function since for large n the convergence at rate $1/n^2$ becomes slower than the one at rate $(1-\gamma\mu)^n$ even for very small μ. This is confirmed in our experiments in Section 7. Thanks to this adaptivity, we can also show using the same tools and considering its weighted average $\tilde{\theta}_n = \frac{2}{n(n+1)} \sum_{k=0}^n k\theta_k$ that the bias term of $\mathbb{E}f(\tilde{\theta}_n) - f(\theta_*)$ has a convergence rate of order $\big(\lambda^2 + \frac{1}{\gamma^2(n+1)^4}\big)\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2$ without any change in the variance term. This has to be compared to the bias of averaged SGD $\big(\lambda + \frac{1}{\gamma(n+1)^2}\big)\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2$ in Section 3 and may lead to faster convergence for the bias in presence of hidden strong-convexity.

- Overall, the bias term is improved whereas the variance term is not degraded and acceleration is thus robust to noise in the gradients. Thereby, while second-order finite difference methods for optimizing quadratic functions in the singular case, such as conjugate gradient (Polyak, 1987, Section 6.1) are notoriously highly sensitive to noise, we are able to propose a version which is robust to stochastic noise.

- Note that when there is no assumption on the covariance of the noise we still have the variance bounded by $\frac{\gamma n}{2} \operatorname{tr}\big[\Sigma(\Sigma+\lambda I)^{-1}V\big]$; setting $\gamma = 1/n^{3/2}$ and $\lambda = 0$ leads to the bound $\frac{\|\theta_0-\theta_*\|^2}{\sqrt{n}} + \frac{\operatorname{tr} V}{\sqrt{n}}$. We recover the usual rate for accelerated stochastic gradient in the non-strongly-convex case (Xiao, 2010). When the values of the bias and the variance are known, we can achieve the optimal trade-off of Lan (2012) $\frac{R^2\|\theta_0-\theta_*\|^2}{n^2} + \frac{\|\theta_0-\theta_*\|\sqrt{\operatorname{tr} V}}{\sqrt{n}}$ for $\gamma = \min\Big\{1/R^2, \frac{\|\theta_0-\theta_*\|}{\sqrt{\operatorname{tr} V} n^{3/2}}\Big\}$.

5. Tighter Dimension-Independent Convergence Rates

We have seen in Corollary 4 above that the averaged accelerated gradient algorithm matches the lower bounds $\tau^2 d/n$ and $\frac{L}{n^2}\|\theta_0 - \theta_*\|^2$ for the prediction error. However the algorithm performs better in almost all cases except the worst-case scenarios corresponding to the lower bounds. For example the algorithm may still predict well when the dimension d is much bigger than n. Similarly the norm of the optimal predictor $\|\theta_*\|^2$ may be huge and the prediction still good, as gradients algorithms happen to be adaptive to the difficulty of the problem: indeed, if the problem is simpler, the convergence rate of the gradient algorithm will be improved. In this section, we provide such a theoretical guarantee.

The following bound stands for the averaged *accelerated* algorithm. It extends previously known bounds in the kernel least-mean-squares setting (Dieuleveut and Bach, 2015).

Theorem 5 *Assume ($\mathcal{A}_4$); for any regularization parameter $\lambda \in \mathbb{R}_+$ and for any constant step-size such that $\gamma(\Sigma + \lambda I) \preccurlyeq I$ we have for $\delta \in \left[\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1\right]$, for the recursion in Eq. (10):*

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leq \min_{r\in[0,1],\ b\in[0,1]} \left[2\|\Sigma^{r/2}(\theta_0 - \theta_*)\|^2\ \lambda^{-r}\left(\frac{36}{\gamma(n+1)^2} + \lambda\right) + 8\frac{\tau^2 \operatorname{tr}(\Sigma^b)\lambda^{-b}}{n+1}\right].$$

The proof is straightforward by upper bounding the terms coming from regularization, depending on $\Sigma(\Sigma+\lambda I)^{-1}$, by a power of λ times the considered quantities. More precisely, the quantity $\operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-1})$ can be seen as an effective dimension of the problem (Gu, 2013), and is upper bounded by $\lambda^{-b}\operatorname{tr}(\Sigma^b)$ for any $b \in [0;1]$. Similarly, $\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}\theta_*\|^2$ can be upper bounded by $\lambda^{-r}\|\Sigma^{r/2}(\theta_0 - \theta_*)\|^2$. A detailed proof of these results is given in Appendix D.

In order to benefit from the acceleration, we choose $\lambda = (\gamma n^2)^{-1}$. With such a choice we have the following corollary:

Corollary 6 *Assume ($\mathcal{A}_4$), for any constant step-size $\gamma(\Sigma + \lambda I) \preccurlyeq I$, we have for $\lambda = \frac{1}{\gamma(n+1)^2}$ and $\delta \in \left[1 - \frac{2}{n+2}, 1\right]$, for the recursion in Eq. (10):*

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leq \min_{r\in[0,1],\ b\in[0,1]} \left[74\ \frac{\|\Sigma^{r/2}(\theta_0 - \theta_*)\|^2}{\gamma^{1-r}(n+1)^{2(1-r)}} + 8\frac{\tau^2\gamma^b \operatorname{tr}(\Sigma^b)}{(n+1)^{1-2b}}\right].$$

We can make the following observations:

- The algorithm is independent of r and b, thus all the bounds for different values of (r, b) are valid. This is a strong property of the algorithm, which is indeed adaptative to the regularity and the effective dimension of the problem (once γ is chosen). In situations in which either d is larger than n or $L\|\theta_0 - \theta_*\|^2$ is larger than n^2, the algorithm can still enjoy good convergence properties, by adapting to the best values of b and r.

- For $b = 0$ we recover the variance term of Corollary 4, but for $b > 0$ and fast decays of eigenvalues of Σ, the bound may be much smaller; note that we lose in the dependency in n, but typically, for large d, this can be advantageous.

- For $r = 0$ we recover the bias term of Corollary 4 and for $r = 1$ (no assumption at all) the bias is bounded by $\|\Sigma^{1/2}\theta_*\|^2 \leq 4R^2$, which is not going to zero. The smaller r is, the stronger the decrease of the bias with respect to n is (which is coherent with the fact that we have a stronger assumption). Moreover, r is only considered between 0 and 1: indeed, if $r < 0$, the constant$\|(\gamma\Sigma)^{r/2}(\theta_0 - \theta_*)\|$ is bigger than $\|\theta_0 - \theta_*\|$, but the dependence on n cannot improve beyond $(\gamma n^2)^{-1}$. This is a classical phenomenon called "saturation" (Engl et al., 1996). It is linked with the uniform averaging scheme: here, the bias term cannot forget the initial condition faster than n^{-2}.

- A similar result happens to hold, for averaged gradient descent, with $\lambda = (\gamma n)^{-1}$:

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leq \min_{\substack{r\in[-1,1]\\ b\in[0,1]}} \left[(18 + \operatorname{Res}(b, r, n, \gamma))\ \frac{\|\Sigma^{r/2}(\theta_0 - \theta_*)\|^2}{\gamma^{1-r}(n+1)^{(1-r)}} + 6\frac{\sigma^2\gamma^b \operatorname{tr}(\Sigma^b)}{(n+1)^{1-b}}\right], \quad (13)$$

where $\mathrm{Res}(b, r, n, \gamma))$ corresponds to a residual term, which is smaller than $\mathrm{tr}(\Sigma^b) n^b \gamma^{1+b}$ if $r \geq 0$ and does not exist otherwise. The bias term's dependence on n is degraded, thus the "saturation" limit is logically pushed down to $r = -1$, which explains the $[-1; 1]$ interval for r. The choice $\lambda = (\gamma n)^{-1}$ arises from Th. 2, in order to balance both components of the bias term $\lambda + (\gamma n)^{-1}$. This result is proved in Appendix D. This recovers the result of Dieuleveut and Bach (2015).

- Considering a non-uniform averaging, as proposed after Theorem 2 the $\min_{0 \leq r \leq 1}$ in Th. 5 and Corollary 6 can be extended to $\min_{-1 \leq r \leq 1}$. Indeed, considering a non-uniform averaging allows to have a faster decreasing bias, pushing the saturation limit observed below.

In finite dimension these bounds for the bias and the variance cannot be said to be optimal independently in any sense we are aware of. Indeed, in finite dimension, the asymptotic rate of convergence for the bias (respectively the variance), when n goes to ∞ is governed by $L\|\theta_0 - \theta_*\|^2/n^2$ (resp. $\tau^2 d/n$). However, we show in the next section that in the setting of non parametric learning in kernel spaces, these bounds lead to the optimal statistical rate of convergence among all estimators (independently of their computational cost). Moving to the infinite-dimensional setting allows to characterize the optimality of the bounds by showing that they achieve the statistical rate when optimizing the bias/variance tradeoff in Corollary 6.

6. Rates of Convergence for Kernel Regression

Computational convergence rates give the speed at which an objective function can decrease depending on the amount of computation which is allowed. Typically, they show how the error decreases with respect to the number of iterations, as in Theorem 2. Statistical rates, however, show how close one can get to some objective given some amount of information which is provided. Statistical rates do not depend on some chosen algorithm: these bounds do not involve computation, on the contrary, they state the best performance that no algorithm can beat, given the information, and without computational limits. In particular, any lower bound on the statistical rate implies a lower bound on the computational rates, if each iteration corresponds to access to some new information, here pairs of observations. Interestingly, many algorithms for the past few years have proved to match, with minimal computations (in general one pass through the data), the statistical rate, emphasizing the importance of carrying together optimization and approximation in large scale learning, as described by Bottou and Bousquet (2008). In a similar flavor, it also appears that regularization can be accomplished through early stopping (Yao et al., 2007; Rudi et al., 2015), highlighting this interplay between computation and statistics.

To characterize the optimality of our bounds, we will show that averaged-accelerated-SGD matches the statistical lower bound in the context of non-parametric estimation. Even if it may be computationally hard or impossible to implement averaged-accelerated-SGD with additive noise in the kernel-based framework below (see remarks following Theorem 8), it leads to the optimal statistical rate for a broader class of problems than averaged-SGD, showing that for a wider set of trade-offs, acceleration is optimal.

A natural extension of the finite-dimensional analysis is the non-parametric setting, especially with reproducing kernel Hilbert spaces. In the setting of non-parametric regression, we consider a probability space $\mathcal{X} \times \mathbb{R}$ with probability distribution ρ, and assume that we are given an i.i.d. sample $(x_i, y_i)_{i=1,\dots,n} \sim \rho^{\otimes n}$, and denote by ρ_X the marginal distribution of x_n in $\mathcal{X}$; the aim of non-parametric least-squares regression is to find a function $g : \mathcal{X} \to \mathbb{R}$, which minimizes the expected risk:

$$f(g) = \frac{1}{2}\mathbb{E}_\rho[(g(x_n) - y_n)^2]. \tag{14}$$

The optimal function g is the conditional expectation $g(x) = \mathbb{E}_\rho(y_n|x)$. In the kernel regression setting, we consider as hypothesis space a reproducing kernel Hilbert space (Aronszajn, 1950; Steinwart and Christmann, 2008; Schölkopf and Smola, 2002) associated with a kernel function K. The space $\mathcal{H}$ is a subspace of the space of squared integrable functions $L^2_{\rho_X}$. We look for a function $g_{\mathcal{H}}$ which satisfies: $f(g_{\mathcal{H}}) = \inf_{g\in\mathcal{H}} f(g)$, and $g_{\mathcal{H}}$ belongs to the closure $\bar{\mathcal{H}}$ of $\mathcal{H}$ (meaning that there exists a sequence of function $g_n \in \mathcal{H}$ such that $\|g_n - g_H\|_{L^2_{\rho_X}} \to 0$). When $\mathcal{H}$ is dense, the minimum is attained for the regression function defined above. This function however *is not* in $\mathcal{H}$ in general. Moreover there exists an operator $\Sigma : \mathcal{H} \to \mathcal{H}$, which extends the finite-dimensional population covariance matrix, that will allow the characterization of the smoothness of $g_{\mathcal{H}}$. This operator is known to be trace class when $\mathbb{E}_{\rho_X}[K(x_n, x_n)] < \infty$.

Data points x_i are mapped into the RKHS, via the feature map: $x \mapsto K_x$, where $K_x : \mathcal{H} \to \mathbb{R}$ is a function in the RKHS, such that $K_x : y \mapsto K(x, y)$. The reproducing property[3] allows to express the minimization problem (14) as a least-squares linear regression problem: for any $g \in \mathcal{H}$, $f(g) = \frac{1}{2}\mathbb{E}_\rho[(\langle g, K_{x_n}\rangle_{\mathcal{H}} - y_n)^2]$, and can thus be seen as an extension to the infinite-dimensional setting of linear least-squares regression.

However, in such a setting, both quantities $\|\Sigma^{r/2}\theta_*\|_{\mathcal{H}}$ (where $\|\cdot\|_{\mathcal{H}}$ stands for the norm associated with the inner product in the Hilbert space $\mathcal{H}$) and $\mathrm{tr}(\Sigma^b)$ *may exist or not*. It thus arises as a natural *assumption* to consider the smaller $r \in [-1; 1]$ and the smaller $b \in [0; 1]$ such that

- $\|\Sigma^{r/2}\theta_*\|_{\mathcal{H}} < \infty$ (meaning that $\Sigma^{r/2}\theta_* \in \mathcal{H}$), $(\mathcal{A}_5)$
- $\mathrm{tr}(\Sigma^b) < \infty$. $(\mathcal{A}_6)$

The quantities considered in Sections 2 and 5 are the natural finite-dimensional twins of these assumptions. However in infinite dimension a quantity may exist or not and it is thus an assumption to consider its existence, whereas it can only be characterized by its value, big or small, in finite dimension. The first assumption is generally called the "source condition", the second one the "capacity condition".

In the last decade, De Vito et al. (2005); Cucker and Smale (2002) studied non-parametric least-squares regression in the RKHS framework. These works were extended to derive rates of convergence depending on assumption $(\mathcal{A}_5)$: Ying and Pontil (2008) studied unregularized stochastic gradient descent and derived asymptotic rate of convergence $O(n^{-\frac{1-r}{2-r}})$, for $r \leq 1$ and proved that one could derive similar rates of convergence for $0 \leq r \leq 1$ from

3. It states that for any function $g \in \mathcal{H}$, $\langle g, K_x\rangle_{\mathcal{H}} = g(x)$, where $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ denotes the scalar product in the Hilbert space.

Zhang (2004), who studies stochastic gradient descent with averaging; whereas Tarrès and Yao (2011) give similar performance for $-1 \leq r \leq 0$. Interestingly, Ying and Pontil (2008) do not have saturation, meaning that the rate still improves for r smaller than -1. As it will appear, any algorithm based on a uniform averaging scheme faces a saturation issue : one cannot forget initial conditions faster than n^{-2}, which makes the algorithm sub-optimal in situations in which the optimal predictor is very smooth ($\mathcal{A}_5$ holds with $r \leq -1$). However, these papers only prove rates in the capacity-independent setting, meaning without assumption on the spectrum of the covariance matrix. Although the rate $O(n^{-\frac{1-r}{2-r}})$ is optimal in this setting, it comes from a worst-case analysis. Considering the capacity-dependent setting is more challenging, but allows to derive tighter and more realistic rates (a capacity condition always stands under the trace class assumption that is made). Moreover, the capacity-independent setting also does not allow to recover finite-dimensional rates. Up to our knowledge, there is no one pass stochastic gradient algorithm which does not have saturation while getting the minimax rate under both the capacity condition and source condition. In a recent work, Lin and Rosasco (2016) achieves optimality without saturation with multiple passes. We show in the next paragraphs that we can derive a tighter and optimal rate for both averaged-SGD (recovering results from Dieuleveut and Bach (2015)) and averaged-accelerated-SGD, for a larger class of kernels for the latter. Note that the averaging scheme for the RKHS setting was originally considered by Yao (2006).

We will first describe results for averaged-SGD, then increase the validity region of these rates (which depends on r, b) using averaged accelerated SGD. We show that the derived rates match statistical rates for our setting and thus our algorithms reach the optimal prediction performance for certain b and r.

6.1 Averaged SGD

We have the following result, proved in Appendix D and following from Theorem 2: for some fixed b, r, we choose the best step-size γ, that optimizes the bias-variance trade-off, while still satisfying the constraint $\gamma \leq 1/(2R^2)$. We get a result for the stochastic oracle (multiplicative/additive noise).

Theorem 7 *With $\lambda = \frac{1}{\gamma n}$, we have, if $r \leq b$, under Assumptions ($\mathcal{A}_{1,2,5,6}$) and the stochastic oracle Eq. (6), for any constant step-size γ such that $2\gamma(R^2 + 2\lambda) \leq 1$, with $\gamma \propto n^{\frac{-b+r}{b+1-r}}$, for the recursion in Eq. (9):*

$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \quad \leq \quad \left((27 + o(1)) \left\| \Sigma^{r/2} (\theta_0 - \theta_*) \right\|^2 + 6\sigma^2 \operatorname{tr}(\Sigma^b) \right) n^{-\frac{1-r}{b+1-r}}.$$

We can make the following remarks:

- The term $o(1)$ stands for a quantity which is decreasing to 0 when $n \to \infty$. More specifically, this constant is smaller than $3 \operatorname{tr}(\Sigma^b)$ divided by n^χ, where χ is bigger than 0 (see Appendix D). The result comes from Eq. (13) (which follows from Theorem 5), with the choice of the optimal step-size.
- We recover the same errors bounds as in Dieuleveut and Bach (2015), but with a simpler analysis resulting from the consideration of the regularized version of the

problem associated with a choice of λ. However, we only recover rates in the finite horizon setting.

- This result shows that we get the optimal rate of convergence under Assumptions $(\mathcal{A}_{5,6})$, for $r \leq b$. This point will be discussed in more details after Theorem 8.

We now turn to the averaged accelerated SGD algorithm. We prove that it enjoys the optimal rate of convergence for a larger class of problems, but only for the additive noise which corresponds to knowing the distribution of x_n.

6.2 Averaged-Accelerated SGD

Similarly, choosing the best step-size γ, it comes from Theorem 5, that in the RKHS setting, under additional Assumptions $(\mathcal{A}_{5,6})$, we have for the the averaged accelerated algorithm the following result:

Theorem 8 *With $\lambda = \frac{1}{\gamma n^2}$, we have, if $r \leq b + 1/2$, under Assumptions $(\mathcal{A}_{4,5,6})$, for any constant step-size $\gamma \leq \frac{1}{L+\lambda}$, with $\gamma \propto n^{\frac{-2b+2r-1}{b+1-r}}$, for the recursion in Eq. (10):*

$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \quad \leq \quad \left(74 \left\| \Sigma^{r/2} (\theta_0 - \theta_*) \right\|^2 + 8\tau^2 \operatorname{tr}(\Sigma^b) \right) n^{-\frac{1-r}{b+1-r}}.$$

We can make the following remarks:

- The rate $\frac{1-r}{b+1-r}$ is always between 0 and 1, and improves when our assumptions gets stronger (r getting smaller, b getting smaller). Ultimately, with $b \to 0$, and $r \to -1$, we recover the finite-dimensional n^{-1} rate.
- We can achieve this optimal rate when $r \leq b + 1/2$. Beyond, if $r > b + 1/2$, the rate is only $n^{-2(1-r)}$. Indeed, the bias term cannot decrease faster than $n^{-2(1-r)}$, as γ is compelled to be upper bounded.
- The same phenomenon appears in the un-accelerated averaged situation, as shown by Theorem 7, but the critical value was then $r \leq b$. There is thus a region (precisely $b < r \leq b + 1/2$) in which only the accelerated algorithm gets the optimal rate of convergence. Note that we increase the optimality region towards optimization problems which are more ill-conditioned, naturally benefiting from acceleration.
- This algorithm cannot be computed in practice (at least with computational limits). Indeed, without any further assumption on the kernel K, it is not possible to compute images of vectors by the covariance operator Σ in the RKHS. However, as explained in the following remark, this is enough to show some form of optimality of our algorithm.

  Note that the easy computability is a great advantage of the multiplicative/additive noise variant of the algorithms, for which the current point θ_n can always be expressed as a finite sum of features $\theta_n = \sum_{i=1}^n \alpha_i K_{x_i}$, with $\alpha_i \in \mathbb{R}$, leading to a tractable algorithm. An accelerated variant of SGD naturally arises from our algorithm, when considering this stochastic oracle from Eq. (6). Such a variant can be implemented but does not behave similarly for large step sizes, say, $\gamma \simeq 1/(2R^2)$. It is an open problem to prove convergence results for averaged accelerated gradient under this multiplicative/additive noise.

- These rates happen to be optimal from a statistical perspective, meaning that no algorithm which is given access to the sample points and the distribution of x_n can perform better for all functions that satisfy assumption $(\mathcal{A}_6)$, for a kernel satisfying $(\mathcal{A}_5)$. Indeed it is equivalent to assuming that the function lives in some ellipsoid in the space of squared integrable functions. Note that the statistical minimization problem (and thus the lower bound) does not depend on the kernel, and is valid without computational limits. The case of learning with kernels is studied by Caponnetto and De Vito (2007) which shows these minimax convergence rates under $(\mathcal{A}_{5,6})$, under assumption that $-1 \leq r \leq 0$ (but state that it can be easily extended to $0 \leq r \leq 1$). They do not assume knowledge of the distribution of the inputs; however, Massart (2007) and Tsybakov (2008) discuss optimal rates on ellipsoids, and Györfi et al. (2006) proves similar results for certain class of functions under a known distribution for the input data, showing that the knowledge of the distribution does not make any difference. This minimax statistical rate stands without computational limits and is thus valid for both algorithms (additive noise that corresponds to knowing Σ, and multiplicative/additive noise). The optimal tradeoff is derived for an extended region of b, r (namely $r \leq b + 1/2$ instead of $r \leq b$) in the accelerated case which shows the improvement upon non-accelerated averaged SGD.

- The choice of the optimal γ is difficult in practice, as the parameters b, r are unknown, and this remains an open problem in general (see, e.g., Birgé, 2001, for some methods for non-parametric regression), even if in the capacity-independent setting, Orabona (2014) has proposed an algorithm that adapts to the unknown parameter r.

- Note that we do not give rates in terms of norm in the RHKS (i.e., an upper bound on $\|\bar{\theta}_n - \theta_*\|_{\mathcal{H}}$), because we mainly aim at extending optimality of prediction error rate to ill-conditioned cases (i.e. situations for which $r \geq b \geq 0$). In such a situation, Hilbert spaces norm bounds would not be relevant as the optimal estimator does not even live in the RKHS.

7. Experiments

We illustrate now our theoretical results on synthetic examples. For $d = 25$ we consider normally distributed inputs x_n with random covariance matrix Σ which has eigenvalues $1/i^3$, for $i = 1, \dots, d$, and random optimum θ_* and starting point θ_0 such that $\|\theta_0 - \theta_*\| = 1$. The outputs y_n are generated from a linear function with homoscedastic noise with unit signal to noise-ratio ($\sigma^2 = 1$), we take $R^2 = \operatorname{tr} \Sigma$ the average radius of the data and a step-size $\gamma = 1/R^2$ and $\lambda = 0$. The additive noise oracle is used. We show results averaged over 10 replications.

We compare the performance of averaged SGD (AvSGD), AccSGD (usual Nesterov acceleration for convex functions) and our novel averaged accelerated SGD from Section 4 AvAccSGD (which is not the averaging of AccSGD because the momentum term is proportional to $1 - 3/n$ for AccSGD instead of being equal to 1 for AvAccSGD), on two different problems: one deterministic ($\|\theta_0 - \theta_*\| = 1$, $\sigma^2 = 0$) which will illustrate how the bias term behaves, and one purely stochastic ($\|\theta_0 - \theta_*\| = 0$, $\sigma^2 = 1$) which will illustrate how the variance term behaves. For the bias (left plot of Figure 1), AvSGD converges at

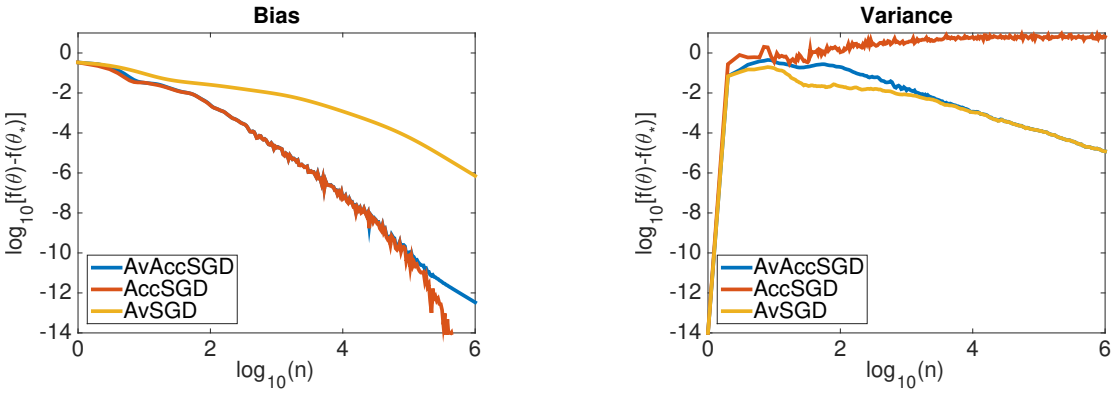

Figure 1: Synthetic problem ($d = 25$) and $\gamma = 1/R^2$. Left: Bias. Right: Variance.

speed $O(1/n)$, while AvAccSGD and AccSGD converge both at speed $O(1/n^2)$. However, as mentioned in the observations following Corollary 4, AccSGD takes advantage of the hidden strong convexity of the least-squares function and starts converging linearly at the end. For the variance (right plot of Figure 1), AccSGD is not converging to the optimum and keeps oscillating whereas AvSGD and AvAccSGD both converge to the optimum at a speed $O(1/n)$. However AvSGD remains slightly faster in the beginning.

Note that for small n, or when the bias $L\|\theta_0 - \theta_*\|^2/n^2$ is much bigger than the variance $\sigma^2 d/n$, the bias may have a stronger effect, although asymptotically, the variance always dominates. It is thus essential to have an algorithm which is optimal in both regimes; this is achieved by AvAccSGD.

8. Conclusion

In this paper, we showed that stochastic *averaged* accelerated gradient descent was robust to structured noise in the gradients present in least-squares regression. Beyond being the first algorithm which is jointly optimal in terms of both bias and finite-dimensional variance, it is also adapted to finer assumptions such as fast decays of the covariance matrices or optimal predictors with large norms.

Our current analysis is performed for least-squares regression. While it could be directly extended to smooth losses through efficient online Newton methods (Bach and Moulines, 2013), an extension to all smooth or self-concordant-like functions (Bach, 2014) would widen its applicability. Moreover, our accelerated gradient analysis is performed for additive noise (i.e., for least-squares regression, with knowledge of the population covariance matrix) and it would be interesting to study the robustness of our results in the contexts of least-mean squares and online learning. Finally, our analysis relies on single observations per iteration and could be made finer by using mini-batches (Cotter et al., 2011; Dekel et al., 2012), which should not change the variance term but could impact the bias term.

Acknowledgments

The authors would like to thank Damien Garreau for interesting discussions and the reviewers for their constructive and helpful comments.

Appendix A. Proofs of Section 3

We first prove the results of Section 3.

A.1 Proof of Lemma 1

We proof here Lemma 1 which is the extension of Lemma 2 of Bach and Moulines (2013) for the regularized case. The proof technique relies on the fact that recursions in Eq. (7) are linear since the cost function is quadratic which allows us to obtain $\theta_n - \theta_*$ in closed form.

For any regularization parameter $\lambda \in \mathbb{R}_+$ and any constant step-size $\gamma(\Sigma + \lambda I) \preccurlyeq I$ we may rewrite the regularized stochastic gradient recursion in Eq. (7) as:

$$\theta_n - \theta_* = \big[I - \gamma\Sigma - \gamma\lambda I\big](\theta_{n-1} - \theta_*) + \gamma\xi_n + \lambda\gamma(\theta_0 - \theta_*).$$

We thus get for $n \geq 1$ the expansion

$$\begin{aligned}
\theta_n - \theta_* &= (I - \gamma\Sigma - \gamma\lambda I)^n(\theta_0 - \theta_*) + \gamma\sum_{k=1}^{n}(I - \gamma\Sigma - \gamma\lambda I)^{n-k}\xi_k \\
&\quad + \gamma\lambda\sum_{k=1}^{n}(I - \gamma\Sigma - \gamma\lambda I)^{n-k}(\theta_0 - \theta_*) \\
&= (I - \gamma\Sigma - \gamma\lambda I)^n(\theta_0 - \theta_*) + \gamma\sum_{k=1}^{n}(I - \gamma\Sigma - \gamma\lambda I)^{n-k}\xi_k \\
&\quad + \lambda\big[I - (I - \gamma\Sigma - \gamma\lambda I)^n\big](\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*) \\
&= (I - \gamma\Sigma - \gamma\lambda I)^n[I - \lambda(\Sigma + \lambda I)^{-1}](\theta_0 - \theta_*) + \gamma\sum_{k=1}^{n}(I - \gamma\Sigma - \gamma\lambda I)^{n-k}\xi_k \\
&\quad + \lambda(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*).
\end{aligned}$$

We then have using the definition of the average

$$\begin{aligned}
n(\bar{\theta}_{n-1} - \theta_*) &= \sum_{j=0}^{n-1}(\theta_j - \theta_*) \\
&= \sum_{j=0}^{n-1}(I - \gamma\Sigma - \gamma\lambda I)^j[I - \lambda(\Sigma + \lambda I)^{-1}](\theta_0 - \theta_*) + \gamma\sum_{j=0}^{n-1}\sum_{k=1}^{j}(I - \gamma\Sigma - \gamma\lambda I)^{n-k}\xi_k \\
&\quad + n\lambda(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*).
\end{aligned}$$

For which we will compute the two sums separately

$$\sum_{j=0}^{n-1}(I-\gamma\Sigma-\gamma\lambda I)^j[I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*)$$
$$=\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big](\Sigma+\lambda I)^{-1}[I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*),$$

and

$$\begin{aligned}\gamma\sum_{j=0}^{n-1}\sum_{k=1}^{j}(I-\gamma\Sigma-\gamma\lambda I)^{j-k}\xi_k &= \gamma\sum_{k=1}^{n-1}\Big(\sum_{j=k}^{n-1}(I-\gamma\Sigma-\gamma\lambda I)^{j-k}\Big)\xi_k\\ &= \gamma\sum_{k=1}^{n-1}\Big(\sum_{j=0}^{n-1-k}(I-\gamma\Sigma-\gamma\lambda I)^{j}\Big)\xi_k\\ &= \sum_{k=1}^{n-1}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big](\Sigma+\lambda I)^{-1}\xi_k.\end{aligned}$$

Gathering the three terms together, we thus have

$$\begin{aligned}n(\bar\theta_{n-1}-\theta_*) &= \frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big](\Sigma+\lambda I)^{-1}[I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*)\\ &\quad+\sum_{k=1}^{n-1}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big](\Sigma+\lambda I)^{-1}\xi_k+n\lambda(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\\ &= \Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big][I-\lambda(\Sigma+\lambda I)^{-1}]+n\lambda I\Big](\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\\ &\quad+\sum_{k=1}^{n-1}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big](\Sigma+\lambda I)^{-1}\xi_k.\end{aligned}$$

Using standard martingale square moment inequalities which amount to consider ξ_i, $i=1,\dots,n$ independent, the variance of the sum is the sum of variances and we have for $V=\mathbb{E}\xi_n\otimes\xi_n$

$$\begin{aligned}&n^2\mathbb{E}\|\Sigma^{1/2}(\bar\theta_{n-1}-\theta_*)\|^2=\textstyle\sum_{k=1}^{n-1}\operatorname{tr}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big]^2\Sigma(\Sigma+\lambda I)^{-2}V\\ &+\Big\|\Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big][I-\lambda(\Sigma+\lambda I)^{-1}]+n\lambda I\Big]\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\Big\|^2. \quad (15)\end{aligned}$$

Since all the matrices in this equality are symmetric positive-definite we are allowed to bound

$$\begin{aligned}\Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big][I-\lambda(\Sigma+\lambda I)^{-1}]+n\lambda I\Big] &\preccurlyeq \Big(\frac{1}{\gamma}+n\lambda\Big)I \qquad (16)\\ \big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big]^2 &\preccurlyeq I.\end{aligned}$$

This concludes proof of the Lemma 1

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_{n-1}-\theta_*)\|^2 \leq \Big(\frac{1}{n\gamma}+\lambda\Big)^2\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2 + \frac{1}{n}\operatorname{tr}\Sigma(\Sigma+\lambda I)^{-2}V. \quad (17)$$

A.2 Proof When Only $\|\theta_0-\theta_*\|$ Is Finite

Unfortunately $\|\Sigma^{-1}(\theta_0-\theta_*)\|$ may not be finite. However we can use that for all $u\in[0,1]$ we have $\frac{1-(1-u)^n}{nu}\leq 1$[4] and have therefore the bound

$$\begin{aligned}
&\Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big][I-\lambda(\Sigma+\lambda I)^{-1}]+n\lambda I\Big][\Sigma+\lambda I]^{-1}\\
&\preccurlyeq \Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big]+n\lambda I\Big][\Sigma+\lambda I]^{-1}\\
&\preccurlyeq \Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big][\Sigma+\lambda I]^{-1}+n\lambda[\Sigma+\lambda I]^{-1}\Big]\\
&\preccurlyeq I+nI.
\end{aligned}$$

Combining with Eq. (16) we have

$$\Big\|\Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big][I-\lambda(\Sigma+\lambda I)^{-1}]+n\lambda I\Big]\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\Big\|^2 \leq (n+1)\Big(\frac{1}{\gamma}+n\lambda\Big)\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\|^2,$$

which implies that

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_{n-1}-\theta_*)\|^2 \leq 2\Big(\frac{1}{n\gamma}+\lambda\Big)\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\|^2 + \frac{1}{n}\operatorname{tr}\Sigma(\Sigma+\lambda I)^{-2}V, \quad (18)$$

which is interesting when only $\|\theta_0-\theta_*\|$ is finite.

A.3 Proof When the Noise Is Not Structured

The bound in Eq. (17) becomes less interesting when the noise is not structured. However using the same technique we have that $\big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big]^2(\Sigma+\lambda I)^{-1}\preccurlyeq(n-k)\gamma I$ and we get the following upper-bound on the variance

$$\begin{aligned}
\sum_{k=1}^n \operatorname{tr}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big]^2\Sigma(\Sigma+\lambda I)^{-2}V &\leq \gamma\sum_{k=1}^n(n-k)\operatorname{tr}\Sigma(\Sigma+\lambda I)^{-1}V\\
&\leq \gamma\frac{n(n+1)}{2}\operatorname{tr}\Sigma(\Sigma+\lambda I)^{-1}V.
\end{aligned}$$

4. since $\frac{1-(1-u)^n}{u}=\sum_{k=0}^n(1-u)^k\leq n$

Therefore we get

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_{n-1}-\theta_*)\|^2 \leq \Big(\frac{1}{n\gamma}+\lambda\Big)^2\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2 + \gamma\,\mathrm{tr}\,\Sigma(\Sigma+\lambda I)^{-1}V, \quad (19)$$

which is meaningful when the noise is not structured.

Appendix B. Proof of Theorem 2

In this section, we will prove Theorem 2. The proof relies on a decomposition of the error as the sum of three main terms which will be studied separately. We state decomposition in Section B.1 then prove upper bounds for the different terms in Sections B.2 and B.3.

B.1 Expansion of the Recursion

We may rewrite the regularized stochastic gradient recursion as:

$$\begin{aligned}\theta_n &= \big[I-\gamma x_n\otimes x_n-\gamma\lambda I\big]\theta_{n-1}+\gamma\varepsilon_n x_n+\gamma\langle x_n,\theta_*\rangle x_n+\lambda\gamma\theta_0\\ \theta_n-\theta_* &= \big[I-\gamma x_n\otimes x_n-\gamma\lambda I\big](\theta_{n-1}-\theta_*)+\gamma\varepsilon_n x_n+\lambda\gamma(\theta_0-\theta_*).\end{aligned}$$

For $i\geqslant k$, let

$$M(i,k)=\big[I-\gamma x_i\otimes x_i-\gamma\lambda I\big]\cdots\big[I-\gamma x_k\otimes x_k-\gamma\lambda I\big]$$

be an operator from $\mathcal{H}$ to $\mathcal{H}$. We have the expansion

$$\theta_n-\theta_*=M(n,1)(\theta_0-\theta_*)+\gamma\sum_{k=1}^n M(n,k+1)\varepsilon_k x_k+\gamma\sum_{k=1}^n M(n,k+1)\lambda(\theta_0-\theta_*).$$

Our goal is to study these three terms separately and bound $\|\Sigma^{1/2}(\bar{\theta}_n-\theta_*)\|$ for each of them.

B.2 Regularization-Based Bias Term

This is the term: $\theta_n-\theta_*=\gamma\sum_{k=1}^n M(n,k+1)\lambda(\theta_0-\theta_*)$, which corresponds to the recursion

$$\theta_n-\theta_*=\big(I-\gamma x_n\otimes x_n-\gamma\lambda I\big)(\theta_{n-1}-\theta_*)+\lambda\gamma(\theta_0-\theta_*), \quad (20)$$

initialized with $\theta_0=\theta_*$, and no noise.

Following the proof technique of Bach and Moulines (2013), we are going to consider a related recursion by replacing in Equation (20) the operator $x_n\otimes x_n$ by its expectation Σ. Thus, we consider η_n defined as

$$\eta_n-\theta_*=\gamma\sum_{k=1}^n(I-\gamma\Sigma-\lambda\gamma I)^{n-k}\lambda(\theta_0-\theta_*),$$

which satisfies the recursion (with initialization $\eta_0=\theta_*$) and

$$\eta_n-\theta_*=\big[I-\gamma\Sigma-\lambda\gamma I\big](\eta_{n-1}-\theta_*)+\lambda\gamma(\theta_0-\theta_*).$$

In order to bound $\|\Sigma^{1/2}(\theta_n-\theta_*)\|$, we will independently bound $\|\Sigma^{1/2}(\eta_n-\theta_*)\|$ and $\|\Sigma^{1/2}(\theta_n-\eta_n)\|$ using Minkowski's inequality.

Bounding $\|\Sigma^{1/2}(\theta_n - \eta_n)\|$. We have $\theta_0 - \eta_0 = 0$, and

$$\theta_n - \eta_n = \big[I - \gamma x_n \otimes x_n - \lambda\gamma I\big](\theta_{n-1} - \eta_{n-1}) + \gamma\big[\Sigma - x_n \otimes x_n\big](\eta_{n-1} - \theta_*).$$

We can now bound the recursion for $\theta_n - \eta_n$ as follows, using standard online learning proofs (Nemirovski et al., 2009):

$$\begin{aligned}\|\theta_n - \eta_n\|^2 \leqslant\ & \|\theta_{n-1} - \eta_{n-1}\|^2 - 2\gamma\big\langle\theta_{n-1} - \eta_{n-1}, (x_n \otimes x_n + \lambda I)(\theta_{n-1} - \eta_{n-1})\big\rangle \\ & + 2\gamma\big\langle\theta_{n-1} - \eta_{n-1}, \big[\Sigma - x_n \otimes x_n\big](\eta_{n-1} - \theta_*)\big\rangle \\ & + \gamma^2\big\|\big[x_n \otimes x_n + \lambda I\big](\theta_{n-1} - \eta_{n-1}) - \big[\Sigma - x_n \otimes x_n\big](\eta_{n-1} - \theta_*)\big\|^2.\end{aligned}$$

By taking conditional expectations given $\mathcal{F}_{n-1}$, we get, using first the fact that $\mathbb{E}(\Sigma - x_n \otimes x_n|\mathcal{F}_{n-1}) = 0$ and the inequality $(a+b)^2 \leqslant 2(a^2+b^2)$, then developing and using $\mathbb{E}[(x_n \otimes x_n)^2] \leqslant R^2\Sigma$, which is assumption $\mathcal{A}_1$.

$$\begin{aligned}\mathbb{E}\big(\|\theta_n - \eta_n\|^2|\mathcal{F}_{n-1}\big) \leqslant\ & \|\theta_{n-1} - \eta_{n-1}\|^2 - 2\gamma\big\langle\theta_{n-1} - \eta_{n-1}, (\Sigma + \lambda I)(\theta_{n-1} - \eta_{n-1})\big\rangle \\ & + 2\gamma^2\mathbb{E}\big(\big\|\big[x_n \otimes x_n + \lambda I\big](\theta_{n-1} - \eta_{n-1})\big\|^2|\mathcal{F}_{n-1}\big) \\ & + 2\gamma^2\mathbb{E}\big(\big\|\big[\Sigma - x_n \otimes x_n\big](\eta_{n-1} - \theta_*)\big\|^2|\mathcal{F}_{n-1}\big) \\ \leqslant\ & \|\theta_{n-1} - \eta_{n-1}\|^2 - 2\gamma\big\langle\theta_{n-1} - \eta_{n-1}, (\Sigma + \lambda I)(\theta_{n-1} - \eta_{n-1})\big\rangle \\ & + 2\gamma^2\big\langle\theta_{n-1} - \eta_{n-1}, (R^2\Sigma + \lambda^2 I + 2\lambda\Sigma)(\theta_{n-1} - \eta_{n-1})\big\rangle \\ & + 2\gamma^2R^2\langle\eta_{n-1} - \theta_*, \Sigma\rangle \\ \leqslant\ & \|\theta_{n-1} - \eta_{n-1}\|^2 - 2\gamma\big[1 - \gamma(R^2 + 2\lambda)\big]\big\langle\theta_{n-1} - \eta_{n-1}, \Sigma(\theta_{n-1} - \eta_{n-1})\big\rangle \\ & + 2\gamma^2R^2\langle\eta_{n-1} - \theta_*, \Sigma(\eta_{n-1} - \theta_*)\rangle.\end{aligned}$$

This leads by taking full expectations and moving terms to

$$\begin{aligned}\mathbb{E}\big\langle\theta_{n-1} - \eta_{n-1}, \Sigma(\theta_{n-1} - \eta_{n-1})\big\rangle \leqslant\ & \frac{1}{2\gamma\big[1 - \gamma(R^2 + 2\lambda)\big]}\big[\mathbb{E}\|\theta_{n-1} - \eta_{n-1}\|^2 - \mathbb{E}\|\theta_n - \eta_n\|^2\big] \\ & + \frac{\gamma R^2}{1 - \gamma(R^2 + 2\lambda)}\langle\eta_{n-1} - \theta_*, \Sigma(\eta_{n-1} - \theta_*)\rangle.\end{aligned}$$

Thus, if $\gamma(R^2 + 2\lambda) \leqslant \frac{1}{2}$

$$\begin{aligned}\mathbb{E}\big\langle\theta_{n-1} - \eta_{n-1}, \Sigma(\theta_{n-1} - \eta_{n-1})\big\rangle \leqslant\ & \frac{1}{\gamma}\big[\mathbb{E}\|\theta_{n-1} - \eta_{n-1}\|^2 - \mathbb{E}\|\theta_n - \eta_n\|^2\big] \\ & + 2\gamma R^2\mathbb{E}\langle\eta_{n-1} - \theta_*, \Sigma(\eta_{n-1} - \theta_*)\rangle.\end{aligned}$$

This leads to, summing and using initial conditions $\theta_0 - \eta_0 = 0$, then using convexity to upper bound $\big\langle\bar\theta_n - \bar\eta_n, \Sigma(\bar\theta_n - \bar\eta_n)\big\rangle \leq \frac{1}{n+1}\sum_{k=0}^n\big\langle\theta_k - \eta_k, \Sigma(\theta_k - \eta_k)\big\rangle$,

$$\mathbb{E}\big\langle\bar\theta_n - \bar\eta_n, \Sigma(\bar\theta_n - \bar\eta_n)\big\rangle \leqslant \frac{2\gamma R^2}{n+1}\sum_{k=0}^n\langle\eta_k - \theta_*, \Sigma(\eta_k - \theta_*)\rangle.$$

Bounding $\|\Sigma^{1/2}(\eta_n - \theta_*)\|$. Moreover we have:

$$\begin{aligned}
\eta_n - \theta_* &= \lambda(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*) - (I - \gamma\Sigma - \lambda\gamma I)^n\big[\lambda(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\big] \\
\bar{\eta}_n - \theta_* &= \lambda(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*) - \frac{1}{n+1}\sum_{k=0}^{n}(I - \gamma\Sigma - \lambda\gamma I)^k\big[\lambda(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\big] \\
&= \lambda(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*) \\
&\qquad - \frac{1}{n+1}\gamma^{-1}(\Sigma + \lambda I)^{-1}\big[I - (I - \gamma\Sigma - \lambda\gamma I)^{n+1}\big]\big[\lambda(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\big].
\end{aligned}$$

This leads using Minkowski inequality to

$$\begin{aligned}
\big(\mathbb{E}\|\Sigma^{1/2}(\eta_n - \theta_*)\|^2\big)^{1/2} &\leqslant \|\lambda\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\| \\
\big(\mathbb{E}\|\Sigma^{1/2}(\bar{\eta}_n - \theta_*)\|^2\big)^{1/2} &\leqslant \|\lambda\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|.
\end{aligned}$$

Thus this part is such that

$$\begin{aligned}
\big(\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|^2\big)^{1/2} &\leqslant \|\lambda\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\| \\
&\qquad + \left(2\gamma R^2\|\lambda\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2\right)^{1/2} \\
&\leqslant \|\lambda\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|\big(1 + \sqrt{2\gamma R^2}\big),
\end{aligned}$$

that gives the first bound on the regularization-based bias

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|^2 \leqslant \|\lambda\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2\big(1 + \sqrt{2\gamma R^2}\big)^2. \tag{21}$$

B.3 Expansion without the Regularization Term

We will follow here the outline of the proof of Györfi and Walk (1996) which considers a full expansion of the function value $\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|^2$. This corresponds to

$$\theta_n - \theta_* = M(n,1)(\theta_0 - \theta_*) + \gamma\sum_{k=1}^{n} M(n,k+1)\varepsilon_k x_k.$$

We have

$$\mathbb{E}\sum_{i=0}^{n}\sum_{j=0}^{n}\langle\theta_i - \theta_*, \Sigma(\theta_j - \theta_*)\rangle \;=\; \mathbb{E}\sum_{i=0}^{n}\langle\theta_i - \theta_*, \Sigma(\theta_i - \theta_*)\rangle + 2\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle\theta_i - \theta_*, \Sigma(\theta_j - \theta_*)\rangle.$$

Moreover,

$$
\begin{aligned}
&\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle\theta_i-\theta_*,\Sigma(\theta_j-\theta_*)\rangle\\
&=\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\Big\langle\theta_i-\theta_*,\Sigma\Big[M(j,i+1)(\theta_i-\theta_*)+\sum_{k=i+1}^{j}M(j,k+1)\gamma\varepsilon_k x_k\Big]\Big\rangle\\
&=\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle\theta_i-\theta_*,\Sigma M(j,i+1)(\theta_i-\theta_*)\rangle\text{ because }\varepsilon_k x_k\text{ and }\theta_i\text{ are independent,}\\
&=\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle\theta_i-\theta_*,\Sigma(I-\gamma\Sigma-\gamma\lambda I)^{j-i}(\theta_i-\theta_*)\rangle\text{ as }M(j,i+1)\text{ and }\theta_i\text{ are independent,}\\
&=\mathbb{E}\sum_{i=0}^{n-1}\Big\langle\theta_i-\theta_*,\gamma^{-1}\Sigma(\Sigma+\lambda I)^{-1}\big[(I-\gamma\Sigma-\gamma\lambda I)-(I-\gamma\Sigma-\gamma\lambda I)^{n-i+1}\big](\theta_i-\theta_*)\Big\rangle\\
&\leqslant\mathbb{E}\sum_{i=0}^{n}\Big\langle\theta_i-\theta_*,\gamma^{-1}\Sigma(\Sigma+\lambda I)^{-1}(I-\gamma\Sigma-\gamma\lambda I)(\theta_i-\theta_*)\Big\rangle\text{ using }(\Sigma+\lambda I)\preccurlyeq I,\\
&=\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle\theta_i-\theta_*,\Sigma(\Sigma+\lambda I)^{-1}(\theta_i-\theta_*)\rangle-\mathbb{E}\sum_{i=0}^{n}\langle\theta_i-\theta_*,\Sigma(\theta_i-\theta_*)\rangle.
\end{aligned}
$$

We thus simply need to bound $\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle\theta_i-\theta_*,\Sigma(\Sigma+\lambda I)^{-1}(\theta_i-\theta_*)\rangle$, to get a bound on $n^2\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n-\theta_*)\|^2$.

Recursion on operators. We have:

$$
\begin{aligned}
\mathbb{E}\big[M(i,k)\Sigma(\Sigma+\lambda I)^{-1}M(i,k)^*\big] &= \mathbb{E}\Big[M(i,k+1)\big[I-\gamma x_k\otimes x_k-\gamma\lambda I\big]\Sigma(\Sigma+\lambda I)^{-1}\\
&\quad\big[I-\gamma x_k\otimes x_k-\gamma\lambda I\big]M(i,k+1)^*\Big]\\
&= \mathbb{E}\Big[M(i,k+1)\Big(\Sigma(\Sigma+\lambda I)^{-1}-2\gamma\Sigma+\gamma^2\big[x_k\otimes x_k\\
&\quad+\lambda I\big]\Sigma(\Sigma+\lambda I)^{-1}\big[x_k\otimes x_k+\lambda I\big]\Big)M(i,k+1)^*\Big]\\
&\preccurlyeq \mathbb{E}\Big[M(i,k+1)\big[\Sigma(\Sigma+\lambda I)^{-1}-2\gamma\Sigma\\
&\quad+\gamma^2(R^2+2\lambda)\Sigma\big]M(i,k+1)^*\Big]\\
&= \mathbb{E}\Big[M(i,k+1)\Sigma(\Sigma+\lambda I)^{-1}M(i,k+1)^*\Big]\\
&\quad-\gamma(2-\gamma(R^2+2\lambda))\mathbb{E}\Big[M(i,k+1)\Sigma M(i,k+1)^*\Big],
\end{aligned}
$$

which leads to

$$
\begin{aligned}
\mathbb{E}\Big[M(i,k+1)\Sigma M(i,k+1)^*\Big] &\preccurlyeq \frac{1}{\gamma(2-\gamma(R^2+2\lambda))}\Big(E\Big[M(i,k+1)\Sigma(\Sigma+\lambda I)^{-1}M(i,k+1)^*\Big]\\
&\quad-E\Big[M(i,k)\Sigma(\Sigma+\lambda I)^{-1}M(i,k)^*\Big]\Big).
\end{aligned}
\tag{22}
$$

Using the operator T on matrices defined below, this corresponds to showing

$$(I - \gamma T)\big[\Sigma(\Sigma + \lambda I)\big] \preccurlyeq \Sigma(\Sigma + \lambda I) - \gamma\Sigma.$$

Noise term. For $\theta_0 - \theta_* = 0$, we have:

$$\begin{aligned}
& \mathbb{E}\langle \theta_i - \theta_*, \Sigma(\Sigma + \lambda I)^{-1}(\theta_i - \theta_*)\rangle \\
= \;& \gamma^2 \mathbb{E} \sum_{k=1}^{i} \sum_{j=1}^{i} \varepsilon_j x_j^* M(i, j+1)^* \Sigma(\Sigma + \lambda I)^{-1} M(i, k+1) \varepsilon_k x_k \text{ by expanding all terms,} \\
= \;& \gamma^2 \mathbb{E} \sum_{k=1}^{i} \varepsilon_k x_k^* M(i, k+1)^* \Sigma(\Sigma + \lambda I)^{-1} M(i, k+1) \varepsilon_k x_k \text{ using independence,} \\
= \;& \gamma^2 \operatorname{tr}\left(\sum_{k=1}^{i} \mathbb{E}\varepsilon_k^2 x_k x_k^* \mathbb{E} M(i, k+1)^* \Sigma(\Sigma + \lambda I)^{-1} M(i, k+1) \right) \\
\leqslant \;& \gamma^2 \sigma^2 \operatorname{tr}\left(\sum_{k=1}^{i} \mathbb{E} M(i, k+1) \Sigma M(i, k+1)^* \Sigma(\Sigma + \lambda I)^{-1} \right) \\
& \text{using our assumption regarding the noise.}
\end{aligned}$$

Using the recurrence between operators

$$\begin{aligned}
& \mathbb{E}\langle \theta_i - \theta_*, \Sigma(\Sigma + \lambda I)^{-1}(\theta_i - \theta_*)\rangle \\
\leqslant \;& \frac{\gamma\sigma^2}{2 - \gamma(R^2 + 2\lambda)} \operatorname{tr} \sum_{k=1}^{i} \Big(E\Big[M(i, k+1) \Sigma(\Sigma + \lambda I)^{-1} M(i, k+1)^* \Sigma(\Sigma + \lambda I)^{-1} \Big] \\
& - E\Big[M(i, k) \Sigma(\Sigma + \lambda I)^{-1} M(i, k)^* \Sigma(\Sigma + \lambda I)^{-1} \Big] \Big) \\
\leqslant \;& \frac{\gamma\sigma^2}{2 - \gamma(R^2 + 2\lambda)} \operatorname{tr} \Big(E\Big[M(i, i+1) \Sigma(\Sigma + \lambda I)^{-1} M(i, i+1)^* \Sigma(\Sigma + \lambda I)^{-1} \Big] \\
& - E\Big[M(i, 1) \Sigma(\Sigma + \lambda I)^{-1} M(i, 1)^* \Sigma(\Sigma + \lambda I)^{-1} \Big] \Big) \text{ by summing,} \\
\leqslant \;& \frac{\gamma\sigma^2}{2 - \gamma(R^2 + 2\lambda)} \operatorname{tr} \Sigma^2 (\Sigma + \lambda I)^{-2}.
\end{aligned}$$

This implies that for the noise process

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|^2 \leqslant \left(\frac{\sigma^2}{n+1} \operatorname{tr}\big[\Sigma^2(\Sigma + \lambda I)^{-2}\big] \right) \frac{1}{1 - \gamma(R^2/2 + \lambda)}.$$

Note that when γ tends to zero, we recover the optimal variance term.

Noiseless term. Without noise, we then need to bound:

$$\gamma^{-1} \mathbb{E} \sum_{i=0}^{n} \langle \theta_i - \theta_*, \Sigma(\Sigma + \lambda I)^{-1}(\theta_i - \theta_*)\rangle,$$

with $\theta_i - \theta_* = M(i,1)(\theta_0 - \theta_*)$, that is

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n} \operatorname{tr}\Big[M(i,1)^*\Sigma(\Sigma+\lambda I)^{-1}M(i,1)(\theta_0-\theta_*)(\theta_0-\theta_*)^*\Big].$$

We follow here the proof of Défossez and Bach (2015) and consider the operator T from symmetric matrices to symmetric matrices defined as

$$TA = (\Sigma+\lambda I)A + A(\Sigma+\lambda I) - \gamma E\big[(x_n\otimes x_n+\lambda I)A(x_n\otimes x_n+\lambda I)\big].$$

of the form $TA = (\Sigma+\lambda I)A + (\Sigma+\lambda I)A - \gamma SA$.

The operator S is self-adjoint and positive. Moreover:

$$\begin{aligned}
\langle A, SA\rangle &= \mathbb{E}\operatorname{tr}\big[A(x_n\otimes x_n+\lambda I)A(x_n\otimes x_n+\lambda I)\big] \\
&= \operatorname{tr}\big[2A^2\lambda\Sigma+\lambda^2A^2\big] + \mathbb{E}\operatorname{tr}\big[\langle x_n, Ax_n\rangle^2\big] \\
&\leqslant \operatorname{tr}\big[2A^2\lambda\Sigma+\lambda^2A^2\big] + \mathbb{E}\operatorname{tr}\big[\|x_n\|^2 x_n\otimes x_n, A^2\big] \text{ using Cauchy-Schwarz inequality,} \\
&\leqslant \operatorname{tr}\big[2A^2\lambda\Sigma+\lambda^2A^2\big] + R^2\operatorname{tr}\Sigma A^2 \\
&\leqslant (R^2+2\lambda)\operatorname{tr}\big[\Sigma+\lambda I]A^2.
\end{aligned}$$

We have for any symmetric matrix A:

$$\mathbb{E}M(i,1)^*AM(i,1) = (I-\gamma T)^i A.$$

Thus,

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\operatorname{tr}\Big[M(i,1)^*\Sigma(\Sigma+\lambda I)^{-1}M(i,1)(\theta_0-\theta_*)(\theta_0-\theta_*)^*\Big] = \gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle\langle(I-\gamma T)^iA, E_0\rangle\rangle$$

with $E_0 = (\theta_0-\theta_*)(\theta_0-\theta_*)^*$ and $A = \Sigma(\Sigma+\lambda I)^{-1}$. This leads to

$$\gamma^{-1}\mathbb{E}\langle\langle\gamma^{-1}T^{-1}(I-(I-\gamma T)^{n+1})A, E_0\rangle\rangle,$$

where $\langle\langle\cdot,\cdot\rangle\rangle$ denote the dot-product between self-adjoint operators.

The sum is less than its limit for $n\to\infty$, and thus, we can get rid of the term $(I-\gamma T)^{n+1}$, and we need to bound

$$\gamma^{-2}\langle\langle M, E_0\rangle\rangle = \gamma^{-2}\langle\langle T^{-1}(\Sigma(\Sigma+\lambda I)^{-1}), E_0\rangle\rangle,$$

with $M := T^{-1}\big[\Sigma(\Sigma+\lambda I)^{-1}\big]$, i.e., such that

$$\begin{aligned}
\Sigma(\Sigma+\lambda I)^{-1} &= (\Sigma+\lambda I)M + M(\Sigma+\lambda I) - \gamma\mathbb{E}(x_n\otimes x_n+\lambda I)M(x_n\otimes x_n+\lambda I) \\
&= (\Sigma+\lambda I)M + M(\Sigma+\lambda I) - \gamma SM. \qquad (23)
\end{aligned}$$

So that :

$$\begin{aligned}
M &= \big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1}\big[\Sigma(\Sigma+\lambda I)^{-1}\big] + \gamma\big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1}SM \\
&= \frac{1}{2}\Sigma(\Sigma+\lambda I)^{-2} + \gamma\big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1}SM.
\end{aligned}$$

The operator $(\Sigma + \lambda I) \otimes I + I \otimes (\Sigma + \lambda I)$ is self adjoint, and so is its inverse, thus:

$$\begin{aligned}\gamma^{-2}\langle\langle M, E_0\rangle\rangle &= \gamma^{-2}\langle\langle \frac{1}{2}\Sigma(\Sigma+\lambda I)^{-2} + \gamma\big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1} SM, E_0\rangle\rangle \\ &= \frac{1}{2}\gamma^{-2}\langle\langle \Sigma(\Sigma+\lambda I)^{-2}, E_0\rangle\rangle + \gamma^{-1}\langle\langle SM, \big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1} E_0\rangle\rangle \\ &= \frac{1}{2}\gamma^{-2}\operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-2}E_0) + \gamma^{-1}\langle\langle SM, \big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1} E_0\rangle\rangle.\end{aligned}$$

Moreover,

$$\begin{aligned}E_0 &= (\theta_0-\theta_*)(\theta_0-\theta_*)^* \\ &= (\Sigma+\lambda I)^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1/2}(\Sigma+\lambda I)^{+1/2} \\ &\preccurlyeq [(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)]\ \ (\Sigma+\lambda I), \\ &\quad \text{as } (\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1/2} \preccurlyeq (\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*) I.\end{aligned}$$

Thus, as $[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)]^{-1}$ is an non-decreasing operator on $(S_n(\mathbb{R}), \preccurlyeq)$ (see technical Lemma 15 in Appendix E):

$$\begin{aligned}&\big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1} E_0 \\ \preccurlyeq\ &\big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1} \big([(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)](\Sigma+\lambda I)\big) \\ =\ &\frac{(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)}{2} I.\end{aligned}$$

Thus as SM is positive :

$$\gamma^{-2}\langle\langle M, E_0\rangle\rangle \ \leq\ \frac{1}{2\gamma^2}\operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-2}E_0) + \frac{(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)}{2\gamma}\operatorname{tr}(SM).$$

Moreover we can upper bound $\operatorname{tr}(SM)$: using Equation (23) we have

$$\operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-1}) = 2\operatorname{tr}(\Sigma+\lambda I)M - \gamma \operatorname{tr}\mathbb{E}(x_n\otimes x_n + \lambda I)M(x_n\otimes x_n+\lambda I)$$

then, using Assumption $(\mathcal{A}_1)$:

$$\operatorname{tr}\mathbb{E}(x_n\otimes x_n+\lambda I)M(x_n\otimes x_n+\lambda I) \leqslant R^2 \operatorname{tr} M\Sigma + 2\operatorname{tr} M\Sigma\lambda + \lambda^2 \operatorname{tr} M \leqslant (R^2+2\lambda)\operatorname{tr} M(\Sigma+\lambda I).$$

This implies

$$\begin{aligned}\operatorname{tr}\big[\Sigma(\Sigma+\lambda I)^{-1}\big] &\geqslant \big(\frac{2}{R^2+2\lambda} - \gamma\big)\operatorname{tr}\mathbb{E}(x_n\otimes x_n+\lambda I)M(x_n\otimes x_n+\lambda I), \\ &\geqslant \frac{1}{R^2+2\lambda}\operatorname{tr}\mathbb{E}(x_n\otimes x_n+\lambda I)M(x_n\otimes x_n+\lambda I) \text{ since } \gamma(R^2+2\lambda)\leqslant 1, \\ &\geqslant \frac{1}{R^2+2\lambda}\operatorname{tr} SM.\end{aligned}$$

Thus finally:

$$\begin{aligned}\gamma^{-2}\langle\langle M, E_0\rangle\rangle \leq \frac{1}{2\gamma^2}\operatorname{tr} E_0\Sigma(\Sigma+\lambda I)^{-2} \\ + \frac{(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)}{2\gamma}(R^2+2\lambda)\operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-1}),\end{aligned}$$

which leads to the desired error term.

B.4 Proof When Only $\|\theta_0 - \theta_*\|$ Is Finite

When $\lambda = 0$, without noise, we then need to bound:

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle\theta_i - \theta_*, \theta_i - \theta_*\rangle,$$

with $\theta_i - \theta_* = M(i,1)(\theta_0 - \theta_*)$, that is

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\operatorname{tr}\Big[M(i,1)^*M(i,1)(\theta_0 - \theta_*)(\theta_0 - \theta_*)^*\Big].$$

By definition of $M(i,1)$ we have that $\mathbb{E}M(i,1)^*M(i,1) \preccurlyeq I$ leading to

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle\theta_i - \theta_*, \theta_i - \theta_*\rangle \leq \frac{(n+1)\|\theta_0 - \theta_*\|^2}{\gamma}.$$

For the regularization-based bias we also have

$$\|\lambda\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2 \leq \lambda\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1/2}(\theta_0 - \theta_*)\|^2.$$

B.5 Proof When the Noise Is Not Structured

For $\|\theta_0 - \theta_*\| = 0$ we have $\theta_n - \theta_* = \gamma\sum_{k=1}^{n} M(n,k+1)\varepsilon_k x_k$ which leads to

$$\mathbb{E}\|\Sigma^{1/2}(\theta_n - \theta_*)\|^2 = \gamma^2\sum_{k=1}^{n}\operatorname{tr}\mathbb{E}M(n,k+1)^*\Sigma M(n,k+1)V,$$

where $V = \mathbb{E}\varepsilon_k^2 x_k x_k^*$. And using the recursion on operators in Eq. (22) by changing order of elements we have

$$\begin{aligned}\mathbb{E}\Big[M(n,k+1)^*\Sigma M(n,k+1)\Big] \preccurlyeq \frac{1}{\gamma(2-\gamma(R^2+2\lambda))}\Big(&E\Big[M(n,k+1)^*\Sigma(\Sigma+\lambda I)^{-1}M(n,k+1)\Big]\\ &- E\Big[M(n,k)^*\Sigma(\Sigma+\lambda I)^{-1}M(n,k)\Big]\Big).\end{aligned}$$

And by adding the terms

$$\mathbb{E}\|\Sigma^{1/2}(\theta_n - \theta_*)\|^2 \preccurlyeq \frac{\gamma^2}{\gamma(2-\gamma(R^2+2\lambda))}\operatorname{tr}\Sigma(\Sigma+\lambda I)^{-1}V,$$

We conclude by convexity

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|^2 \preccurlyeq \frac{\gamma^2}{\gamma(2-\gamma(R^2+2\lambda))}\operatorname{tr}\Sigma(\Sigma+\lambda I)^{-1}V.$$

Appendix C. Convergence of Accelerated Averaged Stochastic Gradient Descent

We now prove Theorem 3. We thus consider iterates satisfying Eq. (10), under Assumptions $(\mathcal{A}_3)$, $(\mathcal{A}_4)$. We consider a fixed step size γ such that $\gamma(\Sigma + \lambda I) \preccurlyeq I$. Seeing Eq. (10) as a linear second order for θ_n, we will derive from exact calculations a decomposition of the errors a sum of three terms that will be studied independently. The proof is organized as follows: in Section C.1, we state the formulation as a second order linear system and derive the three main terms that have to be studied (see Lemma 9). Section C.2 studies asymptotic behaviors of the three terms, ignoring some exponentially decreasing terms, in order to give insight of how they behave. This section is not necessary for the proof, indeed a direct and exact calculation in the eigenbasis of Σ, following O'Donoghue and Candès (2013), is provided in Section C.3. Results are summed up in Section C.4.

C.1 General Expansion

We study the regularized stochastic accelerated gradient descent recursion defined for $n \geq 1$ by

$$\begin{aligned} \theta_n &= \nu_{n-1} - \gamma f'(\nu_{n-1}) - \gamma\lambda(\nu_n - \theta_0) + \gamma\xi_n \\ \nu_n &= \theta_n + \delta(\theta_n - \theta_{n-1}), \end{aligned}$$

starting from $\theta_0 = \nu_0 \in \mathcal{H}$. We may rewrite it for a quadratic function $f : \theta \mapsto \frac{1}{2}\langle \theta - \theta_*, \Sigma(\theta - \theta_*)\rangle$ for $n \geq 2$ as

$$\theta_n = \big[I - \gamma\Sigma - \gamma\lambda I\big]\big[\theta_{n-1} + \delta(\theta_{n-1} - \theta_{n-2})\big] + \gamma\xi_n + \gamma\lambda\theta_0 + \gamma\Sigma\theta_*,$$

with $\theta_0 \in \mathcal{H}$ and $\theta_1 = \big[I - \gamma\Sigma - \gamma\lambda I\big]\theta_0 + \gamma\xi_1 + \gamma\lambda\theta_0 + \gamma\Sigma\theta_*$.

And by centering around the optimum, we get:

$$\theta_n - \theta_* = \big[I - \gamma\Sigma - \gamma\lambda I\big]\big[\theta_{n-1} - \theta_* + \delta(\theta_{n-1} - \theta_* - \theta_{n-2} + \theta_*)\big] + \gamma\xi_n + \lambda\gamma(\theta_0 - \theta_*).$$

Thus this is a second order iterative system which is standard to cast in a linear form

$$\Theta_n = F\Theta_{n-1} + \gamma\Xi_n + \gamma\lambda\Theta_\lambda, \tag{24}$$

with $T = I - \gamma\Sigma - \gamma\lambda I$, $F = \begin{pmatrix} (1+\delta)T & -\delta T \\ I & 0 \end{pmatrix}$, $\Theta_n = \begin{pmatrix} \theta_n - \theta_* \\ \theta_{n-1} - \theta_* \end{pmatrix}$, $\Theta_0 = \begin{pmatrix} \theta_0 - \theta_* \\ \theta_0 - \theta_* \end{pmatrix}$, $\Xi_n = \begin{pmatrix} \xi_n \\ 0 \end{pmatrix}$ and $\Theta_\lambda = \begin{pmatrix} \theta_0 - \theta_* \\ 0 \end{pmatrix}$.

We are interested in the behavior of the average $\bar{\Theta}_n = \frac{1}{n+1}\sum_{k=0}^n \Theta_k$ for which we have the following general convergence result:

Lemma 9 *For all $\lambda \in \mathbb{R}_+$ and γ such that $\gamma(\Sigma + \lambda I) \preccurlyeq I$ and any matrix C the average of the iterates Θ_n defined by Eq. (24) satisfy for $P_k \overset{(def)}{=} C^{1/2}(I - F^k)(I - F)^{-1}$, with*

$\tilde{\Theta}_0 = \Theta_0 - \gamma\lambda(I-F)^{-1}\Theta_\lambda$,

$$\begin{aligned}\mathbb{E}\langle \bar{\Theta}_n, C\bar{\Theta}_n\rangle &\leq 2\left(\gamma\lambda\right)^2 \|C^{1/2}(I-F)^{-1}\Theta_\lambda\|^2 + \frac{2}{(n+1)^2}\|P_{n+1}\tilde{\Theta}_0\|^2 \\ &\quad + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n} \operatorname{tr} P_j V P_j^\top .\end{aligned}$$

The error thus decomposes as the sum of three main terms:

- the two first ones are bias terms, one arising from the regularization (the first one), and one arising computation (the second one),
- a variance term. which is the last one.

We remark that as we have assumed that Σ is invertible, the matrix $I-F$ can be shown to be invertible for all the considered δ.

The regularization-based term will be studied directly whereas the two others will be studied in two stages. First a heuristic will lead to an asymptotic bound then an exact computation will give a non-asymptotic bound. Then using $C = H = \begin{pmatrix}\Sigma & 0\\ 0 & 0\end{pmatrix}$ would give a convergence result on the function value and $C = \begin{pmatrix}I & 0\\ 0 & 0\end{pmatrix}$ a result on the iterate. The end of the section is devoted to the proof of this lemma.

Proof The sequence Θ_n satisfies a linear recursion, from which we get, for all $n \geq 1$:

$$\begin{aligned}\Theta_n &= F^n\Theta_0 + \gamma\sum_{k=1}^{n} F^{n-k}\Xi_k + \gamma\lambda\sum_{k=1}^{n} F^{n-k}\Theta_\lambda \\ &= F^n\Theta_0 + \gamma\sum_{k=1}^{n} F^{n-k}\Xi_k + \gamma\lambda(I-F^n)(I-F)^{-1}\Theta_\lambda.\end{aligned}$$

We study the averaged sequence: $\bar{\Theta}_n = \frac{1}{n+1}\sum_{k=0}^{n}\Theta_k$. Using the identity $\sum_{k=0}^{n-1} F^k = (I-F^n)(I-F)^{-1}$, we get

$$\bar{\Theta}_n = \frac{1}{n+1}\sum_{k=0}^{n} F^k\Theta_0 + \frac{\gamma}{n+1}\sum_{k=1}^{n}\sum_{j=1}^{k} F^{k-j}\Xi_j + \frac{\gamma\lambda}{n+1}\sum_{k=1}^{n}(I-F^k)(I-F)^{-1}\Theta_\lambda.$$

With

$$\tilde{\Theta}_0 = \Theta_0 - \gamma\lambda(I-F)^{-1}\Theta_\lambda,$$

and $\sum_{k=1}^{n}(I-F^k) = \sum_{k=0}^{n}(I-F^k) = [n+1-(I-F^{n+1})(I-F)^{-1}]$.

Using summation formulas for geometric series, we derive:

$$\begin{aligned}
\bar{\Theta}_n &= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{k=1}^{n}\sum_{j=1}^{k} F^{k-j}\Xi_j + \gamma\lambda(I-F)^{-1}\Theta_\lambda \\
&= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{j=1}^{n}\big(\sum_{k=j}^{n} F^{k-j}\big)\Xi_j + \gamma\lambda(I-F)^{-1}\Theta_\lambda \\
&= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{j=1}^{n}\big(\sum_{k=0}^{n-j} F^{k}\big)\Xi_j + \gamma\lambda(I-F)^{-1}\Theta_\lambda \\
&= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{j=1}^{n}(I-F^{n+1-j})(I-F)^{-1}\Xi_j + \gamma\lambda(I-F)^{-1}\Theta_\lambda \\
&= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{j=1}^{n}(I-F^{j})(I-F)^{-1}\Xi_{n+1-j} + \gamma\lambda(I-F)^{-1}\Theta_\lambda.
\end{aligned}$$

Using martingale square moment inequalities which amount to consider $\Xi_i, i = 1, ..., n$ independent, so that the variance of the sum is the sum of variances, and denoting by $V = \mathbb{E}[\Xi_n \otimes \Xi_n]$ we have for any positive semi-definite C,

$$\begin{aligned}
\mathbb{E}\langle \bar{\Theta}_n, C\bar{\Theta}_n\rangle &= \left\| C^{1/2}\left(\frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \gamma\lambda(I-F)^{-1}\Theta_\lambda\right)\right\|^2 \\
&\quad + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n}\operatorname{tr}(I-F^j)(I-F)^{-1}V(I-F^\top)^{-1}(I-F^j)^\top C,
\end{aligned}$$

where $C^{1/2}$ denotes a symmetric square root of C. Define $P_k \overset{(def)}{=} C^{1/2}(I-F^k)(I-F)^{-1}$, we have, Using Minkowski's inequality and inequality $(a+b)^2 \leq 2(a^2+b^2)$ for any $a, b \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\langle \bar{\Theta}_n, C\bar{\Theta}_n\rangle &= \left\|\frac{1}{n+1}P_{n+1}\tilde{\Theta}_0 + \gamma\lambda C^{1/2}(I-F)^{-1}\Theta_\lambda\right\|^2 + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n}\operatorname{tr} P_j V P_j^\top \\
&\leq 2\,(\gamma\lambda)^2\,\|C^{1/2}(I-F)^{-1}\Theta_\lambda\|^2 + \frac{2\|P_{n+1}\tilde{\Theta}_0\|^2}{(n+1)^2} + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n}\operatorname{tr} P_j V P_j^\top.
\end{aligned}$$

This concludes proof of Lemma 9. ■

C.2 Asymptotic Expansion

To give the main terms that we expect, we first provide an asymptotic analysis, which shall only be understood as an insight and is not necessary for the proof. Operator F will have only eigenvalues smaller than 1, thus $\left|\!\left|\!\left|F^j\right|\!\right|\!\right|$ will decrease exponentially to 0 as $j \to \infty$ (even if $|||F|||$[5] might be bigger than 1). The asymptotic analysis relies on ignoring all terms in

5. $|||F|||$ denotes the operator norm of F, i.e., $\sup_{\|x\|\leq 1}\|Fx\|$.

which F^j appears. We thus approximately have:

$$\begin{aligned}
\mathbb{E}\langle \bar{\Theta}_n, C\bar{\Theta}_n\rangle &\leq 2(\gamma\lambda)^2 \|C^{1/2}(I-F)^{-1}\Theta_\lambda\|^2 + 2\left\|C^{1/2}\frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0\right\|^2 \\
&\quad + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^n \operatorname{tr}(I-F^j)(I-F)^{-1}V(I-F^\top)^{-1}(I-F^j)^\top C \\
&\approx 2(\gamma\lambda)^2 \|C^{1/2}(I-F)^{-1}\Theta_\lambda\|^2 + 2\left\|C^{1/2}\frac{1}{n+1}(I-F)^{-1}\tilde{\Theta}_0\right\|^2 \\
&\quad + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^n \operatorname{tr}(I-F)^{-1}V(I-F^\top)^{-1}C,
\end{aligned}$$

where, as it has been explained $\approx$ stands for an equality up to terms that will decay exponentially. However, these terms have to be studied very carefully, what will be done in the Section C.3.

Using the matrix inversion lemma we have for $C = \begin{pmatrix} c & 0 \\ 0 & 0 \end{pmatrix}$,

$$\begin{aligned}
I-F &= \begin{pmatrix} (1+\delta)(\gamma\Sigma+\gamma\lambda I) - \delta I & \delta(I-(\gamma\Sigma+\gamma\lambda I)) \\ -I & I \end{pmatrix} \\
(I-F)^{-1} &= \begin{pmatrix} (\gamma\Sigma+\gamma\lambda I)^{-1} & \delta\big(I-(\gamma\Sigma+\gamma\lambda I)^{-1}\big) \\ (\gamma\Sigma+\gamma\lambda I)^{-1} & (1+\delta)I - \delta(\gamma\Sigma+\gamma\lambda I)^{-1} \end{pmatrix} \\
C^{1/2}(I-F)^{-1} &= \begin{pmatrix} c^{1/2}(\gamma\Sigma+\gamma\lambda I)^{-1} & \delta c^{1/2}\big(I-(\gamma\Sigma+\gamma\lambda I)^{-1}\big) \\ 0 & 0 \end{pmatrix}.
\end{aligned} \tag{25}$$

Regularization based term. This gives for the regularization based term

$$\begin{aligned}
\left\|C^{1/2}(I-F)^{-1}\Theta_\lambda\right\|^2 &= \left\| \begin{pmatrix} c^{1/2}(\gamma\Sigma+\gamma\lambda I)^{-1} & \delta c^{1/2}\big(I-(\gamma\Sigma+\gamma\lambda I)^{-1}\big) \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \theta_0-\theta_* \\ 0 \end{pmatrix} \right\|^2 \\
&= \left(\frac{1}{\gamma}\right)^2 \|(c^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*))\|^2.
\end{aligned} \tag{26}$$

The computation of this term is exact (not asymptotic).

Bias term. For the bias term we have

$$\begin{aligned}
\tilde{\Theta}_0 &= \Theta_0 - \gamma\lambda(I-F)^{-1}\Theta_\lambda \\
&= \begin{pmatrix} \theta_0-\theta_* \\ \theta_0-\theta_* \end{pmatrix} - \gamma\lambda \begin{pmatrix} (\gamma\Sigma+\gamma\lambda I)^{-1} & \delta\big(I-(\gamma\Sigma+\gamma\lambda I)^{-1}\big) \\ (\gamma\Sigma+\gamma\lambda I)^{-1} & (1+\delta)I-\delta(\gamma\Sigma+\gamma\lambda I)^{-1} \end{pmatrix} \begin{pmatrix} \theta_0-\theta_* \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} \theta_0-\theta_* \\ \theta_0-\theta_* \end{pmatrix} - \gamma\lambda \begin{pmatrix} (\gamma\Sigma+\gamma\lambda I)^{-1}(\theta_0-\theta_*) \\ (\gamma\Sigma+\gamma\lambda I)^{-1}(\theta_0-\theta_*) \end{pmatrix} \\
&= \begin{pmatrix} [I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*) \\ [I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*) \end{pmatrix}.
\end{aligned}$$

Thus this gives for the dominant term

$$\begin{aligned}\left\|C^{1/2}(I-F)^{-1}\tilde{\Theta}_0\right\|^2 &= \left\|\begin{pmatrix} c^{1/2}(\gamma\Sigma+\gamma\lambda I)^{-1} & \delta c^{1/2}\big(I-(\gamma\Sigma+\gamma\lambda I)^{-1}\big) \\ 0 & 0\end{pmatrix}\tilde{\Theta}_0\right\|^2 \\ &= \|(c^{1/2}[(1-\delta)(\gamma\Sigma+\gamma\lambda I)^{-1}+\delta I][I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*)\|^2.\end{aligned}$$

And if c commutes with Σ we have the bound for $\delta \in [\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1]$

$$\begin{aligned}\left\|C^{1/2}(I-F)^{-1}\tilde{\Theta}_0\right\|^2 &\leq (\frac{(1-\delta)}{\gamma\lambda}+\delta)\|(c^{1/2}[I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*)\|^2 \\ &\leq (\frac{2}{\sqrt{\gamma\lambda}}+1)\|(c^{1/2}[I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*)\|^2.\end{aligned}$$

Variance term. And for the variance term with $V = \begin{pmatrix} v & 0 \\ 0 & 0\end{pmatrix}$, we have $C^{1/2}(I-F)^{-1}V^{1/2} = \begin{pmatrix} c^{1/2}(\gamma\Sigma+\gamma\lambda I)^{-1}v^{1/2} & 0 \\ 0 & 0\end{pmatrix}$, and

$$\operatorname{tr} C^{1/2}(I-F)^{-1}V(I-F^\top)^{-1}C^{1/2} = \operatorname{tr} c(\gamma\Sigma+\gamma\lambda I)^{-1}v(\gamma\Sigma+\gamma\lambda I)^{-1}.$$

This gives the three dominant terms. However in order to control the remainders we have to compute the eigenvalues more carefully, as done in the next section.

C.3 Direct Computation without the Regularization Based Term

We derive now direct computation both the bias and variance terms. This is not required for the regularization based term whose previous expression in Eq. (26) is already non-asymptotic. Following O'Donoghue and Candès (2013) we consider an eigen-decomposition of the matrix F, in order to study independently the recursion on eigenspaces. We assume Σ has eigenvalues (s_i) and we decompose vectors in an eigenvector basis of Σ we denote by (p_i), with $\theta_n^i = p_i^\top \theta_n$ and $\xi_n^i = p_i^\top \xi_n$ and we have the reduced equation:

$$\Theta_{n+1}^i = F_i\Theta_n^i + \gamma\Xi_{n+1}^i.$$

with $\Theta_0^i = \tilde{\Theta}_0^i$, $F_i = \begin{pmatrix}(1+\delta)T_i & -\delta T_i \\ 1 & 0\end{pmatrix}$, with $T_i = 1-\gamma s_i - \gamma\lambda$.

Computing initial point $\tilde{\Theta}_0^i$. $\tilde{\Theta}_0^i = \Theta_0^i - \gamma\lambda(I-F_i)^{-1}\Theta_\lambda^i$, with $\Theta_0^i = \begin{pmatrix}\theta_0^i-\theta_*^i \\ \theta_0^i-\theta_*^i\end{pmatrix}$, $\Theta_\lambda^i = \begin{pmatrix}\theta_0^i-\theta_*^i \\ 0\end{pmatrix}$ and $(I-F_i)^{-1}$ given in Eq. (25). Thus

$$\begin{aligned}\tilde{\Theta}_0^i &= \begin{pmatrix}\theta_0^i-\theta_*^i \\ \theta_0^i-\theta_*^i\end{pmatrix} - \frac{\gamma\lambda}{(\gamma s_i+\gamma\lambda)}\begin{pmatrix}1 & \delta((\gamma s_i+\gamma\lambda)-1) \\ 1 & (1+\delta)(\gamma s_i+\gamma\lambda)-\delta\end{pmatrix}\begin{pmatrix}\theta_0^i-\theta_*^i \\ 0\end{pmatrix} \\ &= \begin{pmatrix}(1-\frac{\lambda}{\lambda+s_i})(\theta_0^i-\theta_*^i) \\ (1-\frac{\lambda}{\lambda+s_i})(\theta_0^i-\theta_*^i)\end{pmatrix}. \qquad (27)\end{aligned}$$

Study of spectrum of F_i. Depending on δ, F_i may have two distinct complex eigenvalues of same modulus, only one (double) eigenvalue, or two real eigenvalues. We only consider the two former cases, which we detail below.

Indeed, the characteristic polynomial

$$\chi_{F_i}(X) \stackrel{def}{=} \det(XI - F_i) = X^2 - (1+\delta)(1-\gamma(s_i+\lambda))X + \delta(1-\gamma(s_i+\lambda))$$

has discriminant $\Delta_i = (1-\gamma(s_i+\lambda))((1+\delta)^2(1-\gamma(s_i+\lambda)) - 4\delta)$ which is non positive as far as $\delta \in [\delta_-; \delta_+]$, with $\delta_- = \frac{1-\sqrt{\gamma(s_i+\lambda)}}{1+\sqrt{\gamma(s_i+\lambda)}}$, $\delta_+ = \frac{1+\sqrt{\gamma(s_i+\lambda)}}{1-\sqrt{\gamma(s_i+\lambda)}}$.

C.3.1 Two Distinct Eigenvalues

We first assume that F_i has two distinct complex eigenvalues $r_\pm = \frac{(1+\delta)(1-\gamma(s_i+\lambda)) \pm \sqrt{-1}\sqrt{-\Delta_i}}{2}$ which are conjugate. Thus the roots are of the form $\rho_i e^{\pm i\omega_i}$ with $\rho_i = \sqrt{\delta(1-\gamma(s_i+\lambda))}$, $\cos(\omega_i) = \frac{(1+\delta)(1-\gamma(s_i+\lambda))}{2\rho_i}$, $\omega_i \in [-\pi/2; \pi/2]$ and $\sin(\omega_i) = \frac{\sqrt{-\Delta_i}}{2\rho_i}$.

Let $Q_i = \begin{pmatrix} r_i^- & r_i^+ \\ 1 & 1 \end{pmatrix}$ be the transfer matrix into an eigenbasis of F_i, i.e., $F_i = Q_i D_i Q_i^{-1}$ with $D_i = \begin{pmatrix} r_i^- & 0 \\ 0 & r_i^+ \end{pmatrix}$ and $Q_i^{-1} = \frac{1}{r_i^- - r_i^+}\begin{pmatrix} 1 & -r_i^+ \\ -1 & r_i^- \end{pmatrix}$.

Computing $P_{i,k}$. We first compute the matrix $P_{i,k}$: With

$$C_i^{1/2} = \begin{pmatrix} \sqrt{c_i} & 0 \\ 0 & 0 \end{pmatrix}, C_i^{1/2} Q_i = \begin{pmatrix} r_i^- \sqrt{c_i} & r_i^+ \sqrt{c_i} \\ 0 & 0 \end{pmatrix}$$

we have

$$C_i^{1/2} Q_i (I - D_i^k)(I - D_i)^{-1} = \sqrt{c_i} \begin{pmatrix} \frac{1-(r_i^-)^k}{1-r_i^-} r_i^- & \frac{1-(r_i^+)^k}{1-r_i^+} r_i^+ \\ 0 & 0 \end{pmatrix},$$

and, when developing and regrouping terms which depend on k, we get :

$$\begin{aligned} P_{i,k} &= C_i^{1/2} Q_i (I - D_i^k)(I - D_i)^{-1} Q_i^{-1} \\ &= \frac{\sqrt{c_i}}{r_i^- - r_i^+} \begin{pmatrix} \frac{1-(r_i^-)^k}{1-r_i^-} r_i^- - \frac{1-(r_i^+)^k}{1-r_i^+} r_i^+ & \frac{1-(r_i^+)^k}{1-r_i^+} r_i^- r_i^+ - \frac{1-(r_i^-)^k}{1-r_i^-} r_i^+ r_i^- \\ 0 & 0 \end{pmatrix} \\ &= \sqrt{c_i} \begin{pmatrix} \frac{1}{(1-r_i^-)(1-r_i^+)} & \frac{-r_i^+ r_i^-}{(1-r_i^-)(1-r_i^+)} \\ 0 & 0 \end{pmatrix} \\ &\quad - \frac{\sqrt{c_i}}{r_i^- - r_i^+} \begin{pmatrix} \frac{(r_i^-)^{k+1}}{1-r_i^-} - \frac{(r_i^+)^{k+1}}{1-r_i^+} & \frac{(r_i^+)^{k+1}}{1-r_i^+} r_i^- - \frac{(r_i^-)^{k+1}}{1-r_i^-} r_i^+ \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

We also have $P_{i,k} = C_i^{1/2} Q_i (I - D_i^k)(I - D_i)^{-1} Q_i^{-1} = \sum_{j=0}^{k-1} R_{i,j}$ with

$$\begin{aligned} R_{i,j} &= C_i^{1/2} Q_i D_i^j Q_i^{-1} \\ &= \sqrt{c_i} \begin{pmatrix} (r_i^-)^{j+1} & (r_i^+)^{j+1} \\ 0 & 0 \end{pmatrix} Q_i^{-1} \\ &= \frac{\sqrt{s_i}}{r_i^- - r_i^+} \begin{pmatrix} (r_i^-)^{j+1} - (r_i^+)^{j+1} & -r_i^+ (r_i^-)^{j+1} + r_i^- (r_i^+)^{j+1} \\ 0 & 0 \end{pmatrix}, \end{aligned}$$

but computing error terms based in $R_{i,j}$ before summing these errors gives a looser error bound than a tight calculation using $P_{i,k}$. More precisely, if we use $P_{i,k}\Theta_0^i = \sum_{j=0}^{k-1} R_{i,j}\Theta_0^i$ to upper bound $\|P_{i,k}\Theta_0^i\| \leq \sum_{j=0}^{k-1} \|R_{i,j}\Theta_0^i\|$, we end up with a worse bound.

Bias term. Thus, for the bias term:

$$\begin{aligned} P_{i,k}\Theta_0^i &= \sqrt{c_i}\theta_0^i \frac{1 - r_i^+ r_i^-}{(1-r_i^-)(1-r_i^+)} - \frac{\sqrt{c_i}\theta_0^i}{r_i^- - r_i^+} \begin{pmatrix} \left[(r_i^-)^{k+1}\frac{1-r_i^+}{1-r_i^-} - (r_i^+)^{k+1}\frac{1-r_i^-}{1-r_i^+}\right] \\ 0 \end{pmatrix} \\ &= \frac{\sqrt{c_i}\theta_0^i}{\sqrt{(1-r_i^-)(1-r_i^+)}} \begin{pmatrix} \frac{\left[(1-r_i^+r_i^-) - \rho_i^k A_1\right]}{\sqrt{(1-r_i^-)(1-r_i^+)}} \\ 0 \end{pmatrix}, \end{aligned}$$

where

$$\rho_i^k A_1 = \frac{(r_i^-)^{k+1}(1-r_i^+)^2 - (r_i^+)^{k+1}(1-r_i^-)^2}{r_i^- - r_i^+}.$$

This can be bound with the following lemma

Lemma 10 *For all $\rho \in (0,1)$ and $\omega \in [-\pi/2; \pi/2]$ and $r^\pm = \rho(\cos(\omega) \pm \sqrt{-1}\sin(\omega))$ we have:*

$$\left|\frac{1 - r^+r^- - \rho^k|A_1|}{|1-r^+|}\right| \leq 3 + 3\rho^k \leq 6 \tag{28}$$

We note that the exact constant seems empirically to be 2. This lemma is proved as Lemma 16 in Appendix E. This gives for the bias term

$$\begin{aligned} \|P_{i,k}\Theta_0^i\| &= \frac{\sqrt{c_i}(\theta_0^i)}{\sqrt{(1-r_i^-)(1-r_i^+)}} \Big[\frac{1}{\sqrt{(1-r_i^-)(1-r_i^+)}} \Big((1-r_i^+r_i^-) - \rho_i^k A_1\Big)\Big] \\ &\leq 6\frac{\sqrt{c_i}(\theta_0^i)}{\sqrt{\gamma(s_i+\lambda)}}, \end{aligned}$$

since:

$$\begin{aligned} (1-r_i^-)(1-r_i^+) &= 1 - 2\,\mathfrak{Re}\,(r_i^+) + |r_i^+|^2 \\ &= 1 - (1+\delta)(1-\gamma(s_i+\lambda)) + \delta(1-\gamma(s_i+\lambda)) \\ &= \gamma(s_i+\lambda). \end{aligned}$$

We also have a looser bound using $P_{i,k}\Theta_0^i = \sum_{j=0}^{k-1} R_{i,j}\Theta_0^i$.

$$\begin{aligned}
R_{i,j}\Theta_0^i &= \frac{\sqrt{c_i}\theta_0^i}{r_i^- - r_i^+}\big((1-r_i^+)(r_i^-)^{j+1} - (1-r_i^-)(r_i^+)^{j+1}\big) \\
&= \sqrt{c_i}\theta_0^i\left(\frac{(r_i^-)^{j+1} - (r_i^+)^{j+1}}{r_i^- - r_i^+} - \frac{r_i^+(r_i^-)^{j+1} - r_i^-(r_i^+)^{j+1}}{r_i^- - r_i^+}\right) \text{ using De Moivre's formula,} \\
&= \sqrt{c_i}\theta_0^i\left(\frac{\rho_i^{j+1}\sin(\omega_i(j+1))}{\rho_i\sin(\omega_i)} - \frac{\rho_i e^{i\omega_i}\rho_i^{j+1}e^{-i\omega_i(j+1)} - \rho_i e^{-i\omega_i}\rho_i^{j+1}e^{+i\omega_i(j+1)}}{\rho_i e^{-i\omega_i} - \rho_i e^{i\omega_i}}\right) \\
&= \sqrt{c_i}\theta_0^i\left(\frac{\rho_i^{j+1}\sin(\omega_i(j+1))}{\rho_i\sin(\omega_i)} - \rho_i^{j+1}\frac{e^{-i\omega_i j} - e^{+i\omega_i j}}{e^{-i\omega_i} - e^{i\omega_i}}\right) \\
&= \sqrt{c_i}\theta_0^i\left(\frac{\rho_i^{j}\sin(\omega_i(j+1))}{\sin(\omega_i)} - \rho_i^{j+1}\frac{\sin(\omega_i j)}{\sin(\omega_i)}\right) \\
&\leq (1+e^{-1})\sqrt{c_i}\theta_0^i \quad \text{using Lemma 17 (see proof in Appendix E),}
\end{aligned}$$

which also gives for the bias term

$$\|P_{i,k}\Theta_0^i\| \leq (1+e^{-1})\sqrt{c_i}\theta_0^i k.$$

Thus we have the final bound:

$$\|P_{i,k}\Theta_0^i\|^2 \leq \min\left\{36\frac{c_i(\theta_0^i)^2}{\gamma(s_i+\lambda)}, 6n(1+e^{-1})\frac{c_i(\theta_0^i)^2}{\sqrt{\gamma(s_i+\lambda)}}, n^2(1+e^{-1})^2 c_i(\theta_0^i)^2\right\}. \tag{29}$$

Variance term. As for the variance term, with $V_i = \begin{pmatrix} v_i & 0 \\ 0 & 0 \end{pmatrix}$, we have $\operatorname{tr} P_{i,k}V_iP_{i,k} = \left\|P_{i,k}\begin{pmatrix}\sqrt{v_i}\\0\end{pmatrix}\right\|^2$.

$$\begin{aligned}
\left\|P_{i,k}\begin{pmatrix}\sqrt{v_i}\\0\end{pmatrix}\right\| &= \frac{\sqrt{v_i c_i}}{(1-r_i^-)(1-r_i^+)}\left[1 + \frac{(r_i^-)^{k+1}(1-r_i^+) - (r_i^+)^{k+1}(1-r_i^-)}{r_i^+ - r_i^-}\right] \\
&= \frac{\sqrt{v_i c_i}}{\gamma(s_i+\lambda)}\Big[1 - \rho_i^k B_{i,k}\Big],
\end{aligned}$$

where

$$\rho_i^k B_{i,k} = -\frac{(r_i^-)^{k+1}(1-r_i^+) - (r_i^+)^{k+1}(1-r_i^-)}{r_i^+ - r_i^-},$$

which we can bound using the following Lemma:

Lemma 11 *For all $\rho \in (0,1)$ and $\omega \in [-\pi/2;\pi/2]$ and $r^\pm = \rho(\cos(\omega) \pm \sqrt{-1}\sin(\omega))$ we have:*

$$\Big|\rho^k B_k\Big| \leq 1.75.$$

Where we note that the exact majoration seems to be 1.3. This Lemma is proved as Lemma 18 in Appendix E.

We can also have a looser bound using $P_{i,k}\begin{pmatrix} v_i^{1/2} \\ 0 \end{pmatrix} = \sum_{j=0}^{k-1} R_{i,j}\begin{pmatrix} v_i^{1/2} \\ 0 \end{pmatrix}$ and

$$\begin{aligned} R_{i,j}\begin{pmatrix} v_i^{1/2} \\ 0 \end{pmatrix} &= \frac{\sqrt{c_i v_i}}{r_i^- - r_i^+}\left((r_i^-)^{j+1} - (r_i^+)^{j+1}\right) \\ &= \sqrt{c_i v_i}\ \frac{\rho_i^{j+1}\sin(\omega_i(j+1))}{\rho_i \sin(\omega_i)} \\ &\leq (j+1)\sqrt{c_i v_i}, \text{ using the inequality } |\sin(k\omega_i)| \leq k|\sin(\omega_i)| \end{aligned}$$

and $\left\| P_{i,k}\begin{pmatrix} v_i^{1/2} \\ 0 \end{pmatrix} \right\| \leq \frac{\sqrt{c_i v_i}(k+1)k}{2}$.

This gives for the Variance term

$$\begin{aligned} \sum_{k=1}^n \operatorname{tr} P_{i,k} V_i P_{i,k} &\leq v_i c_i \sum_{k=1}^n \min\left\{ \frac{\left[1-\rho_i^k B_{1,k}\right]^2}{\gamma^2(s_i+\lambda)^2}, \frac{\left[1-\rho_i^k B_{1,k}\right]k(k+1)}{2\gamma(s_i+\lambda)}, \frac{k^2(k+1)^2}{4} \right\} \\ &\leq v_i c_i \min\left\{ \frac{8n}{\gamma^2(s_i+\lambda)^2}, \frac{(n+1)^3}{2\gamma(s_i+\lambda)}, \frac{(n+1)^5}{20} \right\}. \end{aligned} \tag{30}$$

C.3.2 One Coalescent Eigenvalue

We now turn to the case where F has two coalescent eigenvalues, which happens when the discriminant $\Delta = 0$. We assume that F_i has one coalescent eigenvalue $r_i = \frac{(1+\delta)(1-\gamma(s_i+\lambda))}{2}$. Then, with $\delta = \frac{1-\sqrt{\gamma(s_i+\lambda)}}{1+\sqrt{\gamma(s_i+\lambda)}}$, $r_i = \frac{(1+\delta)(1-\gamma(s_i+\lambda))}{2} = 1 - \sqrt{\gamma(s_i+\lambda)}$. Then F_i can be trigonalized as $F_i = Q_i D_i Q_i^{-1}$ with $Q_i = \begin{pmatrix} r_i & 1 \\ 1 & 0 \end{pmatrix}$, $D_i = \begin{pmatrix} r_i & 1 \\ 0 & r_i \end{pmatrix}$ and $Q_i^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -r_i \end{pmatrix}$. We note that for all $k \geq 0$, then $D_i^k = r_i^{k-1}\begin{pmatrix} r_i & k \\ 0 & r_i \end{pmatrix}$.

Computing $P_{i,k}$. We first compute $P_{i,k}$:

$$(I_2 - D_i)^{-1} = \begin{pmatrix} \frac{1}{1-r_i} & \frac{1}{(1-r_i)^2} \\ 0 & \frac{1}{1-r_i} \end{pmatrix}$$

and

$$(I_2 - D_i^k)(I_2 - D_i)^{-1} = \begin{pmatrix} \frac{1-r_i^k}{1-r_i} & \frac{1-r_i^k}{(1-r_i)^2} - \frac{k r_i^{k-1}}{1-r_i} \\ 0 & \frac{1-r_i^k}{1-r_i} \end{pmatrix}.$$

Thus with $C_i^{1/2} Q_i = \begin{pmatrix} \sqrt{c_i} r_i & \sqrt{c_i} \\ 0 & 0 \end{pmatrix}$ we have

$$C_i^{1/2} Q_i (I_2 - D_i^k)(I_2 - D_i)^{-1} = \sqrt{c_i}\begin{pmatrix} \frac{1-r_i^k}{1-r_i} r_i & \frac{1-r_i^k}{(1-r_i)^2} - \frac{k r_i^k}{1-r_i} \\ 0 & 0 \end{pmatrix}.$$

And, computing as previously the matrices products, we derive:

$$\begin{aligned}
P_{i,k} &= C_i^{1/2} Q_i (I_2 - D_i^k)(I_2 - D_i)^{-1} Q_i^{-1} \\
&= \sqrt{c_i} \begin{pmatrix} \frac{1-r_i^k}{(1-r_i)^2} - \frac{kr_i^k}{1-r_i} & \frac{1-r_i^k}{1-r_i} r_i - (\frac{1-r_i^k}{(1-r_i)^2} - \frac{kr_i^k}{1-r_i}) r_i \\ 0 & 0 \end{pmatrix} \\
&= \sqrt{c_i} \begin{pmatrix} \frac{1-r_i^k}{(1-r_i)^2} - \frac{kr_i^k}{1-r_i} & \frac{1-r_i^k}{(1-r_i)^2}(r_i)^2 + \frac{kr_i^{k+1}}{1-r_i} \\ 0 & 0 \end{pmatrix} \\
&= \frac{\sqrt{c_i}}{1-r_i} \begin{pmatrix} \frac{1-r_i^k}{1-r_i} - kr_i^k & -\frac{1-r_i^k}{1-r_i}(r_i)^2 + kr_i^{k+1} \\ 0 & 0 \end{pmatrix}.
\end{aligned}$$

Bias term. We thus have:

$$\begin{aligned}
P_{i,k}\Theta_0^i &= \frac{\sqrt{c_i}}{1-r_i} \begin{pmatrix} \frac{1-r_i^k}{1-r_i} - kr_i^k & -\frac{1-r_i^k}{1-r_i}(r_i)^2 + kr_i^{k+1} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \theta_0^i \\ \theta_0^i \end{pmatrix} \\
&= \theta_0^i \sqrt{c_i} \begin{pmatrix} (1-r_i^k)\frac{1+r_i}{1-r_i} - kr_i^k \\ 0 \end{pmatrix},
\end{aligned}$$

and this gives for the bias term:

$$\begin{aligned}
&\|P_{i,k}\Theta_0^i\|^2 \\
=&(\theta_0^i)^2 c_i \Big[(1-r_i^k)\frac{1+r_i}{1-r_i} - kr_i^k\Big]^2 \\
=&(\theta_0^i)^2 c_i \Big[\frac{1+r_i}{1-r_i} - \Big(k + \frac{1+r_i}{1-r_i}\Big) r_i^k\Big]^2 \text{ developing the product, then using formulas for } r_i \\
=&(\theta_0^i)^2 c_i \Big[\frac{2-\sqrt{\gamma(s_i+\lambda)}}{\sqrt{\gamma(s_i+\lambda)}} - \Big(k + \frac{2-\sqrt{\gamma(s_i+\lambda)}}{\sqrt{\gamma(c_i+\lambda)}}\Big)(1-\sqrt{\gamma(s_i+\lambda)})^k\Big]^2 \\
=&\frac{(\theta_0^i)^2 c_i}{\gamma(s_i+\lambda)} \Big[2 - \sqrt{\gamma(s_i+\lambda)} - \big(k\sqrt{\gamma(s_i+\lambda)} + 2 - \sqrt{\gamma(s_i+\lambda)}\big)(1-\sqrt{\gamma(s_i+\lambda)})^k\Big]^2 \\
=&\frac{(\theta_0^i)^2 c_i}{\gamma(s_i+\lambda)} \Big[2 - \sqrt{\gamma(s_i+\lambda)} - \big(2 + (k-1)\sqrt{\gamma(s_i+\lambda)}\big)(1-\sqrt{\gamma(s_i+\lambda)})^k\Big]^2 \\
\leq&4\frac{(\theta_0^i)^2 c_i}{\gamma(s_i+\lambda)}, \text{ using Lemma 19 in Appendix E.}
\end{aligned}$$

Variance term. With $V = \begin{pmatrix} v_i & 0 \\ 0 & 0 \end{pmatrix}$,

$$
\begin{aligned}
&\operatorname{tr} P_{i,k} V P_{i,k} \\
=&\frac{s_i}{(1-r_i)^2} \begin{pmatrix} \frac{1-r_i^k}{1-r_i} - kr_i^k & -\frac{1-r_i^k}{1-r_i}(r_i)^2 + kr_i^{k+1} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} v_i & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1-r_i^k}{1-r_i} - kr_i^k & -\frac{1-r_i^k}{1-r_i}(r_i)^2 + kr_i^{k+1} \\ 0 & 0 \end{pmatrix}^\top \\
=&\frac{s_i v_i}{(1-r_i)^2}\Big[\frac{1-r_i^k}{1-r_i} - kr_i^k\Big]^2 \\
=&\frac{v_i h_i}{\gamma(s_i+\lambda)}\Big[\frac{1-r_i^k}{1-r_i} - kr_i^k\Big]^2 \\
=&\frac{v_i h_i}{\gamma(s_i+\lambda)(1-r_i)^2}\Big[1 - r_i^k - (1-r_i)kr_i^k\Big]^2 \\
=&\frac{v_i h_i}{\gamma^2(s_i+\lambda)^2}\Big[1 - (1+k\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^k\Big]^2,
\end{aligned}
$$

and

$$
\begin{aligned}
\sum_{k=1}^n \operatorname{tr} P_{i,k} V P_{i,k} &= \frac{v_i s_i}{\gamma^2(s_i+\lambda)^2} \sum_{k=1}^n \Big[1 - (1+k\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^k\Big]^2 \\
&\leq n\frac{v_i s_i}{\gamma^2(s_i+\lambda)^2} \text{ using Lemma 19 in Appendix E.} \qquad (31)
\end{aligned}
$$

Alternative bounds for the bias and the variance term, as in Equations(26), (29) may be derived as well. Combining all these results, we are now able to state Theorem 3.

C.4 Conclusion

Combining results from Lemma 9, and Equations (26), (29), (30), with $c = \Sigma$, and using the following simple facts:

- For the least squares regression function, with $c = \Sigma$, $\mathbb{E}\langle \bar{\Theta}_n, C\bar{\Theta}_n\rangle = \mathbb{E}f(\bar{\theta}_n) - f(\theta_*)$.
- Under assumption $\mathcal{A}_3$, $\mathcal{A}_4$, we have $V \preccurlyeq \tau^2\Sigma$.
- The squared norm of a vector is the sum of its squared components on the orthonormal eigenbasis. For example $\|P_{n+1}\Theta_0\|^2 = \sum_{i=1}^d \|P_{i,n+1}\Theta_0^i\|^2$.
- For any regularization parameter $\lambda \in \mathbb{R}_+$ and for any constant step-size $\gamma(\Sigma+\lambda I) \preccurlyeq I$, for any $\delta \in \big[\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1\big]$, matrix F will have only two distinct complex eigenvalues or two coalescent eigenvalues.

Proposition 12 *Under ($\mathcal{A}_{4,5}$), for any regularization parameter $\lambda \in \mathbb{R}_+$ and for any constant step-size $\gamma(\Sigma + \lambda I) \preccurlyeq I$ we have for any $\delta \in \left[\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1\right]$, for the recursion in Eq. (10):*

$$\begin{aligned}\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leq\ & 2\lambda\|\lambda^{1/2}\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2 \\ & + \sum_{i=1}^d \frac{2}{(n+1)^2}\min\left\{36\frac{c_i(\tilde{\theta}_0^i)^2}{\gamma(s_i+\lambda)}, 6n(1+e^{-1})\frac{c_i(\tilde{\theta}_0^i)^2}{\sqrt{\gamma(s_i+\lambda)}}, n^2(1+e^{-1})^2 c_i(\tilde{\theta}_0^i)^2\right\} \\ & + \sum_{i=1}^d \frac{\gamma^2}{(n+1)^2} v_i c_i \min\left\{\frac{8n}{\gamma^2(s_i+\lambda)^2}, \frac{(n+1)^3}{2\gamma(s_i+\lambda)}, \frac{(n+1)^5}{20}\right\}.\end{aligned}$$

This implies, using the Equation (27) for the initial point, using $c_i = \sigma_i$ and regrouping sums as traces or norms:

$$\begin{aligned}\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \ \leq\ & 2\lambda\|\lambda^{1/2}\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2 \\ & + \ 2\min\left\{\frac{36\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\|^2}{\gamma(n+1)^2}, (1+e^{-1})^2\|\Sigma^{1/2}(\theta_0-\theta_*)\|^2\right\} \\ & + \ \min\left\{\frac{8\operatorname{tr}(V\Sigma(\Sigma+\lambda I)^{-2})}{n+1}, n\gamma\operatorname{tr}(V\Sigma(\Sigma+\lambda I)^{-1})\right\},\end{aligned}$$

which gives exactly Theorem 3 using $V \preccurlyeq \tau^2\Sigma$ in the Variance term, and $\lambda^{1/2}(\Sigma+\lambda I)^{-1/2} \preccurlyeq I$ in the first term.

Appendix D. Tighter Bounds

D.1 Simple Upper-Bounds

In this section, we chow how tighter bounds naturally appear from the regularized quantities appearing in Theorems. It only relies on simple algebraic majorations, even if one has to be careful with the allowed intervals for r, b.

Lemma 13 *For any $\lambda \geq 0$, for any $b \in [0;1]$, if $\operatorname{tr}(\Sigma^b)$ exists, we have :*

$$\begin{aligned}\operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-1}) \ &\leq\ \frac{\operatorname{tr}(\Sigma^b)}{\lambda^b} \\ \operatorname{tr}(\Sigma^{-2}(\Sigma+\lambda I)^{-2}) \ &\leq\ \frac{\operatorname{tr}(\Sigma^b)}{\lambda^b}.\end{aligned}$$

Proof As all operators can be diagonalized in a same eigenbasis with positive eigenvalues, we have,

$$\begin{aligned}\operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-1}) \ &\leq\ \left|\left|\left|\Sigma^{1-b}(\Sigma+\lambda I)^{-1}\right|\right|\right| \operatorname{tr}(\Sigma^b) \\ |||\Sigma^{1-b}(\Sigma+\lambda I)^{-1}||| \ &\leq\ \sup_{0\leq x}\frac{x^{1-b}}{(x+\lambda)} \\ &\leq\ \sup_{0\leq x} x^{1-b}\left(\frac{1}{\lambda}\wedge\frac{1}{x}\right) \\ &\leq\ \sup_{0\leq x} x^{1-b}\left(\frac{1}{\lambda}\right)^b\left(\frac{1}{x}\right)^{1-b} = \lambda^{-b}.\end{aligned}$$

The calculations are exactly the same for $\mathrm{tr}(\Sigma^{-2}(\Sigma+\lambda I)^{-2}) \leq \frac{\mathrm{tr}(\Sigma^b)}{\lambda^b}$. ■

As for the bias term, we need to bound the following quantities :

Lemma 14 *For any $\lambda \geq 0$, for any $r \in [-1;1]$, we have :*

$$\left\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\right\|^2 \leq \lambda^{-(1+r)}\left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\|^2.$$

For any $\lambda \geq 0$, for any $r \in [-1;0]$, we have :

$$\left\|(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\right\|^2 \leq \lambda^{-(1+r)}\left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\|^2.$$

For any $\lambda \geq 0$, for any $r \in [0;1]$, we have :

$$\left\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\right\|^2 \leq \lambda^{-r}\left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\|^2$$

(No result when $r \leq 0$ because of saturation effect).

Proof Proof relies of simple following calculations:

$$\begin{aligned}
\left\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\right\| &\leq \left|\left|\left|\Sigma^{1/2-r/2}(\Sigma+\lambda I)^{-1}\right|\right|\right| \left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\| \\
&\leq \left(\frac{1}{\lambda}\right)^{1-(1/2-r/2)} \left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\| \\
&\leq \lambda^{-\frac{1+r}{2}}\left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\|
\end{aligned}$$

$$\begin{aligned}
\left\|(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\right\| &\leq \left|\left|\left|\Sigma^{-r/2}(\Sigma+\lambda I)^{-1/2}\right|\right|\right| \left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\| \\
&\leq \left(\frac{1}{\lambda}\right)^{\frac{1+r}{2}} \left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\| \\
&\leq \lambda^{-\frac{1+r}{2}}\left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\|
\end{aligned}$$

$$\begin{aligned}
\left\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\right\| &\leq \left|\left|\left|\Sigma^{1/2-r/2}(\Sigma+\lambda I)^{-1/2}\right|\right|\right| \left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\| \\
&\leq \left(\frac{1}{\lambda}\right)^{\frac{1-(1-r)}{2}} \left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\| \\
&\leq \lambda^{-\frac{r}{2}}\left\|\Sigma^{r/2}(\theta_0-\theta_*)\right\|.
\end{aligned}$$

■

D.2 Theorem 5 and Equation (13)

Theorem 5 and Equation (13) are directly derived from Theorem 2 and Theorem 3, using Lemmas 13 and 14.

To derive corollaries for the optimal γ, one has to find the γ that balances the bias and variance term and to compute the products for such a step size.

D.2.1 Equation (13)

We derive from Theorem 2, when choosing $\gamma = (\lambda n)^{-1}$, and using Lemmas 13 and 14, the following bound, under assumptions of Theorem 2 :

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leq \frac{(18 + \mathrm{Res}(n,b,r,\gamma))\|\Sigma^{r/2}(\theta_0 - \theta_*)\|^2}{(\gamma n)^{\frac{1-r}{2}}} + \frac{6\sigma^2 \operatorname{tr}(\Sigma^b)\gamma^b}{n^{1-b}}.$$

Where $\mathrm{Res}(n,b,r,\gamma) := 3\gamma^{1+b}n^b \operatorname{tr}(\Sigma^b)$ if $-1 \leq r \leq 0$ and $\mathrm{Res}(n,b,r,\gamma) := 0$ if $0 \leq r \leq 1$. When choosing the optimal $\gamma \propto n^{\frac{-b+r}{b+1-r}}$, we have that $\gamma^{1+b}n^b = n^{-1+\frac{1+b}{1+b-r}} = n^{\chi}$, with $\chi = \frac{-r}{1+b-r} \geq 0$ if $r \leq 0$. Thus the residual term is always vanishing for $r \leq 0$ and does not exist for $r \geq 0$.

D.2.2 Theorem 5

Theorem 5 directly follows from Lemmas 13 and 14 and the choice of $\gamma \propto n^{\frac{-2b+2r-1}{b+1-r}}$.

Appendix E. Technical Lemmas

The following sequence of Lemmas appear in the proof. They are mostly independent and rely on simple calculations.

Lemma 15 *The operator* $\big[(\Sigma + \lambda I) \otimes I + I \otimes (\Sigma + \lambda I)\big]^{-1}$ *is a non-decreasing operator on* $(S_n, \preccurlyeq)$.

Proof Lemma means that for two matrices $M, N \in S_n(\mathbb{R})$ such that $M \preccurlyeq N$, then

$$\big[(\Sigma + \lambda I) \otimes I + I \otimes (\Sigma + \lambda I)\big]^{-1} M \preccurlyeq \big[(\Sigma + \lambda I) \otimes I + I \otimes (\Sigma + \lambda I)\big]^{-1} N.$$

It is equivalent to show that for any symmetric positive matrix $A \in S_n^+$,

$$\big[(\Sigma + \lambda I) \otimes I + I \otimes (\Sigma + \lambda I)\big]^{-1} A \in S_n^+(\mathbb{R}).$$

We consider a matrix $A \in S_n^+(\mathbb{R})$. A can be decomposed as a sum of (at most) n rank one matrices $A = \sum_{i=1}^n \omega_i \omega_i^\top$, with $\omega_i \in \mathbb{R}^n$. We thus just have to prove that for some $\omega \in \mathbb{R}^n$, $\big[(\Sigma + \lambda I) \otimes I + I \otimes (\Sigma + \lambda I)\big]^{-1} \omega\omega^\top \in S_n^+(\mathbb{R})$.

Let $\Sigma = \sum_{i \geqslant 0} \mu_i e_i \otimes e_i$ is the eigenvalue decomposition of Σ, then

$$\big[(\Sigma + \lambda I) \otimes I + I \otimes (\Sigma + \lambda I)\big]^{-1} \omega\omega^\top = \sum_{i,j \geqslant 0} \frac{\langle \omega, e_i\rangle\langle \omega, e_j\rangle}{\mu_i + \mu_j + 2\lambda} e_i \otimes e_j.$$

Thus, in the orthonormal basis of eigenvectors, this is thus Hadamard product between

$$\sum_{i,j\geqslant 0} \langle \omega, e_i\rangle\langle \omega, e_j\rangle e_i \otimes e_j = \omega\omega^\top$$

and the matrix $C = \Big(\big(\frac{1}{\mu_i+\mu_j+2\lambda}\big)_{i,j\geqslant 0}\Big)$. Matrix C is a Cauchy matrix and is thus positive. Moreover the Hadamard product of two positive matrices is positive, which concludes the proof. ■

Remark: surprisingly, the inverse operator $(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)$ is not non-decreasing. Indeed, $\preccurlyeq$ is not a total order on S_n so we may have that an operator is non-decreasing and its inverse is not.

Lemma 16 *For all $\rho \in (0,1)$ and $\omega \in [-\pi/2;\pi/2]$ and $r^\pm = \rho(\cos(\omega) \pm \sqrt{-1}\sin(\omega))$ we have:*

$$\left|\frac{1-r^+r^- - \rho^k|A_1|}{|1-r^+|}\right| \leq \min\{1+\rho+e^{-1}+4\rho^k, 2+\rho+\sqrt{5}\rho^{k+1}\} \leq 6. \tag{32}$$

Proof We note that $\rho_i^k A_1$ is a real number as is is a quotient of pure complex numbers, which come from the difference between a complex and its conjugate. We first write A_1 as a combination of sine and cosine functions:

$$\begin{aligned}
\rho_i^k A_1 &= \frac{(r_i^-)^{k+1}(1-r_i^+)^2 - (r_i^+)^{k+1}(1-r_i^-)^2}{r_i^- - r_i^+} \\
&= -\frac{(r_i^-)^{k+1} - (r_i^+)^{k+1} - 2r_i^-r_i^+((r_i^-)^k - (r_i^+)^k) + (r_i^-r_i^+)(r_i^-)^{k-1} - (r_i^+)^{k-1})}{\rho_i \sin\omega_i} \\
&= -\frac{\rho_i^{k+1}\sin((k+1)\omega_i) - 2\rho_i^{k+2}\sin(k\omega_i) + \rho_i^{k+3}\sin((k-1)\omega_i)}{\rho_i\sin\omega_i}.
\end{aligned}$$

This quantity can be simplified when $\rho \to 1$ or $\omega \to 0$. We thus modify the expression of A_1 to make these dependencies clearer:

$$\begin{aligned}
-A_1 &= \frac{\sin((k+1)\omega_i) - 2\rho_i\sin(k\omega_i) + \rho_i^2\sin((k-1)\omega_i)}{\sin\omega_i} \\
&= \frac{(\cos(\omega)-\rho)(\sin(k\omega) - \rho\sin((k-1)\omega)) + \cos(k\omega)\sin(\omega) - \rho\cos((k-1)\omega)\sin(\omega)}{\sin\omega_i} \\
&\text{developing } \sin(a+b) = \sin(a)\cos(b)+\cos(a)\sin(b) \text{ and regrouping terms,} \\
&= \frac{(\cos(\omega)-\rho)^2\sin((k-1)\omega) + (\cos(\omega)-\rho)\sin(\omega)\cos((k-1)\omega) + \cos(k\omega)\sin(\omega)}{\sin\omega_i} \\
&\qquad + \frac{-\rho\cos((k-1)\omega)\sin(\omega)}{\sin\omega_i} \\
&= \frac{(\cos(\omega)-\rho)^2\sin((k-1)\omega)}{\sin\omega_i} + (\cos(\omega)-\rho)\cos((k-1)\omega) + \cos(k\omega) - \rho\cos((k-1)\omega) \\
&\text{simplifying expression, then developing the cosine,} \\
&= \frac{(\cos(\omega)-\rho)^2\sin((k-1)\omega)}{\sin\omega_i} + 2(\cos(\omega)-\rho)\cos((k-1)\omega) + \sin(\omega)\sin((k-1)\omega). \qquad (33)
\end{aligned}$$

So that in that final expression all the terms behave relatively simply when $\rho \to 1$ or $\omega \to 0$. We want to upper bound:

$$\left| \frac{1 - r^+ r^- - \rho^k |A_1|}{|1 - r^+|} \right|.$$

We thus consider separately the first and second term.

$$\frac{1 - r_i^+ r_i^-}{|1 - r_i^+|} = \frac{1 - \rho^2}{|1 - r_i^+|} \leq 1 + \rho \qquad (\text{exact if } \omega = 0).$$

Then, using Equation (33):

$$\frac{-\rho_i^k |A_1|}{|1 - r_i^+|} = \rho^k \frac{\frac{(\cos(\omega) - \rho)^2 \sin((k-1)\omega)}{\sin \omega_i} + 2(\cos(\omega) - \rho)\cos((k-1)\omega) + \sin(\omega)\sin((k-1)\omega)}{\sqrt{(1 - \rho\cos\omega)^2 + \rho^2 \sin^2(\omega)}}.$$

Considering separately the three terms in the numerator, using numerous times that for any $a, b \in [0; 1]$, $|a - b| \leq 1 - ab$:

$$\begin{aligned}
\left| \frac{\rho^k \frac{(\cos(\omega) - \rho)^2 \sin((k-1)\omega)}{\sin \omega_i}}{\sqrt{(1 - \rho\cos\omega)^2 + \rho^2 \sin^2(\omega)}} \right| &\leq \rho^k \frac{(\cos(\omega) - \rho)\sin((k-1)\omega)}{\sin \omega_i} \\
&\qquad \text{as } |(\cos(\omega) - \rho)| \leq 1 - \rho\cos(\omega), \\
&\leq \rho^k \frac{(\cos(\omega) - 1)\sin((k-1)\omega)}{\sin \omega_i} + \rho^k \frac{(1 - \rho)\sin((k-1)\omega)}{\sin \omega_i} \\
&\qquad \text{writing } \cos(\omega) - \rho = \cos(\omega) - 1 + 1 - \rho \\
&\leq \rho^k (1 - \rho)(k - 1) + \rho^k \frac{(\cos(\omega) - 1)\sin((k-1)\omega)}{\sin \omega_i} \\
&\qquad \text{as } |\sin((k-1)\omega)| \leq |(k-1)\sin(\omega)|, \\
&\leq \rho^k (1 - \rho)k - (1 - \rho)\rho^k + \rho^k \frac{(\cos(\omega) - 1)\sin((k-1)\omega)}{\sin \omega_i} \\
&\qquad \text{writing } \cos(\omega) - 1 = 2\sin^2(\omega/2), \\
&\leq \rho^k (1 + (1 - \rho))^k - \rho^k - (1 - \rho)\rho^k + \rho^k \frac{2\sin^2(\omega/2)}{\sin \omega_i} \\
&\qquad \text{using } 1 + (1 - \rho)k \leq (1 + (1 - \rho))^k, \\
&\leq \rho^k (1 + (1 - \rho))^k - \rho^k - (1 - \rho)\rho^k + \rho^k \tan(\omega/2) \\
&\qquad \text{and as } \tan(\omega/2) \leq 1 \text{ for } |\omega| \leq \pi/2, \\
&\leq 1 - (1 - \rho)\rho^k \\
&\qquad \text{using } \rho^k (1 + (1 - \rho))^k = (1 - (1 - \rho)^2)^k \leq 1,
\end{aligned}$$

And for the second and third term:

$$\begin{aligned} 2\left|\rho^k \frac{(\cos(\omega)-\rho)\cos((k-1)\omega)}{\sqrt{(1-\rho\cos\omega)^2+\rho^2\sin^2(\omega)}}\right| &\leq 2\rho^k, \\ \left|\rho^k \frac{+\sin(\omega)\sin((k-1)\omega)}{\sqrt{(1-\rho\cos\omega)^2+\rho^2\sin^2(\omega)}}\right| &\leq \rho^k. \end{aligned}$$

Thus:

$$\left|\frac{1-r_i^+r_i^- - \rho_i^k|A_1|}{|1-r_i^+|}\right| \leq 1+\rho+1+3\rho^k.$$

We also have

$$\begin{aligned} |\rho^k \frac{\frac{(\cos(\omega)-\rho)^2\sin((k-1)\omega)}{\sin\omega_i}}{\sqrt{(1-\rho\cos\omega)^2+\rho^2\sin^2(\omega)}}| &\leq \rho^k \frac{(\cos(\omega)-\rho)\sin((k-1)\omega)}{\sin\omega_i} \\ &\leq \rho^k(1-\rho)(k-1) + \rho^k \frac{(\cos(\omega)-1)\sin((k-1)\omega)}{\sin\omega_i} \\ &\leq (1-\frac{1}{k+1})^{k+1} - (1-\rho)\rho^k + \rho^k \frac{(\cos(\omega)-1)\sin((k-1)\omega)}{\sin\omega_i} \\ &\leq e^{-1} - (1-\rho)\rho^k + \rho^k \frac{\sin^2(\omega/2)}{\sin\omega_i}. \end{aligned}$$

Using that

$$k \sup_{x\in[0;1]} x^k(1-x) = k\frac{1}{k+1}(1-\frac{1}{k+1})^k = (1-\frac{1}{k+1})^{k+1} = \exp((k+1)\ln((1-\frac{1}{k+1})) \leq e^{-1}, \quad (34)$$

we get

$$\left|\frac{1-r_i^+r_i^- - \rho_i^k|A_1|}{|1-r_i^+|}\right| \leq 1+\rho+e^{-1}+4\rho^k$$

We can also change $3\rho^k$ into $\sqrt{5}\rho^k$ We have used that $|(\rho-\cos(\omega))| \leq (1-\rho\cos(\omega))$. ■

Lemma 17 *For any $\rho_i \in (0;1)$, for any $\omega_i \in [-\pi/2;\pi/2]$*

$$\frac{\rho_i^j \sin(\omega_i(j+1))}{\sin(\omega_i)} - \rho_i^{j+1}\frac{\sin(\omega_i j)}{\sin(\omega_i)} \leq 1+e^{-1}.$$

Proof

$$\begin{aligned} \frac{\rho_i^j \sin(\omega_i(j+1))}{\sin(\omega_i)} - \rho_i^{j+1}\frac{\sin(\omega_i j)}{\sin(\omega_i)} &= \rho_i^j\left(\frac{\sin(\omega_i(j+1)) - \rho_i\sin(\omega_i j)}{\sin(\omega_i)}\right) \\ &= \rho_i^j\left(\frac{(\cos(\omega_i)-\rho_i)\sin(\omega_i j)}{\sin(\omega_i)} + \cos(j\omega_i)\right) \\ &\leq \rho_i^j\left((1-\rho_i)j+1\right) \\ &\leq 1+e^{-1} \text{ using (34).} \end{aligned}$$

■

Lemma 18 *For all $\rho \in (0,1)$ and $\omega \in [-\pi/2;\pi/2]$ and $r^{\pm} = \rho(\cos(\omega) \pm \sqrt{-1}\sin(\omega))$ we have:*

$$\left|\rho_i^k B_{1,k}\right| \leq 1.75 \tag{35}$$

Proof Once again, as the considered quantity is real, we first express it as a combination of sine and cosine functions. We then use some simple trigonometric trics to upper bound the quantity.

$$\begin{aligned}
\rho_i^k B_{1,k} &= -\frac{(r_i^-)^{k+1}(1-r_i^+) - (r_i^+)^{k+1}(1-r_i^-)}{r_i^+ - r_i^-} \\
&= -\frac{2\mathfrak{Im}\big[(r_i^-)^{k+1}(1-r_i^+)\big]}{\sqrt{-\Delta_i}} \text{ as it is the difference between a complex and its conjugate,} \\
&= -\frac{\mathfrak{Im}\big[\rho_i^k e^{-(k+1)i\omega_i}(1-\rho_i\cos(\omega_i) - i\rho_i\sin(\omega_i))\big]}{\sin\omega_i\rho_i} \text{ developing the product,} \\
&= \rho_i^k \frac{\cos((k+1)\omega_i)\sin(\omega_i)\rho_i + \sin((k+1)\omega_i)(1-\rho_i\cos(\omega_i))}{\sin\omega_i\rho_i} \\
&= \rho_i^k\Big[\rho_i\cos((k+1)\omega_i) + (1-\rho_i\cos(\omega_i))\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\Big] \text{ and simplifying.}
\end{aligned}$$

Let's turn our interest to the second part of the quantity:

$$\begin{aligned}
\left|\rho_i^k(1-\rho_i\cos(\omega_i))\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\right| &= \left|\rho_i^k(1-\rho_i+\rho_i(1-\cos(\omega_i)))\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\right| \\
&\text{introducing an artificial } +\rho_i - \rho_i, \\
&\leq \rho_i^k\left|(1-\rho_i)\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\right| + \rho_i^k\left|\rho_i(1-\cos(\omega_i))\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\right| \\
&\text{by triangular inequality,} \\
&\leq \rho_i^k\left|(1-\rho_i)(k+1)\right| + \rho_i^k\left|\rho_i\sin^2(\frac{\omega}{2})\frac{1}{2\cos(\frac{\omega}{2})\sin(\frac{\omega}{2})}\right| \\
&\text{using } 1-\cos(\omega_i) = 2\sin^2(\frac{\omega}{2}) \\
&\leq \rho_i^k(1-\rho_i)k + \rho_i^k(1-\rho) + \rho_i^k\left|\rho_i\sin^2(\frac{\omega}{2})\frac{1}{2\cos(\frac{\omega}{2})\sin(\frac{\omega}{2})}\right| \\
&\leq (1-(1-\rho_i))^k(1+(1-\rho_i))^k - \rho_i^k + \frac{1}{2(k+1)} + \rho_i^k\left|\frac{\rho_i}{2}\tan(\frac{\omega}{2})\right| \\
&\leq (1-(1-\rho_i)^2)^k + \frac{1}{4} + \frac{1}{2} \leq 1 + \frac{1}{4} + \frac{1}{2} - \rho_i^k.
\end{aligned}$$

Thus

$$\left|\rho_i^k B_{1,k}\right| = \rho_i^k + 1 + \frac{1}{4} + \frac{1}{2} - \rho_i^k \leq 1 + \frac{1}{4} + \frac{1}{2} = 1.75.$$

■

Lemma 19 *For any $s_i, \gamma, \lambda \in \mathbb{R}_+^3$ such that $\gamma(s_i + \lambda) \leq 1$, for any $k \in \mathbb{N}$, we have the two following highly related identities:*

$$\begin{aligned} 0 \leq 2 - \sqrt{\gamma(s_i+\lambda)} - \big(2 + (k-1)\sqrt{\gamma(s_i+\lambda)}\big)(1 - \sqrt{\gamma(s_i+\lambda)})^k &\leq 2 \\ 0 \leq 1 - (1 + k\sqrt{\gamma(s_i+\lambda)})(1 - \sqrt{\gamma(s_i+\lambda)})^k &\leq 1. \end{aligned}$$

Proof Proof relies on the trick, for any $\alpha \in \mathbb{R}, n \in \mathbb{N}$: $1 + n\alpha \leq (1+\alpha)^n$. For the first one:

$$\begin{aligned} & \sqrt{\gamma(s_i+\lambda)} + \big(2 + (k-1)\sqrt{\gamma(s_i+\lambda)}\big)(1 - \sqrt{\gamma(s_i+\lambda)})^k = \\ = \;& \sqrt{\gamma(s_i+\lambda)} + (1 - \sqrt{\gamma(s_i+\lambda)})^k + \big(1 + (k-1)\sqrt{\gamma(s_i+\lambda)}\big)(1 - \sqrt{\gamma(s_i+\lambda)})^k \\ \leq \;& \sqrt{\gamma(s_i+\lambda)} + (1 - \sqrt{\gamma(s_i+\lambda)}) + \big(1 + (k-1)\sqrt{\gamma(s_i+\lambda)}\big)(1 - \sqrt{\gamma(s_i+\lambda)})^{k-1} \\ \leq \;& 1 + (1 - \gamma(s_i+\lambda))^{k-1} \leq 2. \end{aligned}$$

For the second one:

$$0 \leq (1 + k\sqrt{\gamma(s_i+\lambda)})(1 - \sqrt{\gamma(s_i+\lambda)})^k \leq (1 - \gamma(s_i+\lambda))^k \leq 1.$$

■

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.*, 15(1):595–627, 2014.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

L. Birgé. An Alternative Point of View on Lepski's Method. *Lecture Notes-Monograph Series*, 36:113–133, 2001.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

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

A. Cotter, O. Shamir, N. Srebro, and K. Sridharan. Better mini-batch algorithms via accelerated gradient methods. In *Advances in Neural Information Processing Systems (NIPS)*. 2011.

F. Cucker and S. Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Found. Comput. Math.*, 2:413–418, 2002.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM J. Optim.*, 19(3): 1171–1183, 2008.

E. De Vito, A. Caponetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Found. Comput. Math.*, 5:59–85, 2005.

A. Défossez and F. Bach. Averaged least-mean-squares: bias-variance trade-offs and optimal sampling distributions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics, (AISTATS)*, 2015.

O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *J. Mach. Learn. Res.*, 13:165–202, 2012.

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1-2, Ser. A):37–75, 2014.

A. Dieuleveut and F. Bach. Non-parametric stochastic approximation with large step sizes. *Annals of Statistics*, 44(4):1363–1399, 2015.

M. Duflo. *Random Iterative Models*. Springer, 1997.

H. W. Engl, M. Hanke, and Neubauer A. Regularization of Inverse Problems. *Klüwer Academic Publishers*, 1996.

N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2015.

C. Gu. *Smoothing Spline ANOVA Models*. Springer, 2013.

L. Györfi and H. Walk. On the averaged stochastic approximation for linear regression. *SIAM J. Optim.*, 34(1):31–61, 1996.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2006.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, second edition, 2009.

D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

H. Kushner and G G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

G. Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 133 (1-2, Ser. A):365–397, 2012.

G. Lecué and S. Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.

J. Lin and L. Rosasco. Optimal learning for multi-pass stochastic gradient methods. *arXiv preprint arXiv:1605.08882*, 2016.

P. Massart. *Concentration Inequalities and Model Selection*. Lecture Notes in Mathematics. Springer, 2007.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, second edition, 1989.

A. S. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Y. Nesterov. *Introductory Lectures on Convex Optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

B. O'Donoghue and E. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, pages 1–18, 2013.

R. I. Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probab. Theory Related Fields*, 166(3-4):1175–1194, 2016.

F. Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems (NIPS)*. 2014.

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. {*USSR*} *Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

B. T. Polyak. *Introduction to Optimization*. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York, 1987.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

H. Robbins and S. Monro. A stochastic approxiation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

A. Rudi, R. Camoriano, and L. Rosasco. Less is More: Nyström Computational Regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2009.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Series in Information Science and Statistics. Springer, 2008.

P. Tarrès and Y. Yao. Online learning as stochastic approximation of regularization paths. *EEE Transactions in Information Theory*, (99):5716–5735, 2011.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, 2010.

Y. Yao. *A Dynamic Theory of Learning*. PhD thesis, University of California at Berkeley, 2006.

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

Y. Ying and M. Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 2008.

T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. *Proceedings of the Conference on Machine Learning (ICML)*, 2004.