

Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model

Raphaël Berthier, Francis Bach, Pierre Gaillard

▶ **To cite this version:**

Raphaël Berthier, Francis Bach, Pierre Gaillard. Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model. NeurIPS '20 - 34th International Conference on Neural Information Processing Systems, Dec 2020, Vancouver, Canada. pp.2576–2586. hal-02866755v2

HAL Id: hal-02866755
https://hal.science/hal-02866755v2

Submitted on 26 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model

Raphaël Berthier
INRIA, Ecole Normale Supérieure
PSL Research University, Paris, France
raphael.berthier@inria.fr

Francis Bach
INRIA, Ecole Normale Supérieure
PSL Research University, Paris, France
francis.bach@inria.fr

Pierre Gaillard
INRIA, Ecole Normale Supérieure
PSL Research University, Paris, France
pierre.gaillard@inria.fr

Abstract

In the context of statistical supervised learning, the noiseless linear model assumes that there exists a deterministic linear relation $Y = \langle \theta_*, \Phi(U) \rangle$ between the random output Y and the random feature vector $\Phi(U)$, a potentially non-linear transformation of the inputs U. We analyze the convergence of single-pass, fixed step-size stochastic gradient descent on the least-square risk under this model. The convergence of the iterates to the optimum θ_* and the decay of the generalization error follow polynomial convergence rates with exponents that both depend on the regularities of the optimum θ_* and of the feature vectors $\Phi(U)$. We interpret our result in the reproducing kernel Hilbert space framework. As a special case, we analyze an online algorithm for estimating a real function on the unit hypercube from the noiseless observation of its value at randomly sampled points; the convergence depends on the Sobolev smoothness of the function and of a chosen kernel. Finally, we apply our analysis beyond the supervised learning setting to obtain convergence rates for the averaging process (a.k.a. gossip algorithm) on a graph depending on its spectral dimension.

1 Introduction

Linear regression is widely used in statistical supervised learning, sometimes in the implicit form of kernel regression. A large theory describes the performance (reconstruction and generalization errors) of various algorithms (penalized least-squares, stochastic gradient descent, ...) under various data models (e.g., noisy or noiseless linear model) and the corresponding minimax bounds. In nonparametric estimation theory, one seeks bounds independent of the dimension of the underlying feature space [16, 30]: these bounds describe best the observed behavior in many modern linear regressions, where the data are inherently high-dimensional or where a kernel associated to a high-dimensional feature map is used. In this paper, we provide nonparametric bounds for stochastic gradient descent under the noiseless linear model and under small perturbations of this model.

Under the noiseless linear model, we assume that there exists a ground-truth linear relation $Y = \langle \theta_*, X \rangle$ between the feature vector X and the output $Y \in \mathbb{R}$. The feature vector X may be itself a non-linear transformation of the inputs U, explicitly computed through a feature map $X = \Phi(U)$ or implicitly defined through a positive-definite kernel $k(U, U')$ [17]. The noiseless linear model

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

assumes that there exists a linear predictor in feature space with zero generalization error. The difficulty to approximate this optimal prediction rule θ_* from independent identically distributed (i.i.d.) samples $(X_1, Y_1), \dots, (X_n, Y_n)$ depends on some measure of the complexity of θ_*.

The noiseless assumption is relevant for some basic vision or sound recognition tasks, where there is no ambiguity of the output Y given the input U, but the rule determining the output from the input can be complex. An example from [18, Section 6] is the classification of images of cats versus dogs. For typical images, the output is unambiguous; humans indeed achieve a near-zero error. In sound recognition, one could think of the recovery of the melody from a tune, an unambiguous (but tremendously complex!) task.

Note that in the noiseless model, there is still the randomness of the sampling of $X_1, \dots, X_n$, sometimes called multiplicative noise because algorithms end up multiplying random matrices [13]. Given those inputs, the outputs $Y_1, \dots, Y_n$ are deterministic: there is no additive noise, and thus the noiseless linear model we consider in this paper is a simplification of problems with low additive noise.

The large dimension and number of samples in modern datasets motivate the use of first-order online methods [8, 7]. We study the archetype of these methods: single-pass, constant step-size stochastic gradient descent with no regularization, referred to as simply "SGD" in the following.

Contributions. Our theoretical results and simulation agree to the following: under the noiseless linear model, the iterates of SGD converge to the optimum θ_* and the generalization error of SGD vanishes as the number of samples increases. Moreover, the convergence rate of SGD is determined by the minimum of two parameters: the regularity of the optimum θ_* and the regularity of the feature vectors X, where regularities are measured in terms of power norms of the covariance matrix $\Sigma = \mathbb{E}[X \otimes X]$, see Section 2 for precise definitions and statements. Our analysis of the convergence is tight as we prove upper and lower bounds on the performance of SGD that almost match. Thus SGD shows some adaptivity to the complexity of the problem. In Section 4, we study the robustness of our results when the noiseless linear assumption does not hold, but the generalization error of the optimal linear regression is small. We prove that the asymptotic generalization error of SGD deteriorates by a constant factor, proportional to this optimal generalization error.

Two extensions of our results are studied. First, in Section 3.1, the extension to kernel regression is derived, with, as a special case, the application to the interpolation of a real function f_* on the torus $[0, 1]^d$ from the observation of its value at randomly uniformly sampled points. In the latter case, we show that the rate of convergence depends on the Sobolev smoothness of the function f_* and of the interpolating kernel. Second, beyond supervised learning, our abstract result can be seen as a result on products of i.i.d. linear operators on a Hilbert space. In Section 3.2, we use this result to study a linear stochastic process, the averaging process on a graph, which models a key algorithmic step in decentralized optimization, the gossip algorithm [28, 24]. We prove polynomial convergence rates depending on the spectral dimension of the graph. Finally, in Section 3.3, a toy application instantiates our results in the special case of Gaussian features.

Comparison to the existing literature on linear / kernel regression. There is an extensive research on the performance of different estimators in nonparametric supervised learning, however almost all of them do not consider the special case of the noiseless linear model [16, 9, 30, 15]. The difference is significant; for instance, rates faster than $O(n^{-1})$ for the least-square risk are impossible with additive noise, while in this paper we prove that SGD can converge with arbitrarily fast polynomial rates. Some of these works analyse the performance of SGD [34, 4, 29, 26, 12, 13, 20, 25, 23]. However, because of the additive noise of the data, convergence requires averaging or decaying step sizes. As a notable exception, [18] studies a variant of kernel regularized least-squares and notices that the rate of convergence improves on noiseless data compared to noisy data. However, their rates are not directly comparable to ours as they assume that the optimal predictor is outside of the kernel space while we focus in Section 3.1 on the attainable case where the optimal predictor is in this space. We make a more precise comparison of this work with our results in Remark 2.

While our work focuses on the test error, a recent trend studies the ability of SGD to reach zero training error in the so-called "interpolation regime", that is in over-parametrized models where a perfect fit on the training data is possible [27, 21, 10]. Even with a fixed step size, SGD is shown to achieve zero-training error. However, these results are significantly different from ours: zero training error does not give any information on the generalization ability of the learned models, and

the "interpolation regime" does not imply the noiseless model. The authors of [31] study a mixed framework that includes both the interpolation regime and the noiseless model, depending on whether SGD is seen as a stochastic algorithm minimizing the generalization error or the training error. An acceleration of SGD is studied, depending on the convexity property of the loss, but not on the nonparametric regularity of the problem.

The field of scattered data approximation [33] studies the estimation of a function from the observation of its values at (possibly random) points, considered in Section 3.1. Again, most of the work focuses on the case where the observation of the values is noisy. We found two exceptions that consider the noiseless case. In [5], a minimax rate of $\Omega((\log n/n)^{p/d})$ is shown for estimating a p-smooth function on $[0,1]^d$ in L^∞ norm using n independent uniformly distributed points; the minimax rate is reached with a spline estimate. In [19], a minimax rate of $\Omega(1/n^p)$ in shown for the same problem, but in the special case of $d=1$ and estimation in L^1 norm; the minimax rate is reached with some nearest neighbor polynomial interpolation. Our results are not rigorously comparable with these as we consider the approximation in L^2 norm and a definition of smoothness different from theirs. However, roughly speaking, our convergence rate in Section 3.1 of $\Omega(1/n^{1-d/(2p+d)})$ when $p > d/2$ is much slower than theirs. Note that previous estimators could not be computed in an online fashion, and thus have a significantly larger running time. But in general, this suggests that SGD might not achieve the nonparametric minimax rates under the noiseless linear model.

2 Linear regression

2.1 Setting and main results

We consider the regression problem of learning the linear relationship between a random feature variable $X \in \mathcal{H}$ and a random output variable $Y \in \mathbb{R}$. The feature space $\mathcal{H}$ is assumed to be a Hilbert space with scalar product $\langle .,. \rangle$ and norm $\|.\|$. We assume a noiseless linear model: there exists $\theta_* \in \mathcal{H}$ such that $Y = \langle \theta_*, X \rangle$ almost surely (a.s.). In the online regression setting, we learn θ_* from i.i.d. observations $(X_1, Y_1), (X_2, Y_2), \dots$ of (X, Y). SGD proceeds as follows: it starts with the non-informative initialization $\theta_0 = 0$ and at iteration n, with current estimate θ_{n-1}, it estimates the risk function on the observation (X_n, Y_n), $\mathcal{R}_n(\theta) = \left(\langle \theta, X_n \rangle - Y_n\right)^2/2$ and it performs one step of gradient descent on $\mathcal{R}_n$:

$$\begin{aligned}\theta_n &= \theta_{n-1} - \gamma \nabla \mathcal{R}_n(\theta_{n-1}) \\ &= \theta_{n-1} - \gamma \left(\langle \theta_{n-1}, X_n \rangle - Y_n\right) X_n \\ &= \theta_{n-1} - \gamma \left\langle \theta_{n-1} - \theta_*, X_n \right\rangle X_n \,. \end{aligned} \tag{1}$$

The risk $\mathcal{R}_n(\theta)$ is an unbiased estimate of the population risk, also called generalization error,

$$\mathcal{R}(\theta) = \frac{1}{2}\mathbb{E}\left[\left(\langle \theta, X \rangle - Y\right)^2\right] = \frac{1}{2}\mathbb{E}\left[\left\langle \theta - \theta_*, X \right\rangle^2\right] .$$

We assume the feature variable to be uniformly bounded, namely that there exists a constant $R_0 < \infty$ such that

$$\|X\|^2 \leqslant R_0 \qquad \text{a.s.} \tag{2}$$

We can then define the covariance operator $\Sigma = \mathbb{E}\left[X \otimes X\right]$ of X, where if $x \in \mathcal{H}$, $x \otimes x$ is the bounded linear operator $\theta \in \mathcal{H} \mapsto \langle \theta, x \rangle x$. Finally, note that,

$$\mathcal{R}(\theta) = \frac{1}{2}\left\langle \theta - \theta_*, \Sigma\left(\theta - \theta_*\right)\right\rangle .$$

We do not assume that the linear operator Σ is inversible as this is incompatible in infinite dimension with the boundedness assumption in Eq. (2). Throughout this paper, we use the following convenient notation: if α is a positive real and θ a vector, $\left\|\Sigma^{-\alpha/2}\theta\right\|^2 = \langle \theta, \Sigma^{-\alpha}\theta \rangle := \inf\left\{\|\theta'\|^2 \,\middle|\, \theta' \text{ such that } \theta = \Sigma^{\alpha/2}\theta'\right\}$, with the convention that it is equal to ∞ when $\theta \notin \Sigma^{\alpha/2}(\mathcal{H})$. We have two theorems (upper and lower bounds) showing tight convergence rates for SGD.

Theorem 1 (upper bound)**.** *Assume that there exists a non-negative real number $\underline{\alpha}$ such that*

- *(a) (regularity of the optimum) $\theta_* \in \Sigma^{\underline{\alpha}/2}(\mathcal{H})$, i.e., $\|\Sigma^{-\underline{\alpha}/2}\theta_*\| < \infty$, and*
- *(b) (regularity of the feature vector) $X \in \Sigma^{\underline{\alpha}/2}(\mathcal{H})$ a.s., and there exists a constant $R_{\underline{\alpha}} < \infty$ such that $\|\Sigma^{-\underline{\alpha}/2}X\|^2 \leqslant R_{\underline{\alpha}}$ a.s.*

Assume further $0 < \gamma \leqslant 1/R_0$. *The iterates* θ_n *of SGD with step-size* γ *satisfy for all* $n \geqslant 1$,

1. *(reconstruction error)* $\mathbb{E}\left[\|\theta_n - \theta_*\|^2\right] \leqslant \dfrac{C}{n^{\underline{\alpha}}}$,
2. *(generalization error)* $\min_{k=0,\dots,n} \mathbb{E}\left[\mathcal{R}(\theta_k)\right] \leqslant \dfrac{C'}{n^{\underline{\alpha}+1}}$,

where $C = \dfrac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}}}\left(\|\Sigma^{-\underline{\alpha}/2}\theta_*\|^2 + \dfrac{R_{\underline{\alpha}}}{R_0}\|\theta_*\|^2\right)$ *and* $C' = 2^{\underline{\alpha}}\dfrac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}+1}}\left(\|\Sigma^{-\underline{\alpha}/2}\theta_*\|^2 + \dfrac{R_{\underline{\alpha}}}{R_0}\|\theta_*\|^2\right)$.

Assumption (a) is classical in the non-parametric kernel literature [9]: it is often called *complexity of the optimum*, or *source condition*. Assumption (b) is made in [25]. It implies that

$$\mathrm{Tr}(\Sigma^{1-\underline{\alpha}}) = \mathbb{E}[\mathrm{Tr}(XX^T\Sigma^{-\underline{\alpha}})] = \mathbb{E}[X^T\Sigma^{-\underline{\alpha}}X] \leqslant R_{\underline{\alpha}}\,.$$

This last condition, called *capacity condition* [25], is sometimes stated under the form of a given decay of the eigenvalues of Σ; it is related to the *effective dimension* of the problem [9].

Theorem 2 (lower bound)**.** *Assume that there exists a positive real number* $\overline{\alpha}$ *such that one of the two following conditions holds:*

(a) (irregularity of the optimum) $\theta_* \notin \Sigma^{\overline{\alpha}/2}(\mathcal{H})$, *i.e.,* $\|\Sigma^{-\overline{\alpha}/2}\theta_*\| = \infty$, *or*
(b) (irregularity of the feature vector) with positive probability, $X \notin \Sigma^{\overline{\alpha}/2}(\mathcal{H})$ *and* $\langle X, \theta_*\rangle \neq 0$.

Assume further $0 < \gamma \leqslant 1/R_0$. *The iterates* θ_n *of SGD with step-size* γ *satisfy for all* $\varepsilon > 0$,

1. *(reconstruction error)* $\mathbb{E}\left[\|\theta_n - \theta_*\|^2\right]$ *is not asymptotically dominated by* $1/n^{\overline{\alpha}+\varepsilon}$,
2. *(generalization error)* $\mathbb{E}\left[\mathcal{R}(\theta_n)\right]$ *is not asymptotically dominated by* $1/n^{\overline{\alpha}+1+\varepsilon}$.

The take-home message of Theorems 1, 2 is that the convergence rate of SGD is governed by two real numbers: the regularity α_1 of the optimum, that is the supremum of all $\underline{\alpha}$ such that $\theta_* \in \Sigma^{\underline{\alpha}/2}(\mathcal{H})$, and the regularity α_2 of the features, that is the supremum of all $\underline{\alpha}$ such that $X \in \Sigma^{\underline{\alpha}/2}(\mathcal{H})$ almost surely. The polynomial convergence rate of SGD is roughly of the order of $n^{-\alpha}$ for the reconstruction error and $n^{-\alpha-1}$ for the generalization error with $\alpha = \min(\alpha_1, \alpha_2)$: one of the two regularities is a bottleneck for fast convergence. See Section 3.1 for an application to the optimal choice of a reproducing kernel Hilbert space. The exponent α_1 corresponds to the decay of the errors of the gradient descent on the population risk $\mathcal{R}$. However, due to the multiplicative noise, the convergence of SGD is slowed down by the irregularity of the feature vectors if $\alpha_2 < \alpha_1$.

In the theorems, the constraint on the step-size $0 < \gamma \leqslant 1/R_0$ is independent of the time horizon n and of the regularities α_1, α_2. Thus fixed step-size SGD shows some adaptivity to the regularity of the problem.

In Section 3, we give extensive numerical evidence that the polynomial rates $n^{-\alpha}$ and $n^{-(\alpha+1)}$ in the bounds are indeed sharp in describing convergence rate of SGD.

We end this section with a few remarks on Theorems 1, 2. They articulate the significance of the results, but are non-essential to the rest of this paper.

Remark 1. *Our upper bound and lower bound on the generalization errors do not match exactly. Indeed, we prove an upper bound on the* minimum *risk of the past iterates, where we prove a lower bound on a larger quantity, the risk of the* last *iterate. To the best of our knowledge, it is an open question whether one can prove an upper bound for the last iterate under our assumptions: more precisely, does* $\mathbb{E}\left[\mathcal{R}(\theta_n)\right] \leqslant C''/n^{\underline{\alpha}+1}$ *hold for some constant* C''*?*

Remark 2 (related literature)**.** *In the case* $\underline{\alpha} = 0$, *where no regularity assumption is made on the optimum or the features (apart from being bounded), we upper-bound* $\min_{k=1,\dots,n} \mathbb{E}\left[\mathcal{R}(\theta_k)\right]$ *by* $O(n^{-1})$. *A similar result was shown in [4]: the excess risk for averaged constant-step size SGD is asymptotically dominated by* n^{-1} *on any least-squares problem–not necessarily a noiseless one. It is remarkable that under the noiseless linear setting, no averaging or decay of the step-size is needed to obtain the same convergence rate.*

The article [18] also studies the performance of an algorithm, a variant of kernel regularized least-squares, in the noiseless non-parametric setting. However, they do not exploit when the function is more regular than being in the kernel space, i.e., when $\alpha_1 > 0$ *with our notation,* $\beta > 1/2$ *with*

theirs. In fact, they leave this case as an open problem in their Section 6. Thus, a fair comparison can only be made when $\alpha_1 = 0, \beta = 1/2$. In this case, SGD and the algorithm of [18] both achieve the same rate $O(n^{-1})$.

Remark 3. *The theorems stated above stay true if one weakens the assumptions in the following way, where $\preccurlyeq$ denotes the semi-definite order:*

- *assume $\mathbb{E}\left[\|X\|^2 X \otimes X\right] \preccurlyeq R_0 \Sigma$ instead of $\|X\|^2 \leqslant R_0$ a.s., and*
- *assume $\mathbb{E}\left[\langle X, \Sigma^{-\underline{\alpha}} X\rangle\, X \otimes X\right] \preccurlyeq R_{\underline{\alpha}} \Sigma$ instead of $\langle X, \Sigma^{-\underline{\alpha}} X\rangle \leqslant R_{\underline{\alpha}}$ a.s.*

This weaker set of assumptions is useful in the case of non-bounded features, like the Gaussian features of Appendix 3.3. We thus take special care in using only these weaker assumptions in the proofs of Theorems 1, 2, 3 and 4. However we prefer stating results with the stronger assumptions for the sake of clarity.

Remark 4 (Application of Theorem 1 in finite dimension)**.** *If $\mathcal{H}$ is finite-dimensional and Σ is of full rank, the assumptions of Theorem 1 hold for any $\underline{\alpha} \geqslant 0$. Thus SGD converges faster than any polynomial; in fact one can check that an exponential upper bounds on the reconstruction and generalization errors of the form $C'' \exp(-\lambda_{\min}(\Sigma) t)$ hold, where $\lambda_{\min}(\Sigma)$ is the smallest eigenvalue of Σ. Although the latter bound is asymptotically better than polynomial rates, for moderate time scales the polynomial rates may describe best the observed behavior; for an illustration of this fact on the averaging process, see Section 3.2 and in particular the discussion following Corollary 1.*

Theorems 1 and 2 are extended in the next section and proved in Appendices A and B respectively. The generalization of Theorem 1 beyond the noiseless linear model is exposed in Section 4. The reader interested mostly by applications of Theorems 1 and 2 can jump directly to Section 3.

2.2 Regularity functions and general results

The main difficulty in the proof of Theorems 1 and 2 is that deriving closed recurrence relations for the expected reconstruction and generalization errors is not straightforward. In this paper, we propose to study the norm of $\theta_n - \theta_*$ associated to different powers of the covariance Σ. More precisely, define

$$\varphi_n(\beta) = \mathbb{E}\left[\left\langle \theta_n - \theta_*, \Sigma^{-\beta}\left(\theta_n - \theta_*\right)\right\rangle\right] \in [0, \infty]\,, \qquad \beta \in \mathbb{R}\,. \tag{3}$$

We call φ_n the regularity function at iteration n. In particular,

$$\varphi_n(0) = \mathbb{E}[\|\theta_n - \theta_*\|^2] \qquad \text{and} \qquad \varphi_n(-1) = 2\mathbb{E}[\mathcal{R}(\theta_n)]\,.$$

The sequence of regularity functions φ_n, $n \geqslant 1$ satisfies a closed recurrence inequality (Property 2 in Appendix A) which is central to our proof strategy. Theorems 1 and 2 can be extended to the following estimates on the regularity functions $\varphi_n(\beta)$ on the full interval $\beta \in [-1, \overline{\alpha}]$ (see proofs in Appendices A and B respectively). .

Theorem 3 (upper bound)**.** *Under the assumptions of Theorem 1, we have for all $n \geqslant 1$,*

1. *for all $\beta \in [0, \underline{\alpha}]$,* $\displaystyle \varphi_n(\beta) \leqslant \frac{C}{n^{\underline{\alpha}-\beta}}\,,$
2. *for all $\beta \in [-1, 0)$,* $\displaystyle \min_{k=0,\dots,n} \varphi_k(\beta) \leqslant \frac{C'}{n^{\underline{\alpha}-\beta}}\,,$

where $C = \dfrac{\underline{\alpha}^{\underline{\alpha}-\beta}}{\gamma^{\underline{\alpha}-\beta}}\left(\|\Sigma^{-\underline{\alpha}/2}\theta_*\|^2 + \dfrac{R_{\underline{\alpha}}}{R_0}\|\theta_*\|^2\right)$, $C' = 2^{\underline{\alpha}-\beta}\dfrac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}-\beta}}\left(\|\Sigma^{-\underline{\alpha}/2}\theta_*\|^2 + \dfrac{R_{\underline{\alpha}}}{R_0}\|\theta_*\|^2\right)$.

Theorem 4 (lower bound)**.** *Under the assumptions of Theorem 2, for all $\beta \in [-1, \overline{\alpha}]$, for all $\varepsilon > 0$, $\varphi_n(\beta)$ is not asymptotically dominated by $1/n^{\overline{\alpha}-\beta+\varepsilon}$.*

3 Applications

3.1 Kernel methods and interpolation in Sobolev spaces

A main case of application of our results is the reproducing kernel Hilbert space (RKHS) setting [17]. In this setting, the space $\mathcal{H}$ is typically large or infinite-dimensional, and we do not have a direct access to the feature variable $X \in \mathcal{H}$. Instead, we have access to some random input variable $U \in \mathcal{U}$

such that $X = \Phi(U)$ for some fixed feature map $\Phi : \mathcal{U} \to \mathcal{H}$. It is then natural to associate a vector $\theta \in \mathcal{H}$ with the function $f_\theta \in L^2(\mathcal{U})$ defined by

$$f_\theta(u) = \langle \theta, \Phi(u) \rangle \ .$$

If the positive-definite kernel $k(u, u') = \langle \Phi(u), \Phi(u') \rangle$ can be computed efficiently, SGD can be "kernelized" [34, 29, 26, 12], i.e., the iteration can be written directly in terms of $f_n := f_{\theta_n}$:

$$\begin{aligned} f_n &= f_{n-1} - \gamma(f_{n-1}(U_n) - Y_n)k(U_n, .) \\ &= f_{n-1} - \gamma(f_{n-1} - f_*)(U_n)k(U_n, .) \end{aligned}$$

where $X_n = \Phi(U_n)$ and $f_*(u) := f_{\theta_*}(u) = \langle \theta_*, \Phi(u) \rangle$. Note that in the kernel literature, the mapping $\theta \mapsto f_\theta$ is used to identify $\mathcal{H}$ with a subspace of $L^2(\mathcal{U})$; indeed, if $\Sigma = \mathbb{E}\left[\Phi(U) \otimes \Phi(U)\right]$ has dense range, the mapping is injective. Using this identification, Theorems 1 and 2 can be applied to obtain bounds in the "attainable" case, meaning that the optimal predictor $f_* \in L^2(\mathcal{U})$ is in the RKHS $\mathcal{H}$. This gives decay rates for the RKHS norm $\|f_n - f_*\| := \|\theta_n - \theta_*\|$ which is inherited from $\mathcal{H}$, but also for the population risk $\mathcal{R}(\theta_n)$ which is reinterpreted as the half squared L^2-distance between the associated f_n and the optimal predictor f_*. Indeed,

$$\mathcal{R}(\theta_n) = \frac{1}{2}\mathbb{E}\left[\langle \theta_n - \theta_*, \Phi(U) \rangle^2\right] = \frac{1}{2}\mathbb{E}\left[(f_n(U) - f_*(U))^2\right] = \frac{1}{2}\|f_n - f\|^2_{L^2(U)} \, .$$

Application: interpolation in Sobolev spaces. To illustrate our results, we consider the case where $\mathcal{U}$ is the torus $[0, 1]^d$, U is uniformly distributed on $\mathcal{U}$ and k is a translation-invariant kernel: $k(u, u') = t(u - u')$ where t is a square-integrable 1-periodic function on $[0, 1]^d$. The kernel k is positive-definite if and only if the Fourier transform of t is positive [32]. This imposes, in particular, that t is maximal at 0. Thus the update rule

$$f_n = f_{n-1} - \gamma \left(f_{n-1}(U_n) - f_*(U_n)\right) t(. - U_n) \tag{4}$$

corrects f_n so that the value $f_n(U_n)$ is closer to the observed value $f_*(U_n)$ than $f_{n-1}(U_n)$. Points near U_n are also updated in the same direction, thus the algorithm should converge rapidly if the function f_* is smooth. Our work derives the polynomial convergence rate as a function of the smoothness of f_* and t. The smoothness of functions is measured with the Sobolev spaces H^s_{per}. A function f with Fourier serie $\hat{f}$ belongs to H^s_{per} if

$$\|f\|^2_{H^s_{\text{per}}} = \sum_{k \in \mathbb{Z}^d} |\hat{f}(k)|^2 \left(1 + |k|^2\right)^s < \infty \, .$$

Assume that the Fourier serie of t satisfies a power-law decay: there exists $c, C > 0$ such that:

$$c\left(1 + |k|^2\right)^{-s/2 - d/4} \leqslant \hat{t}(k) \leqslant C\left(1 + |k|^2\right)^{-s/2 - d/4} \ , \qquad k \in \mathbb{Z}^d \, .$$

This condition does not cover C^∞ kernel, including the Gaussian kernel; it is relevant for less regular kernel, that have a power decay in Fourier. This condition is satisfied, for instance, by the Wendland functions [33, Theorem 10.35], or in dimension $d = 1$ by the kernels corresponding to splines of order s, see [32] or [25]. The latter can be computed using the polylogarithm or–for special values of s–the Bernoulli polynomials.

We have $t \in H^{s'}_{\text{per}}$ if and only if $s' < s$, thus s measures the Sobolev smoothness of k. The operator Σ is the convolution with t and thus

$$\|f\|^2 = \left\langle f, \Sigma^{-1} f \right\rangle_{L^2} \asymp \sum_{k \in \mathbb{Z}^d} |\hat{f}(k)|^2 \left(1 + |k|^2\right)^{s/2 + d/4} = \|f\|^2_{H^{s/2+d/4}_{\text{per}}} \, , \tag{5}$$

where $\asymp$ denotes the equality up to positive multiplicative constants. To predict the convergence rate of (4), we check the assumptions of Theorems 1, 2. Computations similar to (5) give

(a) (regularity of the optimum)

$$\left\langle f_*, \Sigma^{-\alpha} f_* \right\rangle \asymp \left\langle f_*, \Sigma^{-\alpha - 1} f_* \right\rangle_{L^2} \asymp \|f_*\|^2_{H^{(s/2+d/4)(\alpha+1)}_{\text{per}}}$$

Assume $f_* \in H^r_{\text{per}}$. We have $\langle f_*, \Sigma^{-\alpha} f_* \rangle < \infty$ if $\alpha \leqslant \frac{2r}{s + d/2} - 1$.

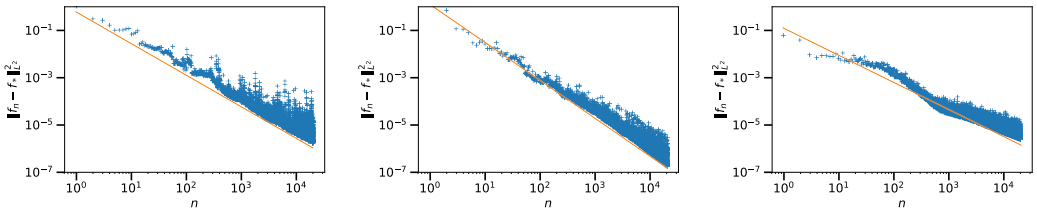


Figure 1: Interpolation of a function of smoothness $r = 2$ using SGD with kernels of smoothness $s = 1$ (left), $s = 2$ (middle) and $s = 3$ (right). Each plot represents one realization of the algorithm (4). The blue crosses represent the square L^2 norms $\|f_n - f_*\|_{L^2}^2$ as a function of the number of iterations n and the orange lines represent the predicted polynomials rates C/n^{α_*+1}, where C is chosen to match best the empirical observations for each plot.

(b) (regularity of the feature vector)

$$\begin{aligned}\left\langle k(u,.), \Sigma^{-\alpha} k(u,.) \right\rangle &= \|k(u,.)\|^2_{H_{\text{per}}^{(s/2+d/4)(\alpha+1)}} = \|t_s\|^2_{H_{\text{per}}^{(s/2+d/4)(\alpha+1)}} \\ &= \sum_{k \in \mathbb{Z}^d} (1+k^2)^{(s/2+d/4)(\alpha-1)} \, .\end{aligned}$$

Thus $\langle k(u,.), \Sigma^{-\alpha} k(u,.) \rangle < \infty$ if and only if $\alpha < 1 - \frac{d}{s+d/2}$.

The regularities of the optimum and of the feature vector are non-negative if the smoothness s of the kernel t satisfies $d/2 < s \leqslant 2r - d/2$, where r is the smoothness of f_*. In this case the polynomial rate of decay of the algorithm is given by the exponent

$$\alpha_* = \min\left(\frac{2r}{s+d/2} - 1, 1 - \frac{d}{s+d/2} \right) . \tag{6}$$

Note that, given a function f_*, this rate is maximal when $s = r$, i.e., the smoothness of the kernel coincides with the smoothness of the function, in which case $\alpha_* = 1 - \frac{d}{r+d/2}$. Theorems 1, 2 give the convergence rates in terms of L^2 norm and RKHS norm, which happens to be a Sobolev norm. The more general Theorems 3 and 4 gives convergence rates in terms of a continuity of fractional Sobolev norms, some weaker and some stronger than the RKHS norm.

In Figure 1, we show the decay of the L^2 norm in the interpolation of a function f_* on $[0,1]$ of smoothness 2 using kernels of smaller, matching and larger smoothness. In each case, the rate predicted by (6) is sharp, and the convergence is indeed fastest when the smoothnesses match.

3.2 Decay rate of the averaging process

The averaging process is a stochastic process on a graph, mostly studied as a model for asynchronous gossip algorithms on networks. Gossip algorithms are subroutines used to diffuse information throughout networks in distributed algorithms [28], in particular in distributed optimization [24].

Let G be a finite undirected connected graph with vertex set $\mathcal{V}$ of cardinality N and edge set $\mathcal{E}$ of cardinality M. The averaging process is a discrete process on functions $x : \mathcal{V} \to \mathbb{R}$ defined as follows. The initial configuration $x_0 = e_{v_\star} : \mathcal{V} \to \mathbb{R}$ is the indicator function of some distinguished vertex $v_\star \in \mathcal{V}$, i.e., $x_0(v_\star) = 1$ and $x_0(v) = 0$ if $v \neq v_\star$. At each iteration, we choose a random edge and replace the values at the ends of the edge by the average of the two current values. In equations, at iterations n, given x_{n-1}, sample an edge $e_n = \{v_n, w_n\}$ uniformly at random from $\mathcal{E}$ and independently from the past, and define

$$x_n(v_n) = x_n(w_n) = \frac{x_{n-1}(v_n) + x_{n-1}(w_n)}{2} \, , \qquad x_n(v) = x_{n-1}(v) \, , \quad v \neq v_n, w_n \, . \tag{7}$$

As the graph is connected, all functions values $x_n(v), v \in \mathcal{V}$ converge to $1/N$ as $n \to \infty$. The study of the averaging process aims at describing how the speed of convergence depends on the graph G.

The averaging process can be seen as a prototype interacting particle system, or finite markov information-exchange process according to Aldous's terminology [1]. However, the linear structure

of the updates of the averaging process makes the analysis simpler than in other interacting particle systems; this property is key in applying the results of Section 2.

In this section, we introduce a quantitive version of the notion of spectral dimension of a graph (see [3] and references therein for other definitions). We use this quantity to build polynomial convergence rates for the expected squared ℓ^2-distance to optimum $\mathbb{E}\big[\sum_{v\in\mathcal{V}}(x_n(v)-1/N)^2\big]$ and for the expected energy $\mathbb{E}\big[\frac{1}{2}\sum_{\{v,w\}\in\mathcal{E}}(x_n(v)-x_n(w))^2\big]$. The comparison with other known convergence bounds is made. We add numerical experiments showing that our bounds describe the observed behavior in some classical large graphs, for an intermediate number of iterations.

Let $L=\sum_{\{v,w\}\in\mathcal{E}}(e_v-e_w)(e_v-e_w)^\top$ be the Laplacian of the graph. It is a positive semi-definite operator. The spectral measure of L at a vertex $v\in\mathcal{V}$ is the unique measure σ_v such that for all continuous real function f,

$$\langle e_v, f(L)e_v\rangle = \int \mathrm{d}\sigma_v(\lambda) f(\lambda)\,.$$

If $0=\lambda_0<\lambda_1\leqslant\ldots\leqslant\lambda_{N-1}$ are the eigenvalues of L and $u_0=\mathbf{1}, u_1,\ldots,u_{N-1}$ are the corresponding normalized eigenvectors, then

$$\sigma_v(\mathrm{d}\lambda)=\sum_{i=0}^{N-1}(u_i(v))^2\delta_{\lambda_i}(\mathrm{d}\lambda)\,.$$

We say that G is of spectral dimension $d\geqslant 0$ with constant $V>0$ if

$$\forall v\in\mathcal{V}\,,\quad \forall E\in(0,\infty)\,,\quad \sigma_v((0,E])\leqslant V^{-1}E^{d/2}\,.$$

A typical example motivating this definition is the following.

Proposition 1. *Let $\mathbb{T}_\Lambda^d$ denote the d-dimensional torus of side length Λ, i.e., the graph with vertex set $\mathcal{V}=(\mathbb{Z}/\Lambda\mathbb{Z})^d$ and edge set $\mathcal{E}=\{\{v,w\}\,|\,v,w\in E,\|v-w\|_2=1\}$. The torus $\mathbb{T}_\Lambda^d$ is of spectral dimension d with some constant $V(d)$ that depends on the dimension d but not on the side length Λ.*

This result is proved in Appendix D. Similar results were proved for supercritical percolation bonds in [22] and for the random geometric graphs in [3].

When the graph is large, the probability of sampling a given edge decays to 0. It is natural to define a rescaled time $t=n/M$ so that the expected number of times a given edge is sampled during a unit time interval does not depend on M (and is equal to 1).

Corollary 1 (of Theorem 1)**.** *Assume that G is of spectral dimension d with constant V, and denote $\delta_{\max}$ the maximal degree of the nodes in the graph. Then, for all $t=n/M\geqslant 2$,*

1. $$\mathbb{E}\left[\sum_{v\in\mathcal{V}}\left(x_{Mt}(v)-\frac{1}{N}\right)^2\right]\leqslant D(d,V,\delta_{\max})\frac{\log t}{t^{d/2}}\,,$$
2. $$\min_{0\leqslant s\leqslant t}\mathbb{E}\left[\frac{1}{2}\sum_{\{v,w\}\in\mathcal{E}}(x_{Ms}(v)-x_{Ms}(w))^2\right]\leqslant D'(d,V,\delta_{\max})\frac{\log t}{t^{d/2+1}}\,,$$

where $D(d,V,\delta_{\max})=\dfrac{2}{\log 2}d^{d/2+1}V^{-1}\delta_{\max}$ and $D'(d,V,\delta_{\max})=\dfrac{2^{d/2+2}}{\log 2}d^{d/2+1}V^{-1}\delta_{\max}$.

See Appendix C for the proof. Note that as G is a finite graph, G can be of any spectral dimension d for some potentially large constant V. However, for many families of graphs of increasing size, such as the toruses $\mathbb{T}_\Lambda^d$, $\Lambda\geqslant 1$, the spectral dimension constant V corresponding to the dimension d and the maximum degree $\delta_{\max}$ remain bounded independently of the size of the graph. In that case, the bounds of Corollary 1 are independent of the size of the graph.

These bounds should be compared to the known exponential convergence bounds of [2] or [28]: they are of the form $O(\exp(-\gamma t))$ where γ is the spectral gap of the Laplacian of the graph, the distance between the two minimal eigenvalues of the Laplacian. Although asymptotically faster, these bounds are only relevant on the typical scale $t\gtrsim 1/\gamma$. In many graphs of interests, the spectral gap γ vanishes as the size of the graph increases; for instance, when $G=\mathbb{T}_\Lambda^d$, γ is of the order of $1/\Lambda^2$. As a consequence, for large graphs and moderate number of iterations, the spectral dimension based

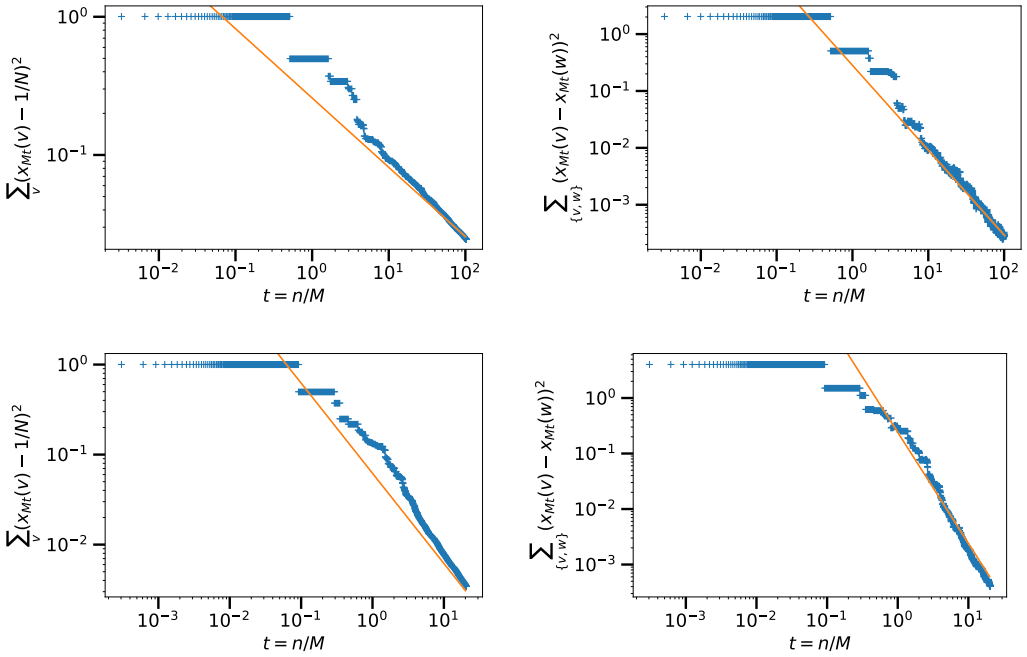


Figure 2: Convergence rates on the circle $\mathbb{T}^1_{300}$ (up) and on the two dimensional torus $\mathbb{T}^2_{40}$ (bottom). The convergence is measured in terms of squared ℓ^2-distance to $\frac{1}{N}\mathbf{1}$ (left) and sum of the squared differences along the edges (right). In orange are the curves of the form $C/n^{d/2}$ and $C'/n^{d/2+1}$ where C and C' are constants chosen to match best the empirical observations for each plot.

bounds describe the observed behavior where spectral gap based bounds do not apply. Indeed, in Figure 2, simulations on a large circle $\mathbb{T}^1_{300}$ and on a large torus $\mathbb{T}^2_{40}$ display polynomial decay rates, with polynomial exponents coinciding with those of the corresponding bounds of Corollary 1. Note that, if pushed on a longer time scale, the simulations would have shown the exponential convergence due to finite graph effects. This incapacity of spectral gap to describe the transient behavior had already motivated the authors of [6] to use the spectral dimension to describe the behavior and to design accelerations of the gossip algorithm. However, the analyses of this paper control only the expected process $\mathbb{E}[x_n]$: the random sampling of the edges is averaged out.

While the polynomial exponents are sharp, we expect the logarithmic factors to be an artifact of the method of proof.

In the case $d = 0$ and $V = 1$, where no assumption on the structure of the graph is made, the fact that the minimal past energy is $O(n^{-1})$ (neglecting the logarithmic factor) has been noticed by Aldous in [2, Proposition 4]. Aldous leaves as an open problem whether one can prove a bound without taking a minimum; this is a special case of our Remark 1.

3.3 Linear regression with Gaussian features

In the setting of Section 2.1, we assume X to be centered Gaussian process of covariance Σ where Σ is a bounded symmetric semidefinite operator. As X is not bounded a.s., we need to use the weaker set of assumptions given in Remark 3. We thus need to compute R_0 such that $\mathbb{E}\left[\|X\|^2 X \otimes X\right] \preccurlyeq R_0\Sigma$ and α, R_α such that $\mathbb{E}\left[\langle X, \Sigma^{-\alpha} X\rangle\, X \otimes X\right] \preccurlyeq R_\alpha \Sigma$. We show here that these conditions are in fact simple trace conditions on Σ, sometimes called *capacity conditions* [25].

Lemma 1. *If $X \sim \mathcal{N}(0, \Sigma)$ and A is a bounded symmetric operator such that $\mathrm{Tr}(\Sigma A) < \infty$,*

$$\mathbb{E}\left[\langle X, AX\rangle\, X \otimes X\right] = 2\Sigma A \Sigma + \mathrm{Tr}(\Sigma A)\Sigma \preccurlyeq \left(2\|\Sigma^{1/2} A \Sigma^{1/2}\|_{\mathcal{H}\to\mathcal{H}} + \mathrm{Tr}(\Sigma A)\right)\Sigma\,.$$

Proof. Diagonalize $\Sigma = \sum_{i\geqslant 1} \lambda_i e_i \otimes e_i$. Then there exists independent standard Gaussian random variables $X_i, i \geqslant 0$ such that $X = \sum_i \lambda_i^{1/2} X_i e_i$.

Let $i, j \geqslant 1$.

$$\langle e_i, \mathbb{E}\left[\langle X, AX\rangle X \otimes X\right] e_j\rangle = \mathbb{E}\left[\langle X, AX\rangle \langle e_i, X \otimes X e_j\rangle\right] = \mathbb{E}\left[\langle X, AX\rangle \lambda_i^{1/2} X_i \lambda_j^{1/2} X_j\right]$$
$$= \lambda_i^{1/2}\lambda_j^{1/2} \sum_{k,l} A_{k,l} \lambda_k^{1/2}\lambda_l^{1/2} \mathbb{E}\left[X_i X_j X_k X_l\right] .$$

As $X_i, i \geqslant 1$ are centered independent random variables, the quantity $\mathbb{E}\left[X_i X_j X_k X_l\right]$ is 0 in many cases. More precisely,

- if $i \neq j$, the general term of the sum in non-zero only when $k = i$ and $l = j$ or $k = j$ and $l = i$. This gives
$$\langle e_i, \mathbb{E}\left[\langle X, AX\rangle X \otimes X\right] e_j\rangle = 2A_{i,j}\lambda_i\lambda_j .$$
- if $i = j$, the general term of the sum is non-zero only when $k = l$. This gives
$$\langle e_i, \mathbb{E}\left[\langle X, AX\rangle X \otimes X\right] e_i\rangle = \lambda_i \sum_k A_{k,k}\lambda_k \mathbb{E}\left[X_i^2 X_k^2\right] = \lambda_i \sum_{k\neq i} A_{k,k}\lambda_k + 3\lambda_i^2 A_{i,i}$$
$$= \lambda_i \sum_k A_{k,k}\lambda_k + 2\lambda_i^2 A_{i,i} .$$

In both cases,

$$\langle e_i, \mathbb{E}\left[\langle X, AX\rangle X \otimes X\right] e_j\rangle = 2\lambda_i\lambda_j A_{i,j} + \left(\sum_k A_{k,k}\lambda_k\right)\lambda_i \mathbf{1}_{i=j} .$$

Note that

$$\mathrm{Tr}(A\Sigma) = \sum_k \langle e_k, \Sigma A e_k\rangle = \sum_k \lambda_k A_{k,k} .$$

Thus we get

$$\langle e_i, \mathbb{E}\left[\langle X, AX\rangle X \otimes X\right] e_j\rangle = 2\lambda_i\lambda_j A_{i,j} + \mathrm{Tr}(A\Sigma)\lambda_i \mathbf{1}_{i=j}$$
$$= 2\langle e_i, \Sigma A \Sigma e_j\rangle + \mathrm{Tr}(A\Sigma)\langle e_i, \Sigma e_j\rangle$$
$$= \langle e_i, \left[2\Sigma A\Sigma + \mathrm{Tr}(\Sigma A)\Sigma\right] e_j\rangle .$$

□

From this lemma with $A = \mathrm{Id}$, we compute $R_0 = 2\|\Sigma\|_{\mathcal{H}\to\mathcal{H}} + \mathrm{Tr}(\Sigma)$, and with $A = \Sigma^{-\alpha}$, we compute $R_\alpha = 2\|\Sigma\|_{\mathcal{H}\to\mathcal{H}}^{1-\alpha} + \mathrm{Tr}(\Sigma^{1-\alpha})$. Thus in the Gaussian case, the condition of (weak) regularity of the features is given by $\mathrm{Tr}(\Sigma^{1-\alpha}) < \infty$.

Simulations. We present simulations in finite but large dimension $d = 10^5$, and we check that dimension-independent bounds describe the observed behavior. We artificially generate regression problems with different regularities by varying the decay of the eigenvalues of the covariance Σ and varying the decay of the coefficients of θ_*.

Choose an orthonormal basis $e_1, \dots, e_d$ of $\mathcal{H}$. We define $\Sigma = \sum_{i=1}^d i^{-\beta} e_i \otimes e_i$ for some $\beta \geqslant 1$ and $\theta_* = \sum_{i=1}^d i^{-\delta} e_i$ for some $\delta \geqslant 1/2$. We now check the condition on α such that the assumptions (a) and (b) are satisfied.

(a) $\langle \theta_*, \Sigma^{-\alpha}\theta_*\rangle = \sum_{i=1}^d \langle \theta_*, e_i\rangle^2 i^{\beta\alpha} = \sum_{i=1}^d i^{-2\delta+\alpha\beta}$, which is bounded independently of the dimension d if and only if $\sum_{i=1}^\infty i^{-2\delta+\alpha\beta} < \infty \Leftrightarrow -2\delta + \alpha\beta < -1 \Leftrightarrow \alpha < \frac{2\delta-1}{\beta}$.

(b) $\mathrm{Tr}(\Sigma^{1-\alpha}) = \sum_{i=1}^d i^{-\beta(1-\alpha)}$, which is bounded independently of the dimension d if and only if $\sum_{i=1}^\infty i^{-\beta(1-\alpha)} < \infty \Leftrightarrow -\beta(1-\alpha) < -1 \Leftrightarrow \alpha < 1 - 1/\beta$.

Thus the corollary gives dimension-independent convergence rates for all $\alpha < \alpha_* = \min\left(1 - \frac{1}{\beta}, \frac{2\delta-1}{\beta}\right)$.

In Figure 3, we show the evolution of $\|\theta_n - \theta_*\|^2$ and $\mathcal{R}(\theta_n)$ for two realizations of SGD. We chose the stepsize $\gamma = 1/R_0 = 1/(2\|\Sigma\|_{\mathcal{H}\to\mathcal{H}} + \mathrm{Tr}(\Sigma))$. The two realizations represent two possible different regimes:

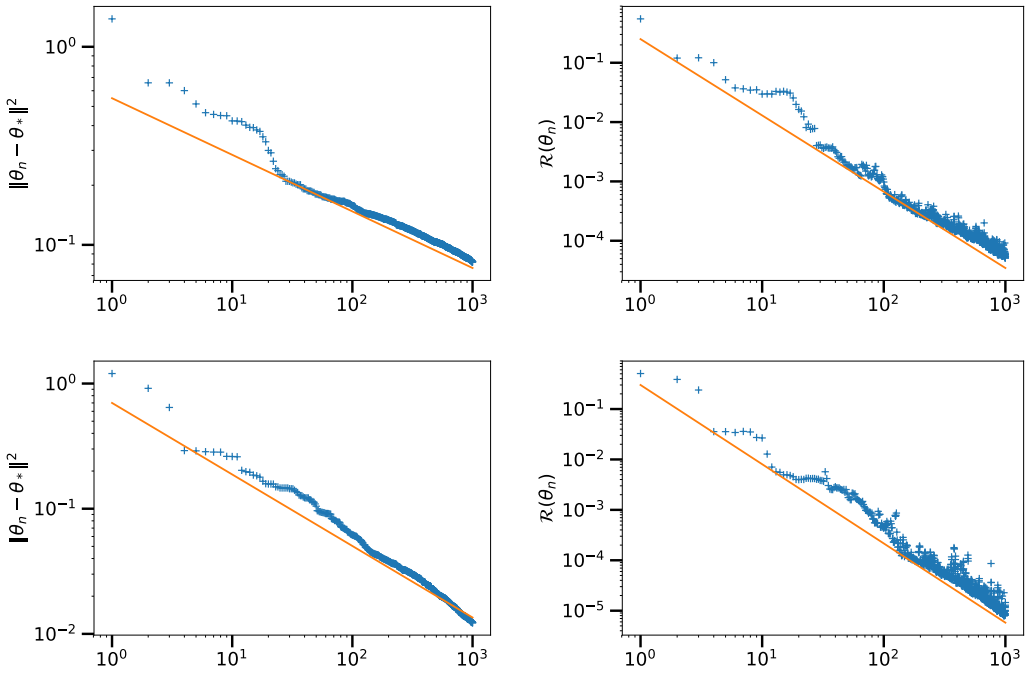


Figure 3: In blue +, evolution of $\|\theta_n - \theta_*\|^2$ (left) and $\mathcal{R}(\theta_n)$ (right) as functions of n, for the problems with parameters $\beta = 1.4, \delta = 1.2$ (up) and $\beta = 3.5, \delta = 1.5$. The orange lines represent the curves D/n^{α_*} (left) and D'/n^{α_*+1} (right).

- In the two upper plots, $\beta = 1.4, \delta = 1.2$. The irregularity of the feature vectors is the bottleneck for fast convergence. We have $\alpha_* = \min\left(1 - \frac{1}{\beta}, \frac{2\delta-1}{\beta}\right) \approx \min(0.29, 1) = 0.29$.
- In the two lower plots, $\beta = 3.5, \delta = 1.5$. The irregularity of the optimum is the bottleneck for fast convergence. We have $\alpha_* = \min\left(1 - \frac{1}{\beta}, \frac{2\delta-1}{\beta}\right) \approx \min(0.71, 0.57) = 0.57$.

We compare with the curves D/n^{α_*} and D'/n^{α_*+1} with hand-tuned constants D and D' to fit best the data for each plot. In both regimes, our theory is sharp in predicting the exponents in the polynomial rates of convergence of $\|\theta_n - \theta_*\|^2$ and $\mathcal{R}(\theta_n)$.

4 Robustness to model mispecification

In this section, we describe how the results of Section 2 are perturbed in the case where a linear relation $Y = \langle \theta_*, X \rangle$ a.s. does not hold. Following the statistical learning framework, we assume a joint law on (X, Y). We further assume that there exists a minimizer $\theta_* \in \mathcal{H}$ of the population risk $\mathcal{R}(\theta)$:

$$\theta_* \in \operatorname*{argmin}_{\theta \in \mathcal{H}} \left\{ \mathcal{R}(\theta) = \frac{1}{2}\mathbb{E}\left[(Y - \langle \theta, X \rangle)^2 \right] \right\} .$$

This general framework encapsulates two types of perturbations of the noiseless linear model:

- (variance) The output Y can be uncertain given X. For instance, under the noisy linear model, $Y = \langle \theta_*, X \rangle + Z$, where Z is centered and independent of X. In this case, $\mathcal{R}(\theta_*) = \mathbb{E}[Z^2] = \mathbb{E}[\mathrm{var}\,(Y|X)]$.
- (bias) Even if Y is deterministic given X, this dependence can be non-linear: $Y = \psi(X)$ for some non-linear function ψ. Then $\mathcal{R}(\theta_*)$ is the squared L^2 distance of the best linear approximation to ψ: $\mathcal{R}(\theta_*) = \frac{1}{2}\mathbb{E}\left[(\psi(X) - \langle \theta_*, X \rangle)^2 \right]$.

In the general framework, the optimal population risk is a combination of both sources

$$\mathcal{R}(\theta_*) = \frac{1}{2}\mathbb{E}\left[\mathrm{var}\left(Y|X\right)\right] + \frac{1}{2}\mathbb{E}\left[\left(\mathbb{E}[Y|X] - \langle \theta_*, X\rangle\right)^2\right] .$$

Given i.i.d. realizations $(X_1, Y_1), (X_2, Y_2), \dots$ of (X, Y), the SGD iterates are defined as

$$\theta_0 = 0\,, \qquad\qquad \theta_n = \theta_{n-1} - \gamma\left(\langle \theta_{n-1}, X_n\rangle - Y_n\right) X_n\,. \tag{8}$$

Apart from the new definition of θ_*, we repeat the same assumptions as in Section 2: let $R_0 < \infty$ be such that $\|X\|^2 \leqslant R_0$ a.s., denote $\Sigma = \mathbb{E}[X \otimes X]$ and $\varphi_n(\beta) = \mathbb{E}\left[\left\langle \theta_n - \theta_*, \Sigma^{-\beta}\left(\theta_n - \theta_*\right)\right\rangle\right]$.

Theorem 5. *Under the assumptions of Theorem 1,*

$$\min_{k=0,\dots,n} \mathbb{E}\left[\mathcal{R}(\theta_k) - \mathcal{R}(\theta_*)\right] \leqslant 2\frac{C'}{n^{\underline{\alpha}+1}} + 2R_0\gamma\mathcal{R}(\theta_*)\,,$$

where C' is the same constant as in Theorem 1.

The take-home message is that if we consider the excess risk $\mathcal{R}(\theta_k) - \mathcal{R}(\theta_*)$, we get the upper bound of the form $2C'n^{-(\underline{\alpha}+1)}$, analog to Theorem 1, but with an additional constant term $2R_0\gamma\mathcal{R}(\theta_*)$. This term can be small if $\mathcal{R}(\theta_*)$ is small, that is if the problem is close to the noiseless linear model, or if the step-size γ is small. In the finite horizon setting setting, one can optimize γ as a function of the scheduled number of steps n in order to balance both terms in the upper bound. As $C' \propto \gamma^{-(\underline{\alpha}+1)}$, the optimal choice is $\gamma \propto n^{-(\underline{\alpha}+1)/(\underline{\alpha}+2)}$ which gives a rate $\min_{k=0,\dots,n}\mathbb{E}\left[\mathcal{R}(\theta_k) - \mathcal{R}(\theta_*)\right] = O\left(n^{-(\underline{\alpha}+1)/(\underline{\alpha}+2)}\right)$.

In the theorem below, we study the SGD iterates θ_n in terms of the power norms $\varphi_n(\beta)$, $\beta \in [-1, \underline{\alpha} - 1]$, in particular in term of the reconstruction error $\varphi_n(0) = \mathbb{E}[\|\theta_n - \theta_*\|^2]$ if $\underline{\alpha} \geqslant 1$. Note that the population risk $\mathcal{R}(\theta)$ is a quadratic with Hessian Σ, minimized at θ_*, thus

$$\mathbb{E}\left[\mathcal{R}(\theta_n) - \mathcal{R}(\theta_*)\right] = \frac{1}{2}\mathbb{E}\left[\langle \theta_n - \theta_*, \Sigma(\theta_n - \theta_*)\rangle\right] = \frac{1}{2}\varphi_n(-1)\,.$$

Thus the theorem below extends Theorem 5.

Theorem 6. *Under the assumptions of Theorem 1,*

1. *for all $\beta \geqslant 0$, $\beta \leqslant \underline{\alpha} - 1$,*

$$\varphi_n(\beta) \leqslant 2\frac{C(\beta)}{n^{\underline{\alpha}-\beta}} + 4R_0^{1-(\beta+1)/\underline{\alpha}} R_{\underline{\alpha}}^{(\beta+1)/\underline{\alpha}}\gamma\mathcal{R}(\theta_*)\,,$$

2. *for all $\beta \in [-1, 0)$, $\beta \leqslant \underline{\alpha} - 1$,*

$$\min_{k01,\dots,n} \varphi_k(\beta) \leqslant 2\frac{C'(\beta)}{n^{\underline{\alpha}-\beta}} + 4R_0^{1-(\beta+1)/\underline{\alpha}} R_{\underline{\alpha}}^{(\beta+1)/\underline{\alpha}}\gamma\mathcal{R}(\theta_*)\,,$$

where C, C' are the same constants as in Theorem 3.

This theorem is proved in Appendix E. We expect the condition $\beta \leqslant \underline{\alpha} - 1$ to be necessary. More precisely, when $\mathcal{R}(\theta_*)$ is positive, we expect the error $\theta_n - \theta_*$ to diverge under the norm $\|\Sigma^{-\beta/2}\,.\,\|$ if $\beta > \underline{\alpha} - 1$. In particular, this would imply that the reconstruction error diverges when $\underline{\alpha} < 1$.

In Figure 4, we show how the simulations of Section 3.3 are perturbed in the presence of additive noise. We consider the noisy linear model $Y = \langle \theta_*, X\,\rangle + \sigma^2 Z$, where $X \sim \mathcal{N}(0, \Sigma)$ and $Z \sim \mathcal{N}(0, 1)$ are independent. As in the previous simulations, we consider the case $\Sigma = \sum_{i=1}^d i^{-\beta} e_i \otimes e_i$ and $\theta_* = \sum_{i=1}^d i^{-\delta} e_i$ with here $d = 10^5$, $\beta = 1.4$, $\delta = 1.2$. In the noiseless case $\sigma^2 = 0$, we have shown that the rate of convergence was given by the polynomial exponent $\alpha_* = \min\left(1 - \frac{1}{\beta}, \frac{2\delta-1}{\beta}\right)$. These predicted rates are represented by the orange lines in the plots. In blue, we show the results of our simulations with some additive noise with variance $\sigma^2 = 2 \times 10^{-4}$. The exponent α_* still describes the behavior of SGD in the initial phase, but in the large n asymptotic the population risk $\mathcal{R}(\theta_n)$ stagnates around the order of σ^2. Both of these qualitative behaviors are predicted by Theorem 5. Moreover, the reconstruction error $\|\theta_n - \theta_*\|$ diverges for large n.

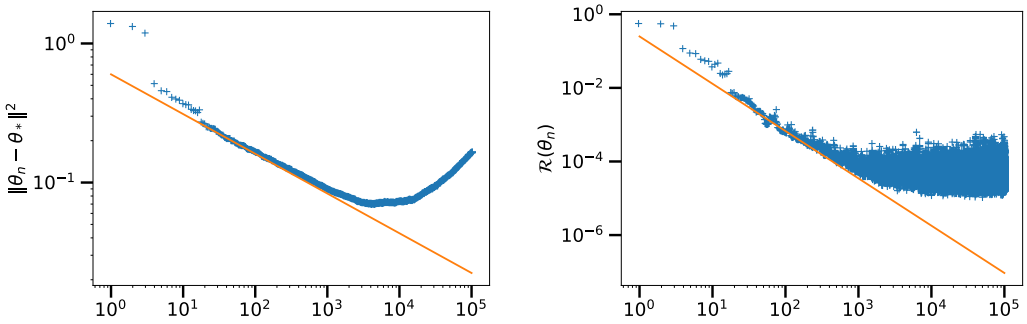


Figure 4: In blue +, evolution of $\|\theta_n - \theta_*\|^2$ (left) and $\mathcal{R}(\theta_n)$ (right) as functions of n, for the problems with parameters $d = 10^5, \beta = 1.4, \delta = 1.2$. The orange lines represent the curves D/n^{α_*} (left) and D'/n^{α_*+1} (right).

5 Conclusion and research directions

In this paper, we give a sharp description of the convergence of SGD under the noiseless linear model and made connexions with the interpolation of a real function and the averaging process. The behavior of SGD is surprisingly different in the absence of additive noise: it converges without any averaging or decay of the step-sizes. To some extent, SGD adapts to the regularity of the problem thanks to the implicit regularization ensured by the initialization at zero and the single pass on the data. However, by comparing with some known estimators for the interpolation of functions [5, 19] (see the end of Section 1), we conjecture that the convergence rate of SGD is suboptimal. What are the minimax rates under the noiseless linear model? Can they be reached with some accelerated online algorithm?

Acknowledgments

This work was greatly improved by detailed comments from Loucas Pillaud-Vivien on earlier versions of the manuscript. We also thank Alessandro Rudi, Nicolas Flammarion and anonymous reviewers for useful discussions. This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grant SEQUOIA 724063) and from the DGA.

References

[1] D. Aldous. Interacting particle systems as stochastic social dynamics. *Bernoulli*, 19(4):1122–1149, 2013.

[2] D. Aldous and D. Lanoue. A lecture on the averaging process. *Probability Surveys*, 9:90–102, 2012.

[3] K. Avrachenkov, L. Cottatellucci, and M. Hamidouche. Eigenvalues and spectral dimension of random geometric graphs in thermodynamic regime. In *International Conference on Complex Networks and Their Applications*, pages 965–975. Springer, 2019.

[4] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

[5] B. Bauer, L. Devroye, M. Kohler, A. Krzyżak, and H. Walk. Nonparametric estimation of a function from noiseless observations at random points. *Journal of Multivariate Analysis*, 160:93–104, 2017.

[6] R. Berthier, F. Bach, and P. Gaillard. Accelerated gossip in networks of given dimension using Jacobi polynomial iterations. *SIAM Journal on Mathematics of Data Science*, 2(1):24–47, 2020.

[7] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems 20*, pages 161–168, 2008.

[8] L. Bottou and Y. Le Cun. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005.

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

[10] V. Cevher and B. C. Vũ. On the linear convergence of the stochastic gradient method with constant step-size. *Optimization Letters*, 13(5):1177–1187, 2019.

[11] F. R. Chung and F. C. Graham. *Spectral Graph Theory*. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Soc., 1997.

[12] A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

[13] A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

[14] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.0.26 of 2020-03-15.

[15] S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv preprint arXiv:1702.07254*, 2017.

[16] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[17] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, pages 1171–1220, 2008.

[18] K.-S. Jun, A. Cutkosky, and F. Orabona. Kernel truncated randomized ridge regression: Optimal rates and low noise acceleration. In *Advances in Neural Information Processing Systems*, pages 15332–15341, 2019.

[19] M. Kohler and A. Krzyżak. Optimal global rates of convergence for interpolation problems with random design. *Statistics & Probability Letters*, 83(8):1871–1879, 2013.

[20] J. Lin and V. Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral-regularization algorithms. *arXiv preprint arXiv:1801.07226*, 2018.

[21] S. Ma, R. Bassily, and M. Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3325–3334, 2018.

[22] P. Mathieu and E. Remy. Isoperimetry and heat kernel decay on percolation clusters. *The Annals of Probability*, 32(1A):100–128, 2004.

[23] N. Mücke, G. Neu, and L. Rosasco. Beating SGD saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, pages 12568–12577, 2019.

[24] A. Nedic, A. Ozdaglar, and P. A. Parrilo. Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, 55(4):922–938, 2010.

[25] L. Pillaud-Vivien, A. Rudi, and F. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018.

[26] L. Rosasco and S. Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

[27] M. Schmidt and N. Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

[28] D. Shah. Gossip algorithms. *Foundations and Trends® in Networking*, 3(1):1–125, 2009.

[29] P. Tarrès and Y. Yao. Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9):5716–5735, 2014.

[30] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

[31] S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *Proceedings of Machine Learning Research*, pages 1195–1204, 2019.

[32] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

[33] H. Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004.

[34] Y. Ying and M. Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

A Proof of Theorems 1 and 3

We recall here the definition of the regularity functions

$$\varphi_n(\beta) = \mathbb{E}\left[\left\langle \theta_n - \theta_*, \Sigma^{-\beta}(\theta_n - \theta_*)\right\rangle\right] \in [0,\infty], \qquad \beta \in \mathbb{R}.$$

A.1 Properties of the regularity functions

We derive here two properties of the sequence of regularity functions $\varphi_n, n \geqslant 1$ that are useful for the proof of Theorem 3. The first one is a simple consequence of the above definition of the regularity function. The second property is the closed recurrence relation of the regularity functions φ_n, $n \geqslant 0$ associated to the iterates of SGD.

Property 1. *For all n, the function φ_n is log-convex, i.e., for all $\beta_1, \beta_2 \in \mathbb{R}$, for all $\lambda \in [0,1]$,*

$$\varphi_n\left((1-\lambda)\beta_1 + \lambda\beta_2\right) \leqslant \varphi_n(\beta_1)^{1-\lambda}\varphi_n(\beta_2)^{\lambda}.$$

Proof. The proof is based on the following lemma, that we state clearly for another use below.

Lemma 2. *Let $\theta \in \mathcal{H}$. Then for all $\beta_1, \beta_2 \in \mathbb{R}$, $\lambda \in [0,1]$,*

$$\left\langle \theta, \Sigma^{-[(1-\lambda)\beta_1+\lambda\beta_2]}\theta\right\rangle \leqslant \left\langle \theta, \Sigma^{-\beta_1}\theta\right\rangle^{1-\lambda}\left\langle \theta, \Sigma^{-\beta_2}\theta\right\rangle^{\lambda}.$$

This lemma follows from Hölder's inequality with $p = (1-\lambda)^{-1}$ and $q = \lambda^{-1}$. Indeed, diagonalize $\Sigma = \sum_i \mu_i e_i \otimes e_i$. Then

$$\begin{aligned}
\left\langle \theta, \Sigma^{-[(1-\lambda)\beta_1+\lambda\beta_2]}\theta\right\rangle &= \sum_i \mu_i^{-[(1-\lambda)\beta_1+\lambda\beta_2]}\langle\theta, e_i\rangle^2 \\
&= \sum_i \left(\mu_i^{-\beta_1}\langle\theta, e_i\rangle^2\right)^{1-\lambda}\left(\mu_i^{-\beta_2}\langle\theta, e_i\rangle^2\right)^{\lambda} \\
&\leqslant \left(\sum_i \mu_i^{-\beta_1}\langle\theta, e_i\rangle^2\right)^{1-\lambda}\left(\sum_i \mu_i^{-\beta_2}\langle\theta, e_i\rangle^2\right)^{\lambda} \\
&= \left\langle \theta, \Sigma^{-\beta_1}\theta\right\rangle^{1-\lambda}\left\langle \theta, \Sigma^{-\beta_2}\theta\right\rangle^{\lambda}.
\end{aligned}$$

We now apply this lemma to prove Property 1.

$$\begin{aligned}
\varphi_n((1-\lambda)\beta_1 + \lambda\beta_2) &= \mathbb{E}\left[\left\langle \theta_n - \theta_*, \Sigma^{-[(1-\lambda)\beta_1+\lambda\beta_2]}(\theta_n - \theta_*)\right\rangle\right] \\
&\leqslant \mathbb{E}\left[\left\langle \theta_n - \theta_*, \Sigma^{-\beta_1}(\theta_n - \theta_*)\right\rangle^{1-\lambda}\left\langle \theta_n - \theta_*, \Sigma^{-\beta_2}(\theta_n - \theta_*)\right\rangle^{\lambda}\right].
\end{aligned}$$

Using again Hölder's inequality, we get

$$\begin{aligned}
\varphi_n((1-\lambda)\beta_1 + \lambda\beta_2) &\leqslant \mathbb{E}\left[\left\langle \theta_n - \theta_*, \Sigma^{-\beta_1}(\theta_n - \theta_*)\right\rangle\right]^{1-\lambda}\mathbb{E}\left[\left\langle \theta_n - \theta_*, \Sigma^{-\beta_2}(\theta_n - \theta_*)\right\rangle\right]^{\lambda} \\
&= \varphi_n(\beta_1)^{1-\lambda}\varphi_n(\beta_2)^{\lambda}.
\end{aligned}$$

□

Property 2. *Under the assumptions of Theorem 3, for all n, the function φ_n is finite on $(-\infty, \underline{\alpha}]$, and if $0 \leqslant \beta \leqslant \underline{\alpha}$,*

$$\varphi_n(\beta) \leqslant \varphi_{n-1}(\beta) - 2\gamma\varphi_{n-1}(\beta-1) + \gamma^2 R_0^{1-\beta/\underline{\alpha}} R_{\underline{\alpha}}^{\beta/\underline{\alpha}}\varphi_{n-1}(-1).$$

Proof. By assumption (a), $\varphi_0(\underline{\alpha}) = \|\Sigma^{-\underline{\alpha}/2}\theta_*\|^2$ is finite, i.e., there exists $\theta \in \mathcal{H}$ such that $\theta_* = \Sigma^{\underline{\alpha}/2}\theta$. Then for any $\beta \leqslant \underline{\alpha}$, $\theta_* = \Sigma^{\beta/2}\left(\Sigma^{(\underline{\alpha}-\beta)/2}\theta\right)$ thus $\varphi_0(\beta) = \|\Sigma^{-\beta/2}\theta_*\|^2$ is finite.

Further, assume that for some n, the function φ_{n-1} is finite on $(\infty, \underline{\alpha}]$. Then we can rewrite the stochastic gradient iteration (1) as

$$\theta_n - \theta_* = (\mathrm{Id} - \gamma X_n \otimes X_n)(\theta_{n-1} - \theta_*).$$

Substituting this expression in the definition of φ_n and expanding the formula, we get

$$\begin{aligned}
\varphi_n(\beta) &= \mathbb{E}\left[\left\langle \theta_n - \theta_*, \Sigma^{-\beta}\left(\theta_n - \theta_*\right)\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle (\mathrm{Id} - \gamma X_n \otimes X_n)(\theta_{n-1} - \theta_*), \Sigma^{-\beta}(\mathrm{Id} - \gamma X_n \otimes X_n)(\theta_{n-1} - \theta_*)\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle \theta_{n-1} - \theta_*, \Sigma^{-\beta}(\theta_{n-1} - \theta_*)\right\rangle\right] && (9) \\
&\qquad - 2\gamma\mathbb{E}\left[\left\langle \theta_{n-1} - \theta_*, \Sigma^{-\beta} X_n \otimes X_n(\theta_{n-1} - \theta_*)\right\rangle\right] && (10) \\
&\qquad + \gamma^2\mathbb{E}\left[\left\langle \theta_{n-1} - \theta_*, X_n \otimes X_n \Sigma^{-\beta} X_n \otimes X_n(\theta_{n-1} - \theta_*)\right\rangle\right] . && (11)
\end{aligned}$$

Note that the first term of this sum is $\varphi_{n-1}(\beta)$. Further, θ_{n-1} is computed using only $(X_1, Y_1), \ldots, (X_{n-1}, Y_{n-1})$, thus it is independent of X_n. It follows that

$$\begin{aligned}
\mathbb{E}\left[\left\langle \theta_{n-1} - \theta_*, \Sigma^{-\beta} X_n \otimes X_n(\theta_{n-1} - \theta_*)\right\rangle\right] &= \mathbb{E}\left[\left\langle \theta_{n-1} - \theta_*, \Sigma^{-\beta}\mathbb{E}\left[X_n \otimes X_n\right](\theta_{n-1} - \theta_*)\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle \theta_{n-1} - \theta_*, \Sigma^{-\beta+1}(\theta_{n-1} - \theta_*)\right\rangle\right] \\
&= \varphi_{n-1}(\beta - 1) . && (12)
\end{aligned}$$

Finally,

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \theta_{n-1} - \theta_*, X_n \otimes X_n \Sigma^{-\beta} X_n \otimes X_n(\theta_{n-1} - \theta_*)\right\rangle\right] && (13) \\
&\quad = \mathbb{E}\left[\left\langle \theta_{n-1} - \theta_*, X_n\right\rangle^2 \left\langle X_n, \Sigma^{-\beta} X_n\right\rangle\right] && (14)
\end{aligned}$$

We now assume that $0 \leqslant \beta \leqslant \underline{\alpha}$. We apply Lemma 2 with $\beta_1 = 0, \beta_2 = \underline{\alpha}, \lambda = \beta/\underline{\alpha}$:

$$\left\langle X_n, \Sigma^{-\beta} X_n\right\rangle \leqslant \|X_n\|^{2(1-\beta/\underline{\alpha})} \left\langle X_n, \Sigma^{-\underline{\alpha}} X_n\right\rangle^{\beta/\underline{\alpha}}$$

Let E_{X_n} denote the expectation with respect to X_n only, while keeping $X_0, \ldots, X_{n-1}$ random. Applying Hölder's inequality, we get

$$\begin{aligned}
&\mathbb{E}_{X_n}\left[\left\langle X_n, \Sigma^{-\beta} X_n\right\rangle \left\langle \theta_{n-1} - \theta_*, X_n\right\rangle^2\right] \\
&\quad \leqslant \mathbb{E}_{X_n}\left[\|X_n\|^{2(1-\beta/\underline{\alpha})} \left\langle X_n, \Sigma^{-\underline{\alpha}} X_n\right\rangle^{\beta/\underline{\alpha}} \left\langle \theta_{n-1} - \theta_*, X_n\right\rangle^2\right] \\
&\quad \leqslant \mathbb{E}_{X_n}\left[\|X_n\|^2 \left\langle \theta_{n-1} - \theta_*, X_n\right\rangle^2\right]^{1-\beta/\underline{\alpha}} \mathbb{E}\left[\left\langle X_n, \Sigma^{-\underline{\alpha}} X_n\right\rangle \left\langle \theta_{n-1} - \theta_*, X_n\right\rangle^2\right]^{\beta/\underline{\alpha}} \\
&\quad = \left\langle \theta_{n-1} - \theta_*, \mathbb{E}\left[\|X_n\|^2 X_n \otimes X_n\right](\theta_{n-1} - \theta_*)\right\rangle^{1-\beta/\underline{\alpha}} \\
&\qquad\qquad \times \left\langle \theta_{n-1} - \theta_*, \mathbb{E}\left[\left\langle X_n, \Sigma^{-\underline{\alpha}} X_n\right\rangle X_n \otimes X_n\right](\theta_{n-1} - \theta_*)\right\rangle^{\beta/\underline{\alpha}} \\
&\quad \leqslant R_0^{1-\beta/\underline{\alpha}} R_{\underline{\alpha}}^{\beta/\underline{\alpha}} \left\langle \theta_{n-1} - \theta_*, \Sigma(\theta_{n-1} - \theta_*)\right\rangle ,
\end{aligned}$$

where in this last step, we use the assumptions that the features X are bounded and regular, in their weak formulation of Remark 3. Returning to the computation of (13)-(14), we get

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \theta_{n-1} - \theta_*, X_n \otimes X_n \Sigma^{-\beta} X_n \otimes X_n(\theta_{n-1} - \theta_*)\right\rangle\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}_{X_n}\left[\left\langle \theta_{n-1} - \theta_*, X_n\right\rangle^2 \left\langle X_n, \Sigma^{-\beta} X_n\right\rangle\right]\right] && (15) \\
&\quad \leqslant R_0^{1-\beta/\underline{\alpha}} R_{\underline{\alpha}}^{\beta/\underline{\alpha}} \mathbb{E}\left[\left\langle \theta_{n-1} - \theta_*, \Sigma(\theta_{n-1} - \theta_*)\right\rangle\right] \\
&\quad = R_0^{1-\beta/\underline{\alpha}} R_{\underline{\alpha}}^{\beta/\underline{\alpha}} \varphi_{n-1}(-1) . && (16)
\end{aligned}$$

The result is obtained by putting together Equations (9)-(11), (12) and (16). □

A.2 Proof of Theorem 1

A remarkable feature of the proof that follows is that only Properties 1 and 2 of the regularity functions are used to derive the theorem. In particular, we do not use the definition of the regularity functions φ_n in this section.

We start with a few preliminary remarks. Using the recurrence Property 2 and that $\gamma R_0 \leqslant 1$,

$$\begin{aligned}
\varphi_k(0) &\leqslant \varphi_{k-1}(0) - \gamma\left(2 - \gamma R_0\right)\varphi_{k-1}(-1) \\
&\leqslant \varphi_{k-1}(0) - \gamma\varphi_{k-1}(-1) .
\end{aligned}$$

Thus the sequence $\varphi_k(0)$, $k \geqslant 0$ decreases, and

$$\gamma \varphi_{k-1}(-1) \leqslant \varphi_{k-1}(0) - \varphi_k(0)\,. \tag{17}$$

By summing this inequality over $k \geqslant 1$, we get

$$\gamma \sum_{k=0}^{\infty} \varphi_k(-1) \leqslant \varphi_0(0)\,. \tag{18}$$

Using again the recurrence Property 2,

$$\begin{aligned} \varphi_k(\underline{\alpha}) &\leqslant \varphi_{k-1}(\underline{\alpha}) - 2\gamma \varphi_{k-1}(\underline{\alpha}-1) + \gamma^2 R_{\underline{\alpha}} \varphi_{k-1}(-1) \\ &\leqslant \varphi_{k-1}(\underline{\alpha}) + \gamma^2 R_{\underline{\alpha}} \varphi_{k-1}(-1)\,. \end{aligned} \tag{19}$$

By summing for $k = 1, \ldots, n$ and using the bound (18),

$$\begin{aligned} \varphi_n(\underline{\alpha}) &\leqslant \varphi_0(\underline{\alpha}) + \gamma^2 R_{\underline{\alpha}} \sum_{k=0}^{n-1} \varphi_k(-1) \\ &\leqslant \varphi_0(\underline{\alpha}) + \gamma R_{\underline{\alpha}} \varphi_0(0) \\ &\leqslant \varphi_0(\underline{\alpha}) + \frac{R_{\underline{\alpha}}}{R_0} \varphi_0(0)\,. \end{aligned} \tag{20}$$

In words, the sequence $\varphi_n(\underline{\alpha})$, $n \geqslant 0$ is bounded by $D := \varphi_0(\underline{\alpha}) + \frac{R_{\underline{\alpha}}}{R_0}\varphi_0(0)$. As a side note, this proves Theorem 3 for $\beta = \underline{\alpha}$.

We can now give a closed recurrence relation $\varphi_k(0)$, $k \geqslant 0$. Using the log-convexity Property 1,

$$\varphi_{k-1}(0) \leqslant \varphi_{k-1}(-1)^{\underline{\alpha}/(\underline{\alpha}+1)} \varphi_{k-1}(\underline{\alpha})^{1/(\underline{\alpha}+1)} \leqslant \varphi_{k-1}(-1)^{\underline{\alpha}/(\underline{\alpha}+1)} D^{1/(\underline{\alpha}+1)}\,.$$

Substituting in (17), we obtain

$$\begin{aligned} \varphi_{k-1}(0) - \varphi_k(0) &\geqslant \gamma \varphi_{k-1}(-1) \\ &\geqslant \gamma D^{-1/\underline{\alpha}} \varphi_{k-1}(0)^{1+1/\underline{\alpha}}\,. \end{aligned}$$

This gives the wanted closed recurrence relation for $\varphi_k(0)$, $k \geqslant 0$. It implies a decay of $\varphi_k(0)$ as follows: consider the real function $f(\varphi) = \frac{1}{\varphi^{1/\underline{\alpha}}}$. It is a convex function on the positive reals, with derivative $f'(\varphi) = -\frac{1}{\underline{\alpha}} \frac{1}{\varphi^{1+1/\underline{\alpha}}}$. Using that a convex function is above its tangents, we obtain

$$\begin{aligned} f\left(\varphi_k(0)\right) - f\left(\varphi_{k-1}(0)\right) &\geqslant f'\left(\varphi_{k-1}(0)\right)\left(\varphi_k(0) - \varphi_{k-1}(0)\right) \\ &= -\frac{1}{\underline{\alpha}} \frac{1}{\varphi_{k-1}(0)^{1+1/\underline{\alpha}}} \left(\varphi_k(0) - \varphi_{k-1}(0)\right) \\ &\geqslant \frac{1}{\underline{\alpha}} \gamma D^{-1/\underline{\alpha}}\,. \end{aligned}$$

By summing this inequality for $k = 1, \ldots, n$, we obtain

$$\frac{1}{\varphi_n(0)^{1/\underline{\alpha}}} = f\left(\varphi_n(0)\right) \geqslant f\left(\varphi_0(0)\right) + \frac{1}{\underline{\alpha}} \gamma D^{-1/\underline{\alpha}} n \geqslant \frac{1}{\underline{\alpha}} \gamma D^{-1/\underline{\alpha}} n\,.$$

This implies conclusion 1 of Theorem 1:

$$\mathbb{E}\left[\|\theta_n - \theta_*\|^2\right] = \varphi_n(0) \leqslant \frac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}}} D \frac{1}{n^{\underline{\alpha}}}\,. \tag{21}$$

Further,

$$\min_{0 \leqslant k \leqslant n} \varphi_k(-1) \leqslant \min_{\lceil n/2 \rceil \leqslant k \leqslant n} \varphi_k(-1) \leqslant \frac{2}{n} \sum_{k=\lceil n/2 \rceil}^{n} \varphi_k(-1) \leqslant \frac{2}{n} \frac{1}{\gamma} \sum_{k=\lceil n/2 \rceil}^{n} \left(\varphi_k(0) - \varphi_{k+1}(0)\right)\,,$$

where in the last step we used (17). Telescoping the sum, we obtain

$$\begin{aligned} \min_{0 \leqslant k \leqslant n} \varphi_k(-1) &\leqslant \min_{\lceil n/2 \rceil \leqslant k \leqslant n} \varphi_k(-1) \leqslant \frac{2}{n} \frac{1}{\gamma} \varphi_{\lceil n/2 \rceil}(0) \\ &\leqslant \frac{2}{n} \frac{1}{\gamma} \frac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}}} D \frac{1}{\lceil n/2 \rceil^{\underline{\alpha}}} \leqslant 2^{\underline{\alpha}+1} \frac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}+1}} D \frac{1}{n^{\underline{\alpha}+1}}\,. \end{aligned} \tag{22}$$

Using that $\varphi_n(-1) = 2\mathbb{E}[\mathcal{R}(\theta_n)]$, this gives conclusion 2 of Theorem 1.

A.3 Proof of Theorem 3

We continue the proof of Theorem 1 to prove Theorem 3. By the log-convexity Property 1, for all $\beta \in [0, \underline{\alpha}]$,

$$\varphi_n(\beta) \leqslant \varphi_n(0)^{1-\beta/\underline{\alpha}} \varphi_n(\underline{\alpha})^{\beta/\underline{\alpha}} .$$

Using Equations (21) and (20), we obtain

$$\varphi_n(\beta) \leqslant \frac{\underline{\alpha}^{\underline{\alpha}-\beta}}{\gamma^{\underline{\alpha}-\beta}} D \frac{1}{n^{\underline{\alpha}-\beta}} .$$

This proves conclusion 1 of the theorem. We now consider the case $\beta \in [-1, 0)$. By the log-convexity Property 1,

$$\min_{0 \leqslant k \leqslant n} \varphi_k(\beta) \leqslant \min_{\lceil n/2 \rceil \leqslant k \leqslant n} \varphi_k(\beta) \leqslant \min_{\lceil n/2 \rceil \leqslant k \leqslant n} \varphi_k(-1)^{-\beta} \varphi_k(0)^{1+\beta}$$

Using that $\varphi_k(0)$, $k \geqslant 0$ is decreasing and the inequality (22), we obtain

$$\begin{aligned}
\min_{\lceil n/2 \rceil \leqslant k \leqslant n} \varphi_k(-1)^{-\beta} \varphi_k(0)^{1+\beta} &\leqslant \varphi_{\lceil n/2 \rceil}(0)^{1+\beta} \left(\min_{\lceil n/2 \rceil \leqslant k \leqslant n} \varphi_k(-1) \right)^{-\beta} \\
&\leqslant \varphi_{\lceil n/2 \rceil}(0)^{1+\beta} \left(\frac{2}{n} \frac{1}{\gamma} \varphi_{\lceil n/2 \rceil}(0) \right)^{-\beta} \\
&\leqslant \frac{2^{-\beta}}{n^{-\beta}} \frac{1}{\gamma^{-\beta}} \varphi_{\lceil n/2 \rceil}(0) .
\end{aligned}$$

Using finally (21), we obtain conclusion 2 of the theorem

$$\min_{0 \leqslant k \leqslant n} \varphi_k(\beta) \leqslant \frac{2^{-\beta}}{n^{-\beta}} \frac{1}{\gamma^{-\beta}} \frac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}}} D \frac{1}{\lceil n/2 \rceil^{\underline{\alpha}}} \leqslant 2^{\underline{\alpha}-\beta} \frac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}-\beta}} D \frac{1}{n^{\underline{\alpha}-\beta}} .$$

B Proof of Theorems 2 and 4

We start in the case (a) where the optimum is irregular: $\theta_* \notin \Sigma^{-\overline{\alpha}/2}(\mathcal{H})$. In that case, we give a lower bound in the convergence rate by studying the expected process $\overline{\theta}_n := \mathbb{E}[\theta_n]$. Indeed, by Jensen's inequality,

$$\varphi_n(\beta) = \mathbb{E}\left[\left\langle \theta_n - \theta_*, \Sigma^{-\beta} (\theta_n - \theta_*) \right\rangle \right] \geqslant \left\langle \overline{\theta}_n - \theta_*, \Sigma^{-\beta} \left(\overline{\theta}_n - \theta_* \right) \right\rangle . \tag{23}$$

The expectation $\overline{\theta}_n$ can be interpreted as the (non-stochastic) gradient descent on the population risk $\mathcal{R}(\theta)$. Indeed, by taking the expectation in (1), we obtain

$$\overline{\theta}_n - \theta_* = (\mathrm{Id} - \gamma\Sigma)(\overline{\theta}_{n-1} - \theta_*) = -(\mathrm{Id} - \gamma\Sigma)^n \theta_* . \tag{24}$$

Note that as $\gamma \leqslant 1/R_0$, $I - \gamma\Sigma$ is a positive definite matrix. Indeed, by the weak definition of R_0 in Remark 3,

$$R_0 \Sigma \succcurlyeq \mathbb{E}\left[\|X\|^2 X \otimes X \right] = \mathbb{E}\left[(X \otimes X)(X \otimes X) \right] \succcurlyeq \mathbb{E}[X \otimes X]^2 = \Sigma^2 ,$$

thus R_0 is larger than the operator norm of Σ. Thus $\gamma\Sigma \preccurlyeq \frac{1}{R_0}\Sigma \preccurlyeq \mathrm{Id}$.

In the following, if $\alpha \in \mathbb{R}$ and $k \in \mathbb{N}$, $\binom{\alpha}{k}$ denotes the generalized binomial coefficient: $\binom{\alpha}{k} = \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}$. Fix now $\alpha \geqslant 0$. We have the (formal) power series

$$\begin{aligned}
(1+x)^{-\alpha} &= \sum_{k=0}^{\infty} \binom{-\alpha}{k} x^k \\
(1-x)^{-\alpha} &= \sum_{k=0}^{\infty} \binom{-\alpha}{k} (-1)^k x^k = \sum_{k=0}^{\infty} \binom{\alpha+k-1}{k} x^k \\
y^{-\alpha} &= \sum_{k=0}^{\infty} \binom{\alpha+k-1}{k} (1-y)^k .
\end{aligned}$$

This last equality holds in $[0,\infty]$ for $y \in [0,1]$. In that case, all terms of the serie are positive, thus the meaning of the sum is unambiguous.

Note that $0 \preccurlyeq \gamma\Sigma \preccurlyeq \mathrm{Id}$, thus we have, formally,

$$\gamma^{-\alpha}\Sigma^{-\alpha} = \sum_{k=0}^{\infty} \binom{\alpha+k-1}{k} (\mathrm{Id} - \gamma\Sigma)^k \,.$$

The rigorous meaning of this equality is that for all $\theta \in \mathcal{H}$,

$$\gamma^{-\alpha}\langle \theta, \Sigma^{-\alpha}\theta\rangle = \sum_{k=0}^{\infty} \binom{\alpha+k-1}{k} \langle \theta, (\mathrm{Id} - \gamma\Sigma)^k \theta\rangle \,.$$

Both terms of the equality can be infinite: here we are using the convention stated in Section 2.1 that implies that $\langle \theta, \Sigma^{-\alpha}\theta\rangle = \infty \Leftrightarrow \theta \notin \Sigma^{\alpha/2}(\mathcal{H})$. In particular, take $\alpha = \overline{\alpha} - \beta$ and $\theta = \Sigma^{-\beta/2}\theta_*$:

$$\begin{aligned}
\infty = \gamma^{\beta-\overline{\alpha}} \left\langle \theta_*, \Sigma^{-\overline{\alpha}}\theta_* \right\rangle &= \sum_{k=0}^{\infty} \binom{\overline{\alpha}-\beta+k-1}{k} \left\langle \theta_*, \Sigma^{-\beta}(\mathrm{Id} - \gamma\Sigma)^k \theta_* \right\rangle \\
&= \sum_{n=0}^{\infty} \left[\binom{\overline{\alpha}-\beta+2n-1}{2n} \left\langle \theta_*, \Sigma^{-\beta}(\mathrm{Id} - \gamma\Sigma)^{2n} \theta_* \right\rangle \right. \\
&\qquad \left. + \binom{\overline{\alpha}-\beta+2n}{2n+1} \left\langle \theta_*, \Sigma^{-\beta}(\mathrm{Id} - \gamma\Sigma)^{2n+1} \theta_* \right\rangle \right] .
\end{aligned}$$

Using that $\binom{\overline{\alpha}-\beta+2n-1}{2n} \leqslant \binom{\overline{\alpha}-\beta+2n}{2n+1}$ and $\left\langle \theta_*, \Sigma^{-\beta}(\mathrm{Id} - \gamma\Sigma)^{2n} \theta_* \right\rangle \geqslant \left\langle \theta_*, \Sigma^{-\beta}(\mathrm{Id} - \gamma\Sigma)^{2n+1} \theta_* \right\rangle$ and then (24), (23),

$$\begin{aligned}
\infty &\leqslant 2 \sum_{n=0}^{\infty} \binom{\overline{\alpha}-\beta+2n}{2n+1} \left\langle \theta_*, \Sigma^{-\beta}(\mathrm{Id} - \gamma\Sigma)^{2n} \theta_* \right\rangle \\
&= 2 \sum_{n=0}^{\infty} \binom{\overline{\alpha}-\beta+2n}{2n+1} \left\langle \overline{\theta}_n - \theta_*, \Sigma^{-\beta}(\overline{\theta}_n - \theta_*) \right\rangle \\
&\leqslant 2 \sum_{n=0}^{\infty} \binom{\overline{\alpha}-\beta+2n}{2n+1} \varphi_n(\beta) \,.
\end{aligned}$$

From [14, Equation 5.8.1], we have the formula $\Gamma(z) = \lim_{k\to\infty} \frac{k! k^z}{z(z+1)\cdots(z+k)}$ where Γ denotes the Gamma function. Thus as $n \to \infty$

$$\binom{\overline{\alpha}-\beta+2n}{2n+1} = \frac{(\overline{\alpha}-\beta)(\overline{\alpha}-\beta+1)\cdots(\overline{\alpha}-\beta+2n)}{(2n+1)(2n)!} \sim \frac{(2n)^{\overline{\alpha}-\beta}}{(2n+1)\Gamma(\overline{\alpha}-\beta)} \,.$$

As a consequence, the serie $\sum_n n^{\overline{\alpha}-\beta-1}\varphi_n(\beta)$ diverges. The criteria for the convergence of Riemann series implies that $\varphi_n(\beta)$ can not be asymptotically dominated by $1/n^{\overline{\alpha}-\beta+\varepsilon}$ for $\varepsilon > 0$.

We now turn to the case (b) where the features are irregular: with positive probability $p > 0$, $X \notin \Sigma^{\overline{\alpha}/2}(\mathcal{H})$ and $\langle X, \theta_*\rangle \neq 0$. With probability p, the second iterate $\theta_1 = -\gamma\langle X_1, \theta_*\rangle X_1$ is irregular, i.e., $\theta_1 \notin \Sigma^{\overline{\alpha}/2}(\mathcal{H})$. By a simple shift of the iterates, we show that the effect of the irregularity of the initial condition for this iteration started from θ_1 has an effect equivalent to the irregularity of the optimum, thus we can apply the result above to lower bound the convergence rate. More precisely, consider the iterates $\tilde{\theta}_n = \theta_{n+1} - \theta_1$ and $\tilde{\theta}_* = \theta_* - \theta_1$. The iteration (1) can be rewritten as $\tilde{\theta}_n = \tilde{\theta}_{n-1} - \gamma\langle \tilde{\theta}_{n-1} - \tilde{\theta}_*, X_n\rangle X_n$ and $\tilde{\theta}_0 = 0$, thus the new sequence $\tilde{\theta}_n$ satisfies our framework. We can assume that (a) is satisfied, i.e., $\theta_* \in \Sigma^{\overline{\alpha}/2}(\mathcal{H})$. In that case, with probability p, $\tilde{\theta}_* = \theta_* - \theta_1 \notin \Sigma^{\overline{\alpha}/2}(\mathcal{H})$. Thus by the case above,

$$\begin{aligned}
\varphi_n(\beta) &= \mathbb{E}\left[\left\langle \theta_n - \theta_*, \Sigma^{-\beta}\left(\theta_n - \theta_*\right) \right\rangle \right] \\
&= \mathbb{E}\left[\left\langle \tilde{\theta}_{n-1} - \tilde{\theta}_*, \Sigma^{-\beta}\left(\tilde{\theta}_{n-1} - \tilde{\theta}_*\right) \right\rangle \right]
\end{aligned}$$

is not asymptotically dominated by $1/n^{\overline{\alpha}-\beta+\varepsilon}$, for $\varepsilon > 0$.

C Proof of Corollary 1

We apply Theorem 1 in the following way. Denote $\theta_n = x_n - x_0$, $\theta_* = x_* - x_0$, where $x_* = \frac{1}{N}\mathbf{1}$ is the function identically equal to $\frac{1}{N}$. These vectors belong to the Hilbert space $\mathcal{H} = \ell^2(\mathcal{V})$. Denote $\langle ., . \rangle$ and $\|.\|$ the $\ell^2(\mathcal{V})$ scalar product and norm. Denote also $X_n = e_{v_n} - e_{w_n} \in \mathcal{H}$ and $\gamma = 1/2$. Note that $\Sigma = \mathbb{E}[X_n X_n^\top] = \frac{1}{M} L$. The graph is connected thus $\lambda_0 = 0$ is the unique zero eigenvalue of L [11, Lemma 1.7]. The corresponding eigenspace is the space of constant functions. The vectors θ_n, X_n, θ_* are orthogonal to the null space of Σ, thus the quantities of the form $\langle \theta_n, \Sigma^{-\alpha}\theta_n \rangle$, $\langle X_n, \Sigma^{-\alpha} X_n \rangle$,$\langle \theta_*, \Sigma^{-\alpha}\theta_* \rangle$ are finite.

We have $\theta_0 = 0$ and the averaging update step (7) can be written as

$$\theta_n = \theta_{n-1} - \gamma \langle \theta_{n-1} - \theta_*, X_n \rangle X_n \, .$$

The last form makes explicit the parallel with Equation (1). To apply Theorem 1, we check that its assumptions are satisfied. First, $\|X_n\|^2 = 2$ a.s. thus can take $R_0 = 2$ and then $\gamma = 1/R_0$. Second, we seek $\alpha > 0$ such that $\|\Sigma^{-\alpha/2}\theta_*\| < \infty$ and $R_\alpha = \sup_{\{v,w\}\in\mathcal{E}} \langle e_v - e_w, \Sigma^{-\alpha}(e_v - e_w)\rangle < \infty$. In the following, we bound these constants for all $\alpha < d/2$, thus giving decay rates for the expected squared distance to optimum of the form $n^{-\alpha}$ for all $\alpha < d/2$. However, our bounds of the constants $\|\Sigma^{-\alpha/2}\theta_*\|$ and R_α diverge as $\alpha \to d/2$. Nevertheless, by estimating how fast the bounds diverge as $\alpha \to d/2$, we obtain a decay rate of $n^{-d/2}$ by paying an additional logarithmic factor.

Fix $0 < \alpha < d/2$. We check assumptions (a) and (b).

(a)

$$\|\Sigma^{-\alpha/2}\theta_*\|^2 = M^\alpha \left\langle x_* - x_0, L^{-\alpha}(x_* - x_0) \right\rangle = M^\alpha \sum_{i=1}^{N-1} \lambda_i^{-\alpha} \langle x_* - x_0, u_i \rangle^2 \, .$$

First, as x_* is a constant vector, $\langle x_*, u_i \rangle$ is zero for all $i \geqslant 1$. Second, $x_0 = e_{v_\star}$. Thus

$$\begin{aligned}
\|\Sigma^{-\alpha/2}\theta_*\|^2 &= M^\alpha \sum_{i=1}^{N-1} \lambda_i^{-\alpha} u_i(v_\star)^2 \\
&= M^\alpha \int_{(0,\infty)} \mathrm{d}\sigma_{v_\star}(\lambda)\, \lambda^{-\alpha} \\
&= M^\alpha \int_{(0,\infty)} \mathrm{d}\sigma_{v_\star}(\lambda) \int_0^\infty \mathrm{d}s\, \mathbf{1}_{\{s \leqslant \lambda^{-\alpha}\}} \\
&= M^\alpha \int_0^\infty \mathrm{d}s \int_{(0,\infty)} \mathrm{d}\sigma_{v_\star}(\lambda)\, \mathbf{1}_{\{\lambda \leqslant s^{-1/\alpha}\}} \\
&= M^\alpha \int_0^\infty \mathrm{d}s\, \sigma_{v_\star}((0, s^{-1/\alpha}]) \, .
\end{aligned}$$

The graph G is of spectral dimension d with constant V, thus $\sigma_{v_\star}((0, s^{-1/\alpha}]) \leqslant V^{-1} s^{-\frac{d}{2\alpha}}$. However, if $s < \delta_{\max}^{-\alpha}$, it is better to use a more naive bound. As all eigenvalues of L are smaller or equal than $\delta_{\max}$, $\sigma_{v_\star}((0, s^{-1/\alpha}]) \leqslant \sigma_{v_\star}((0, \delta_{\max}]) \leqslant V^{-1}\delta_{\max}^{d/2}$. Then

$$\begin{aligned}
\|\Sigma^{-\alpha/2}\theta_*\|^2 &\leqslant M^\alpha \left[\int_0^{\delta_{\max}^{-\alpha}} \mathrm{d}s\, V^{-1}\delta_{\max}^{d/2} + \int_{\delta_{\max}^{-\alpha}}^\infty \mathrm{d}s\, V^{-1} s^{-\frac{d}{2\alpha}} \right] \\
&= M^\alpha V^{-1} \delta_{\max}^{d/2-\alpha} \frac{d}{d-2\alpha} \, .
\end{aligned}$$

(b) Let $\{v, w\} \in E$. As $\|\Sigma^{-\alpha/2}.\|$ is a norm, by the triangle inequality,

$$\begin{aligned}
\|\Sigma^{-\alpha/2}(e_v - e_w)\|^2 &= \|\Sigma^{-\alpha/2}\left[(x_* - e_w) - (x_* - e_v)\right]\|^2 \\
&\leqslant \left(\|\Sigma^{-\alpha/2}(x_* - e_w)\| + \|\Sigma^{-\alpha/2}(x_* - e_v)\| \right)^2 \\
&\leqslant 2 \left(\|\Sigma^{-\alpha/2}(x_* - e_w)\|^2 + \|\Sigma^{-\alpha/2}(x_* - e_v)\|^2 \right) \, .
\end{aligned}$$

We bound the two quantities as above. We obtain

$$R_\alpha = \sup_{v,w\in E} \|\Sigma^{-\alpha/2}(e_v - e_w)\|^2 \leqslant 2M^\alpha V^{-1}\delta_{\max}^{d/2-\alpha}\frac{d}{d-2\alpha}\,.$$

Theorem 1 gives

$$\begin{aligned}\mathbb{E}\left[\|x_n - x_*\|^2\right] = \mathbb{E}\left[\|\theta_n - \theta_*\|^2\right] &\leqslant \frac{\alpha^\alpha}{\gamma^\alpha}\left(\|\Sigma^{-\alpha/2}\theta_*\|^2 + \frac{R_\alpha}{R_0}\|\theta_*\|^2\right)\frac{1}{n^\alpha}\\ &\leqslant \frac{(d/2)^\alpha}{(1/2)^\alpha}\left(M^\alpha V^{-1}\delta_{\max}^{d/2-\alpha}\frac{d}{d-2\alpha} + M^\alpha V^{-1}\delta_{\max}^{d/2-\alpha}\frac{d}{d-2\alpha}\|\theta_*\|^2\right)\frac{1}{n^\alpha}\end{aligned}$$

Note that $\|\theta_*\|_2^2 \leqslant 1$ and recall the scaling $t = n/M$:

$$\mathbb{E}\left[\|x_n - x_*\|^2\right] \leqslant d^{d/2+1}V^{-1}\delta_{\max}^{d/2-\alpha}\frac{1}{d/2-\alpha}\frac{1}{t^\alpha}\,.$$

This bound is valid for all $\alpha < \frac{d}{2}$. Choose $\alpha = \frac{d}{2} - \frac{\log 2}{\log t}$.

$$\mathbb{E}\left[\|x_n - x_*\|^2\right] \leqslant d^{d/2+1}V^{-1}\delta_{\max}^{2\log 2/\log t}\frac{\log t}{\log 2}\frac{2}{t^{d/2}}$$

As we assume $t \geqslant 2$, $\delta_{\max}^{2\log 2/\log t} \leqslant \delta_{\max}$. Thus we obtain conclusion 1.

The proof of 2 is similar. Theorem 1 gives

$$\begin{aligned}\min_{0\leqslant k\leqslant n}\mathbb{E}\left[\frac{1}{2}\sum_{\{v,w\}\in\mathcal{E}}\left(x_k(v) - x_k(w)\right)^2\right] &= \min_{0\leqslant k\leqslant n}\mathbb{E}\left[\frac{1}{2}\left\langle x_k - x_*, L(x_k - x_*)\right\rangle\right]\\ &= M\min_{0\leqslant k\leqslant n}\mathbb{E}\left[\frac{1}{2}\left\langle\theta_k - \theta_*, \Sigma(\theta_k - \theta_*)\right\rangle\right]\\ &\leqslant 2^\alpha\frac{\alpha^\alpha}{\gamma^{\alpha+1}}\left(\|\Sigma^{-\alpha/2}\theta_*\|^2 + \frac{R_\alpha}{R_0}\|\theta_*\|^2\right)\frac{1}{n^\alpha}\\ &\leqslant 2^{\alpha+1}d^\alpha V^{-1}\delta_{\max}^{d/2-\alpha}\frac{d}{d/2-\alpha}\frac{1}{t^{\alpha+1}}\,.\end{aligned}$$

Taking again $\alpha = \frac{d}{2} - \frac{1}{2\log t}$ and $t \geqslant 2$,

$$\min_{0\leqslant k\leqslant n}\mathbb{E}\left[\frac{1}{2}\sum_{\{v,w\}\in\mathcal{E}}\left(x_k(v) - x_k(w)\right)^2\right] \leqslant 2^{d/2+1}d^{d/2}V^{-1}\delta_{\max}\frac{d\log t}{\log 2}\frac{2}{t^{d/2+1}}$$

This gives conclusion 2 of the corollary.

D Proof of Proposition 1

The graph $\mathbb{T}_\Lambda^d$ is invariant by translation, thus the spectral measure σ_v is the same for all vertices $v \in \mathcal{V}$. Thus

$$|\mathcal{V}|\sigma_v(\mathrm{d}\lambda) = \sum_{w\in\mathcal{V}}\sigma_w(\mathrm{d}\lambda) = \sum_{w\in\mathcal{V}}\sum_{i=0}^{N-1}u_i(w)^2\delta_{\lambda_i} = \sum_{i=0}^{N-1}\left(\sum_{w\in\mathcal{V}}u_i(w)^2\right)\delta_{\lambda_i} = \sum_{i=0}^{N-1}\delta_{\lambda_i}\,.$$

Thus

$$\sigma_v((0,E]) = \frac{1}{\Lambda^d}\left|\{0 < i \leqslant N-1 | \lambda_i \leqslant E\}\right|\,.$$

We need to bound the number of eigenvalues of the Laplacian of $\mathbb{T}_\Lambda^d$ below some fixed value E. The eigenvalues of the Laplacian of the circle $\mathbb{T}_\Lambda^1$ are $1 - \cos\left(\frac{2\pi i}{\Lambda}\right)$, $i \in \mathbb{Z}$, $-\Lambda/2 < i \leqslant \Lambda/2$ [11,

Example 1.5]. As $\mathbb{T}^d_\Lambda$ is the Cartesian product $\mathbb{T}^1_\Lambda \times \cdots \times \mathbb{T}^1_\Lambda$ (with d terms), the eigenvalues of the Laplacian of the torus $\mathbb{T}^d_\Lambda$ are the

$$1-\cos\left(\frac{2\pi i_1}{\Lambda}\right)+\cdots+1-\cos\left(\frac{2\pi i_d}{\Lambda}\right), \qquad i_1,\dots i_d\in\mathbb{Z}, \quad -\frac{\Lambda}{2}<i_1,\dots,i_d\leqslant\frac{\Lambda}{2}.$$

For $y\in[-\pi,\pi]$, $1-\cos(y)\geqslant\frac{2}{\pi^2}y^2$. Thus

$$\begin{aligned}1-\cos\left(\frac{2\pi i_1}{\Lambda}\right)+\cdots+1-\cos\left(\frac{2\pi i_d}{\Lambda}\right)\leqslant E &\Rightarrow \frac{2}{\pi^2}\left[\left(\frac{2\pi i_1}{\Lambda}\right)^2+\cdots+\left(\frac{2\pi i_d}{\Lambda}\right)^2\right]\leqslant E\\ &\Leftrightarrow i_1^2+\cdots+i_d^2\leqslant\frac{E\Lambda^2}{8}.\end{aligned}$$

We need to count the number of integer points in the Euclidean ball centered at 0 and of radius $\sqrt{E/8}\Lambda$ in $\mathbb{R}^d$. This problem is famously known as Gauss circle problem. For our purposes, a crude estimate suffices: there exists a constant $C(d)$, depending only on the dimension d, such that for all radius R, the number of integer points in the ball of radius R is smaller than $1+C(d)R^d$. This leads to the final estimate

$$\begin{aligned}\sigma_v((0,E]) &= \frac{1}{\Lambda^d}\left|\left\{(i_1,\dots,i_d)\in\left(\mathbb{Z}\cap\left(-\frac{\Lambda}{2},\frac{\Lambda}{2}\right]\right)^d\setminus\{0\}\ \text{ such that }\right.\right.\\ &\qquad\qquad\left.\left.1-\cos\left(\frac{2\pi i_1}{\Lambda}\right)+\cdots+1-\cos\left(\frac{2\pi i_d}{\Lambda}\right)\leqslant E\right\}\right|\\ &\leqslant \frac{1}{\Lambda^d}\left|\left\{(i_1,\dots,i_d)\in\mathbb{Z}^d\setminus\{0\}\,\middle|\, i_1^2+\cdots+i_d^2\leqslant\frac{E\Lambda^2}{8}\right\}\right|\\ &\leqslant \frac{1}{\Lambda^d}C(d)\left(\frac{E\Lambda^2}{8}\right)^{d/2}=\frac{C(d)}{8^{d/2}}E^{d/2}.\end{aligned}$$

This proves the proposition with $V(d)=8^{d/2}/C(d)$.

E Proof of Theorems 5 and 6

Note that in this proof, we use the strong assumptions of regularity of the feature vector X. We do not know whether it is possible to prove the same result under the weak assumptions of Remark 3.

Our proof stategy is the following: we decompose the SGD iterates sequence θ_n as a sum of sequences $\theta_n=\nu_n+\sum_{l=1}^n\eta_n^{(l)}$, where each of the auxiliary sequences is interpreted as the iterates of some SGD iteration under a noiseless linear model. We thus apply the results of Section 2 to control these auxiliary sequences and obtain the presented bound.

Define $\varepsilon_n=Y_n-\langle\theta_*,X_n\rangle$, the error of the best linear estimator. Then Equation (8) can be rewritten as

$$\theta_0=0, \qquad\qquad \theta_n=\theta_{n-1}-\gamma\langle\theta_{n-1}-\theta_*,X_n\rangle X_n+\gamma\varepsilon_n X_n.$$

We see this iteration as an additively perturbed version of the iteration

$$\nu_0=0, \qquad\qquad \nu_n=\nu_{n-1}-\gamma\langle\nu_{n-1}-\theta_*,X_n\rangle X_n,$$

studied in Section 2. To understand the effect of the additive noise, define for all $l\geqslant 1$,

$$\eta_l^{(l)}=\gamma\varepsilon_l X_l, \qquad\qquad \eta_n^{(l)}=\eta_{n-1}^{(l)}-\gamma\langle\eta_{n-1}^{(l)},X_n\rangle X_n, \qquad n>l.$$

Then

$$\theta_n=\nu_n+\sum_{l=1}^n\eta_n^{(l)}. \tag{25}$$

Indeed, this last equation is checked by induction: $\theta_0 = 0 = \nu_0$, and if the equation is satisfied for some $n \geqslant 0$,

$$\begin{aligned}
\theta_{n+1} &= \theta_n - \gamma\langle \theta_n - \theta_*, X_{n+1}\rangle X_{n+1} + \gamma\varepsilon_{n+1}X_{n+1} \\
&= \nu_n + \sum_{l=1}^n \eta_n^{(l)} - \gamma\left\langle \nu_n + \sum_{l=1}^n \eta_n^{(l)} - \theta_*, X_{n+1}\right\rangle X_{n+1} + \eta_{n+1}^{(n+1)} \\
&= [\nu_n - \gamma\langle \nu_n - \theta_*, X_{n+1}\rangle X_{n+1}] + \sum_{l=1}^n \left[\eta_n^{(l)} - \gamma\langle \eta_n^{(l)}, X_{n+1}\rangle X_{n+1}\right] + \eta_{n+1}^{(n+1)} \\
&= \nu_{n+1} + \sum_{l=1}^n \eta_{n+1}^{(l)} + \eta_{n+1}^{(n+1)}\,.
\end{aligned}$$

We use the decomposition (25) to study $\varphi_n(\beta)$. Using the triangle inequality,

$$\begin{aligned}
\varphi_n(\beta) &= \mathbb{E}\left[\left\|\Sigma^{-\beta/2}\left(\nu_n + \sum_{l=1}^n \eta_n^{(l)}\right)\right\|^2\right] \\
&\leqslant \mathbb{E}\left[\left(\left\|\Sigma^{-\beta/2}\nu_n\right\| + \left\|\Sigma^{-\beta/2}\sum_{l=1}^n \eta_n^{(l)}\right\|\right)^2\right] \\
&\leqslant 2\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\nu_n\right\|^2\right] + 2\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\sum_{l=1}^n \eta_n^{(l)}\right\|^2\right] \qquad (26)
\end{aligned}$$

The first term is studied in Section 2. We detail the analysis of the second term. Note that

$$\begin{aligned}
\eta_n^{(l)} &= (I - \gamma X_n \otimes X_n)\eta_{n-1}^{(l)} = \cdots = (I - \gamma X_n \otimes X_n)\cdots(I - \gamma X_{l+1} \otimes X_{l+1})\eta_l^{(l)} \\
&= (I - \gamma X_n \otimes X_n)\cdots(I - \gamma X_{l+1} \otimes X_{l+1})\gamma\varepsilon_l X_l\,. \qquad (27)
\end{aligned}$$

Thus if $l < l'$,

$$\begin{aligned}
\mathbb{E}\left[\left\langle \eta_n^{(l)}, \Sigma^{-\beta}\eta_n^{(l')}\right\rangle\right] &= \mathbb{E}\left[\left\langle \mathbb{E}\left[\eta_n^{(l)}\middle| X_{l+1}, \ldots, X_n\right], \Sigma^{-\beta}\eta_n^{(l')}\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle (I - \gamma X_n \otimes X_n)\cdots(I - \gamma X_{l+1} \otimes X_{l+1})\gamma\mathbb{E}[\varepsilon_l X_l], \Sigma^{-\beta}\eta_n^{(l')}\right\rangle\right]
\end{aligned}$$

Note that by definition of θ_*, $0 = \nabla\mathcal{R}(\theta_*) = -\mathbb{E}\left[(Y_l - \langle\theta_*, X_l\rangle)X_l\right] = -\mathbb{E}\left[\varepsilon_l X_l\right]$ thus we obtain that the cross products $\mathbb{E}\left[\left\langle \eta_n^{(l)}, \Sigma^{-\beta}\eta_n^{(l')}\right\rangle\right]$ are zero. This gives

$$\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\sum_{l=1}^n \eta_n^{(l)}\right\|^2\right] = \sum_{l=1}^n \mathbb{E}\left[\left\|\Sigma^{-\beta/2}\eta_n^{(l)}\right\|^2\right]\,.$$

Note that from Equation (27), $\eta_n^{(l)}$ and $\eta_{n-l+1}^{(1)}$ are equal in law. Thus

$$\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\sum_{l=1}^n \eta_n^{(l)}\right\|^2\right] = \sum_{l=1}^n \mathbb{E}\left[\left\|\Sigma^{-\beta/2}\eta_{n-l+1}^{(1)}\right\|^2\right] = \sum_{l=1}^n \mathbb{E}\left[\left\|\Sigma^{-\beta/2}\eta_l^{(1)}\right\|^2\right]\,. \qquad (28)$$

This last quantity is the sum of the expected squared power norms

$$\varphi_l'(\beta) := \mathbb{E}\left[\left\|\Sigma^{-\beta/2}\eta_l^{(1)}\right\|^2\right]$$

of the SGD iterates $\eta_l^{(1)}, l \geqslant 1$ on a noiseless linear model, with initialization $\eta_1^{(1)} = \gamma\varepsilon_1 X_1$. When $\beta = -1$, this control is given by (18): with our notation here, this gives

$$\sum_{l=1}^n \varphi_l'(-1) \leqslant \sum_{l=1}^\infty \varphi_l'(-1) \leqslant \frac{1}{\gamma}\varphi_1'(0)\,. \qquad (29)$$

When $\beta = \underline{\alpha} - 1$, a similar control can be obtained from (19) which gives:

$$2\gamma\varphi'_{l-1}(\underline{\alpha}-1) \leqslant \varphi'_{l-1}(\underline{\alpha}) - \varphi'_l(\underline{\alpha}) + \gamma^2 R_{\underline{\alpha}}\varphi'_{l-1}(-1)\,.$$

By summing these inequalities for $l = 2, 3, \ldots$, we obtain,

$$\begin{aligned} 2\gamma\sum_{l=1}^{\infty}\varphi'_l(\underline{\alpha}-1) &\leqslant \varphi'_1(\underline{\alpha}) + \gamma^2 R_{\underline{\alpha}}\sum_{l=1}^{\infty}\varphi'_l(-1) \\ &\leqslant \varphi'_1(\underline{\alpha}) + \frac{R_{\underline{\alpha}}}{R_0}\varphi'_1(0) \end{aligned} \tag{30}$$

Note that using the strong assumption of regularity of the feature vectors,

$$\begin{aligned} \varphi'_1(0) &= \mathbb{E}\left[\|\gamma\varepsilon_1 X_1\|^2\right] \leqslant \gamma^2 R_0 \mathbb{E}\left[\varepsilon_1^2\right] = 2\gamma^2 R_0 \mathcal{R}(\theta_*)\,, \\ \varphi'_1(\underline{\alpha}) &= \mathbb{E}\left[\left\|\Sigma^{-\underline{\alpha}/2}\gamma\varepsilon_1^2 X\right\|^2\right] \leqslant \gamma^2 R_{\underline{\alpha}} \mathbb{E}\left[\varepsilon_1^2\right] = 2\gamma^2 R_{\underline{\alpha}} \mathcal{R}(\theta_*)\,. \end{aligned}$$

We use these expressions to simply further (29) and (30):

$$\begin{aligned} \sum_{l=1}^{n}\varphi'_l(-1) &\leqslant 2\gamma R_0 \mathcal{R}(\theta_*)\,, \\ \sum_{l=1}^{\infty}\varphi'_l(\underline{\alpha}-1) &\leqslant 2\gamma R_{\underline{\alpha}} \mathcal{R}(\theta_*)\,. \end{aligned}$$

If $\beta \in [-1, \underline{\alpha} - 1]$, we use the log-convexity Property 1 and Hölder's inequality: decompose $\beta = (1-\lambda)(-1) + \lambda(\underline{\alpha}-1)$ with $\lambda = (\beta+1)/\underline{\alpha}$,

$$\begin{aligned} \sum_{l=1}^{\infty}\varphi'_l(\beta) &\leqslant \sum_{l=1}^{\infty}\varphi'_l(-1)^{1-\lambda}\varphi'_l(\underline{\alpha}-1)^{\lambda} \\ &\leqslant \left(\sum_{l=1}^{n}\varphi'_l(-1)\right)^{1-\lambda}\left(\sum_{l=1}^{\infty}\varphi'_l(\underline{\alpha}-1)\right)^{\lambda} \\ &\leqslant \left(2\gamma R_0 \mathcal{R}(\theta_*)\right)^{1-\lambda}\left(2\gamma R_{\underline{\alpha}}\mathcal{R}(\theta_*)\right)^{\lambda} \\ &= 2\gamma R_0^{1-\lambda} R_{\underline{\alpha}}^{\lambda}\mathcal{R}(\theta_*)\,. \end{aligned} \tag{31}$$

Putting back together Equations (26), (28) and (31), we obtain

$$\varphi_n(\beta) \leqslant 2\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\nu_n\right\|^2\right] + 4\gamma R_0^{1-\lambda} R_{\underline{\alpha}}^{\lambda}\mathcal{R}(\theta_*)$$

The theorem follows the application of Theorem 3 to the sequence ν_n in order to control the first term.