Last iterate convergence of SGD for Least-Squares in the Interpolation regime

Anonymous Author(s)
Affiliation
Address
email

Abstract

Motivated by the recent successes of neural networks that have the ability to fit the data perfectly *and* generalize well, we study the noiseless model in the fundamental least-squares setup. We assume that an optimum predictor perfectly fits the inputs and outputs $\langle \theta_*, \phi(X) \rangle = Y$, where $\phi(X)$ stands for a possibly infinite dimensional non-linear feature map. To solve this problem, we consider the estimator given by the last iterate of stochastic gradient descent (SGD) with constant step-size. In this context, our contribution is two fold: (i) *from a (stochastic) optimization perspective*, we exhibit an archetypal problem where we can show explicitly the convergence of SGD final iterate for a non-strongly convex problem with constant step-size whereas usual results use some form of average and (ii) *from a statistical perspective*, we give explicit non-asymptotic convergence rates in the over-parameterized setting and leverage a *fine-grained* parameterization of the problem to exhibit polynomial rates that can be faster than $O(1/T)$. The link with reproducing kernel Hilbert spaces is established.

1 Introduction

As soon as large-scale statistics and optimization need to work together, stochastic gradient descent (SGD) is the core algorithm everybody tries to build upon [7]. Its versatility, practicability and adaptability make it the workhorse of almost every supervised machine learning problem. Yet, its outstanding efficiency remains mysterious, or at least surprising on certain aspects. Furthermore, the recent successes of deep neural networks (DNN) brought a new paradigm to the classical supervised learning setting with the ability to fit the data perfectly *and* to generalize well [5]. Following this idea, the old statistical modelling where the model suffers from problem-dependent noise has to be revisited: there is a need of analyzing stochastic algorithms in this new light [18]. Whether we call it *over-parameterization* to put emphasis on the large number of neurons needed in DNNs, *interpolation* as in approximation theory or *noiseless model* to stress the absence of noise in this statistical model, all these terminologies refer to the same idea. This regime brings with it new insights that are better to model and hence understand the current machine learning setup.

Hence, the main question: how would SGD profit from this noiseless model? At first glance, the story seems clear: the old problem of variance at optimum making the SGD iterates oscillate asymptotically now disappears. Hence, should also disappear techniques to prevent from this, namely, averaging and decaying step sizes: one should be able to study the convergence of the SGD *final iterate* with *constant step-size*.

However, the study of the last iterate of SGD has always caused some technical problems preventing from a clear theory in this case [25]. Indeed, the convergence of the final iterate is much more difficult to prove than the convergence of the average of the iterates [26, 13, 15]. This counter-intuitive difficulty can be explained by the fact that the interactions coming from the sampling noise of

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

SGD prevent the final iterates from forming a decreasing sequence. Therefore standard Lyapounov strategies often failed in such a setting. Besides, even if averaged SGD has shown *theoretically* some good convergence properties, the final iterate is commonly used *in practice*. Finally, averaging techniques always suffer from saturation coming from the slow forgetting of initial conditions.

To tackle these questions, we consider the simplest setting of the linear regression over features $\phi(X)$ in a Hilbert space $\mathcal{H}$. In this context, the setup corresponds to the existence of linear relationship between the output and the input: $Y = \langle \theta_*, \phi(X) \rangle$. Note that this setting is rich as the features can be a non-linear transformation of the inputs, as it is commonly the case when they are defined through a positive-definite kernel $K(X, X')$ [24, 27]. Note also that our analysis will, unless stated explicitly, be conducted in a non-parametric and dimensionless fashion, enabling the features to come from a infinite dimensional reproducing kernel Hilbert space. In this perspective, the problem we are considering is not strongly convex.

Main contributions. The aim of the present article is to answer at once the two following problems: (i) *from a (stochastic) optimization perspective*, the goal is to exhibit an archetypal problem where we can show explicitly the convergence of SGD final iterate for a non-strongly convex problem with constant step-size and (ii) *from a statistical/machine learning perspective*, the aim is to push deeper the study of the over-parameterized setting for the non-parametric least-squares problem. Contrary to the noisy setting, where (almost) everything is known, the noiseless model suffers from a certain lack of understanding. More precisely, our contributions are the following:

- We show that the final iterate of constant step-size SGD achieves a convergence rate of $O(\ln(T)/T)$ under minimal assumptions. With a slightly stricter hypothesis we improve this convergence rate to $O(1/T)$.
- Going further, we assume the usual non-parametric *capacity and source conditions* respectively on the spectrum of the covariance and the optimum, and derive bounds for this fine-grained model. In this setup, we show the SGD *fast rates* of $O(1/T^{1+\alpha})$.
- We derive an explicit recursion on the eigenspaces of the covariance matrix that is of its own interest. Indeed, it is the cornerstone of our analysis and could be very useful in the future to understand important properties of SGD in this setting.

From a technical standpoint, going from the average to the last iterate is far from anodyne for constant step-size SGD even in the interpolation regime. Indeed, in a similar setting, D.Aldous considered it an open problem [1, Sec. 3.3, Open problem (i)]. Prior to our result, how to directly deal with the fluctuations induced by the stochasticity of the gradients was not known. Our work presents a direct Lyapunov technique that handles them without any explicit variance reduction methods like averaging or step-size decay.

1.1 Related Work

As recalled earlier, it is often easier to show convergence for the averaged iterates of SGD than for the final one. However, there has been a huge effort by the optimization and machine learning communities to work on the final iterate. We report here the different works and contexts for which such results are shown. All the results stated below are for convergence in function value.

Last versus averaged iterates. First, for *non-smooth functions* (variance of the SGD-gradients uniformly bounded), it is easy to show, with proper decrease of the step sizes, that the averaged iterates converge at rate $O(1/\sqrt{T})$ in the non-strongly convex case and $O(1/T)$ when the function is strongly convex. On the other side, convergence of the last iterate has been first addressed by [34, 26] who showed convergence rates of $O(\log(T)/\sqrt{T})$ for non-strongly convex and $O(\log(T)/T)$ for strongly convex. These rates have recently been shown to be tight by [13] (the $\log$ term cannot be suppressed for the last iterate). Secondly, for *smooth functions* (gradient Lipschitz and variance bounded only at optimum), [2] showed in the non-strongly convex case that smoothness does not help; results on the averaged iterate are the same $O(1/\sqrt{T})$ and results on the last iterate are actually worse: $O(T^{-1/3})$. In the strongly convex setting, the optimal rate $O(1/T)$ rate is obtained both by the averaged and last iterates. However the averaged iterates are preferred since they lead to the optimal covariance and the step size is independent of the strong convexity constant [22]. Finally, the final iterate of SGD, despite its ubiquitous use in machine learning, never theoretically performs as well as the averaged, except when used with a geometrically-decaying learning-rate [15, 12].

Over-parameterized setting. As recalled earlier, SGD in the over-parameterization regime corresponds to the assumption that all the gradients of the function vanish at optimum. It has been first studied by [23] who assumed furthermore a strong growth condition (*SGC*) [first introduced by 28, 31]) but often too stringent to match the machine learning practical set-up. Up to our knowledge, [18] was the first to point out that this interpolation regime was particularly relevant in the recent deep learning framework and to show linear convergence in the strongly-convex setting. In the non-strongly convex case, $O(1/T)$-convergence has been shown by [32] for the averaged iterates. Consequently, there has been many papers discussing this setting, proving convergence rates in different contexts: Polyak-Lojasiewicz [4], accelerated [32, 17], second-order [19], line-search [33]. However, no rate for the final iterate has been shown in the non-strongly convex setting for the last iterate as the only convergence rates achieved, $O(1/T)$, corresponds in all these works to the averaged iterates.

The Quadratic Case. More specifically, the quadratic case has been widely studied as a central model in the machine learning literature. As explained above, on the one hand, in the usual noisy setting, averaging or decaying step sizes have been considered, showing well-known $O(1/T)$ convergence rates [3]. The final iterate convergence of SGD for the noisy quadratic case is studied in [12] where they show a $O(d \log T/T)$ rate and advocate the primacy of exponential step decay. On the other hand, the SGD in over-parameterized setting has been considered only lately by [18, 6, 35]. The first work corresponds to the strongly convex case, where the final iterate provably converges whereas our work is more aligned to [6, 35] where the non-strongly convex case is considered. In [6], among other results, it is shown that only the minimum of the function value iterates converge at rate $O(1/T)$ and up to $O(1/T^{1+\alpha})$ for $\alpha > 0$ when usual *capacity and source* conditions are assumed for the problem [8, 29]. A part of our work can be seen as a continuation of [6], strengthening its result by proving that the final iterate converges instead of the $\min$ or some average.

2 Problem Set-up: Stochastic Gradient Descent on the Least-squares Problem

The setting is the classic one for stochastic gradient descent for linear least-squares in some Hilbert space $\mathcal{H}$. The function we would like to minimize over $\theta \in \mathcal{H}$ is

$$\mathcal{R}(\theta) = \frac{1}{2}\mathbb{E}_\rho \left(\langle \theta, \phi(X)\rangle - Y\right)^2, \tag{1}$$

when ρ is the joint law over input/outputs $(X, Y) \in \mathcal{X} \times \mathbb{R}$ and ϕ is a feature map from $\mathcal{X}$ to $\mathcal{H}$. For the sake of simplicity, the reader can assume that $\mathcal{H}$ is a finite-dimensional Euclidean space. However, we take a special care that all the quantities and definitions could be valid when $\mathcal{H}$ is infinite dimensional. Hence, all the theorems are valid in a Reproducing Kernel Hilbert Spaces (RKHS) framework. Without any loss of generality, and to avoid heavy notations, we assume that $\phi(X) = X$, and as soon as we refer the the finite dimensional case, we set $\mathcal{H} = \mathbb{R}^d$.

Covariance matrix. We define the covariance operator on $\mathcal{H}$: $H := \mathbb{E}_\rho[X \otimes X]$. This positive semi-definite operator is diagonalizable and plays a central role in our analysis. Hence, let us denote λ_i its non-negative eigenvalues sorted in non-increasing order and v_i its corresponding eigenvectors. In finite dimension, the covariance operator is simply a $d \times d$ matrix $H = \mathbb{E}_\rho[XX^\top]$ but, once again, notice that the analysis can also afford being infinite dimensional. In either case we denote $\lambda_{\max}$ the largest eigenvalue of H. In the finite dimensional case $\lambda_{\min} > 0$ is always correctly defined but can be arbitrarily small. On the opposite, a standard and important consequence of the infinite dimension is that, in most cases (especially when $\mathrm{Tr} H < +\infty$), the infinite sequence of eigenvalues is no longer lower bounded as it converges to 0: H is no longer invertible and the problem is non-strongly convex.

Noiseless model. We assume that the model does not suffer from any noise: there exists $\theta_* \in \mathcal{H}$ such that, ρ-almost surely, $\langle \theta_*, x\rangle = y$. This means that the model is well specified *and* that there is no noise at optimum. Due to this noiseless condition, we expect our iterates to go to the optimum without decaying step-sizes or averaging. We can also rewrite the Risk in Eq. (1):

$$\mathcal{R}(\theta) = \frac{1}{2}\left(\theta - \theta_*\right)^\top H \left(\theta - \theta_*\right) = \frac{1}{2}\mathrm{Tr}\left[\left(\theta - \theta_*\right)\left(\theta - \theta_*\right)^\top H\right]. \tag{2}$$

SGD with constant step-size. To minimize the function $\mathcal{R}$ defined in Eq. (1), we do not have access to the distribution ρ but to a stream of i.i.d. observations $(x_t, y_t)_{t \geqslant 1}$ sampled from it. We hence

perform a gradient descent in the direction given by one sample at a time with *constant* stepsize $\gamma > 0$. Throughout all this paper, the initial condition is always set to $\theta_0 = 0$ and for $t \geqslant 1$,

$$\theta_{t+1} = \theta_t - \gamma \left(\langle \theta_t, x_t \rangle - y_t \right) x_t. \tag{3}$$

The impossibility of linear rates. We stress that the least-squares setup we consider is a *non-strongly* convex problem. Indeed, in finite dimension d, $\lambda_{\min} > 0$, and SGD converges at linear rate $\sim e^{-\gamma \lambda_{\min} T}$ [18, Theorem 1], this asymptotic regime occurring after a time scale $\tau \sim 1/(\gamma \lambda_{\min})$. However, this apparent strong convexity is a lure, since the large (or infinite) dimension makes this convergence rate vacuous for a arbitrarily small $\lambda_{\min}$. Hence we focus on the non-parametric non-strongly convex setup where we prove non-asymptotic polynomial rates. Yet in finite dimension, it is always possible to see the linear regime after time $1/(\gamma \lambda_{\min})$ as shown in Figure 1, even if this time can be arbitrarily long.

3 Convergence rates of the final iterate of SGD

In large dimension, two quantities govern principally the rates of convergence of least-squares estimators: (i) the spectrum of the covariance matrix, and (ii) how the solution, θ_*, projects on the eigenbasis of the covariance matrix. In this section, we state refined assumptions on the spectrum of the covariance matrix and the decomposition of θ_* over its eigenbasis. Note that in finite-dimension, all these quantities are always finite, but can be extremely large compared to the sample size T. This is why the reader can see these more as a *fine-grained parameterization* of the problem rather than restrictive assumptions. The assumptions we make always go in pairs, (i) one for the features through the covariance matrix (Assumptions 1, 3 and 5) and (ii) one for the target solution (Assumptions 2, 4 and 6).

Summary on the results. As assumptions go stricter, the convergence rates go faster. Every theorem is a bound on the *expected risk* given by the *last iterate* of the SGD recursion with *constant* step size γ, started at $\theta_0 = 0$. Only for Theorem 1, for which the assumptions are the weakest, we allow the step-size to depend on the finite horizon T (through its logarithm). All the theorems are stated with respect to finite constants defined thanks to the different assumptions. The reader can refer to Table 1 for a concise summary of them. Note also that we put a particular effort for the clarity of the bounds, and hence, some numerical constants might appear large. These are simple artifacts on the proofs and could be lowered, but at the price of less clear results. A more detailed summary of our results can be found in Appendix A. All the proofs are deferred to Appendix C.

Theorem	Assumption	Condition	Rate
Theorem **1**	A. **1**	$\mathbb{E}\left[\|X\|^2 \, XX^\top\right] \preccurlyeq RH$	$\dfrac{\ln(T)}{T}$
	A. **2**	$\|\theta_*\|_{\mathcal{H}} < +\infty$	
	$\Uparrow$	$\Uparrow$	
Theorem **2**	A. **3**	$\mathbb{E}\left[\left\langle X, \ln\left(H^{-1}\right)X\right\rangle \, XX^\top\right] \preccurlyeq R_{\ln} H$	$\dfrac{1}{T}$
	A. **4**	$\mathrm{Tr}(M_0 \ln(H^{-1})) < +\infty$	
	$\Uparrow$	$\Uparrow$	
Theorem **3**	A. **5**	$\mathbb{E}\left[\left\|H^{-\alpha/2}X\right\|^2 \, XX^\top\right] \preccurlyeq R_\alpha H, \ \alpha \in (0,1)$	$\dfrac{1}{T^{1+\alpha\wedge\beta}}$
	A. **6**	$\mathrm{Tr}(M_0 H^{-\beta}) < +\infty, \ \beta \geqslant 0$	

Table 1: Table showing the main results of the article: different upper-bounds for the convergence of the SGD final iterate determined by different assumptions.

3.1 Standard Least-Squares setting

As often, when we analyze SGD, we make an 4-th order assumption on the distribution of the features.

Assumption 1 (Fourth Moment Condition) *There exists a finite constant $R \geq 0$, such that*

$$\mathbb{E}\left[\|X\|^2 \, XX^\top\right] \preccurlyeq RH \tag{4}$$

The assumption holds in case of bounded features i.e $\|X\|^2 \leq R$ a.s. It also holds for features with infinite supports, such as sub-Gaussian data, and simple one-hot basis distributions (i.e. $x = e_i$ with probability p_i). This is a standard assumption when analysing SGD for least squares [3, 14, 12] which is weaker than what is assumed in [10, 11, 35].

Assumption 2 (Attainable case) *The target solution θ_* lies in the space $\mathcal{H}$. It ensures that it has a finite norm $\|\theta_*\|_{\mathcal{H}} < +\infty$.*

While this is always true in finite dimension, this assumption draws the attention on the fact that the norm of the optimum θ_* could be very large in high-dimension. As a limit, in infinite-dimensional spaces, θ_* could not belong to $\mathcal{H}$ and hence would have an infinite norm. This is why we refer to this assumption as the *attainable case*. Under these two assumptions we have the following result.

Theorem 1 *Assume Assumptions 1, 2. Then, for $T \geqslant 2$, if we set $\gamma = (4R\ln(T))^{-1}$, we have the following bound for the expected risk of the estimator given by the Tth iterate of SGD:*

$$\mathbb{E}\mathcal{R}(\theta_T) \leqslant 3R\|\theta_*\|^2_{\mathcal{H}} \frac{\ln(T)}{T}. \tag{5}$$

Let us comments upon this result proven in Appendix C.1. The theorem above states that, under mild assumptions, if we allow the step-size to depend on the time horizon, we have a $O(\ln(T)/T)$ convergence rate for the final iterate. First, removing the dependence on finite horizon T for the step-size by considering decaying step-sizes $\gamma_t \propto 1/\ln(t)$ could be done, but we decided to keep this way as we have focused on constant step-size in this paper. Furthermore, Theorem 2 of [6] shows that this bound is optimal up to $\log(T)$ for a SGD. This rate can also be compared to the classical optimization results for the non-strongly convex objective (our case). It is well known that gradient descent and averaged SGD (even for non-quadratic objectives) achieve a $O(1/T)$ rate with constant step size. Similarly [6] achieves a convergence rate of $O(1/T)$ rate with constant step size for the $\min$ of the function value along the iterates. However, the rate of convergence of SGD last iterate was an open problem [see 6, Remark 1]. Note however that our bound suffers from a $\log(T)$ both in function value and for the step-size. Whether this term is necessary is an open question for us. The purpose of the following development is to first remove these $\log(T)$ dependence at the price of a $\log$-scale refinement of the assumptions.

3.2 A logarithm-scale refinement

The second sequence of assumptions is slightly stronger. They reinforce assumptions 1 and 2 at the log-scale. Once again, there is one on the 4-th order moment and the other on the target.

Assumption 3 ($\log$-regularity of the features) *There exists constant $\lambda_{\rm o} > 0$ and $R_{\ln}$ such that the covariance matrix H satisfies the following condition:*

$$\mathbb{E}\left[\left\langle X, \ln\left(\lambda_{\rm o} H^{-1}\right) X\right\rangle \; XX^\top\right] \preccurlyeq R_{\ln} H.$$

Note that $\lambda_{\rm o}$ is a just a reference which lets us present cleaner results. We choose $\lambda_{\rm o}$ such that $7R_{\ln} \leqslant \lambda_{\rm o}$. This is always possible as $R_{\ln}$ scales linearly with $\ln \lambda_{\rm o}$ (see details in A.1). This condition implies in fact $\mathrm{Tr}\big(H \ln(\lambda_{\rm o} H^{-1})\big) \leqslant R_{\ln}$, which consequently implies the following eigenvalue decay: $\lambda_i \leqslant O(1/(i \ln i))$ (see details in A.1).

Assumption 4 ($\log$-regularity of the optimum) *The covariance matrix at optimum $M_0 = \theta_* \theta_*^\top$ satisfies the following condition:*

$$\mathrm{Tr}\left(M_0 \ln\left(H^{-1}\right)\right) < +\infty.$$

We define $C_{\ln} := \sum_i m_i^0 \ln(\lambda_{\rm o}/\lambda_i)$, with the same reference $\lambda_{\rm o}$. To give an order of magnitude in finite dimensions, when the spectrum of the covariance matrix in lower bounded by $\lambda_{\min} > 0$, $C_{\ln}$ can be estimated as $C_{\ln} \leqslant \|\theta_*\|^2_{\mathcal{H}} \ln \lambda_{\rm o}/\lambda_{\min} \sim a\|\theta_*\|^2_{\mathcal{H}} \ln d$ even if $\lambda_{\min}$ is as small as $1/d^a$. When the dimension is not too big (even at the $\log$-scale), this constant $C_{\ln}$ is comparable with $\|\theta_*\|^2_{\mathcal{H}}$. Under these two assumptions we have the following convergence result.

Theorem 2 *Assume Assumptions 1,3,4. Then, for $T \geqslant 3$, if we set $\gamma = (14R_{\ln})^{-1}$, we have the following bound for the expected risk of the estimator given by the Tth iterate of SGD:*

$$\mathbb{E}\mathcal{R}(\theta_T) \leqslant \frac{10R_{\ln}C_{\ln}}{T}. \tag{6}$$

This theorem, proven in Appendix C.2, states that at the price of a log-scale refinement on the features and optimum, the SGD-convergence rate is $O(1/T)$. This naturally restricts the class of problems that suffers from a $\log(T)$ (see Theorem 1) to a very small class: roughly speaking, it is the class of problems for which the eigenvalue decreasing rate is strictly squeezed between: $O(1/(i \ln i)) < \lambda_i < O(1/i)$. The role of the assumptions is fundamental here: Assumption 3 allows to remove the $\ln(T)$ for the step-size and Assumption 4 allows to remove the $\ln(T)$ for the convergence rate. The proof technique is the same as for the previous theorem, the only difference is that the assumptions allow to control more precisely the bias and the variance in the SGD recursion.

3.3 A fine-grained parameterization of the problem: *capacity and source* conditions

The final set of assumptions have been introduced by [8]. They are in the same vein as above assumptions and are often called *capacity* and *source* conditions in the Reproducing Kernel Hilbert Spaces community.

Assumption 5 (capacity condition: α-regularity of the features)) *The covariance matrix H is such that there exists $\alpha > 0$ and a finite constant $R_\alpha \geq 0$ verifying*

$$\mathbb{E}\left[\left\langle X, H^{-\alpha}X\right\rangle XX^\top\right] \preccurlyeq R_\alpha H.$$

Note that Assumption 5 is strictly more demanding than Assumption 3. This is same as *regularity of the features* [6, Remark 3]. Note also that $\alpha = 0$ corresponds to Assumption 1 and the larger the α the stricter the assumption is. The capacity condition stated above implies $\mathrm{Tr}\big(H^{1-\alpha}\big) \leqslant R_\alpha$. which consequently implies a eigenvalue decay as a power law for the sequence of eigenvalue of H: $\lambda_i = O(1/i^{\frac{1}{1-\alpha}})$ (see details in A.1). It is also often related to the *effective dimension* of the problem [8]. Finally, as a limiting case, when $\alpha \to 1$, $R_\alpha \to \mathrm{Tr}(\mathrm{Id}) = d$ and the assumption is only valid in finite dimension. Hence, it is expected that the bound would blow in large dimension when α grows to one. For this reason, in the literature [8, 29], the bound is often reparameterized as $\alpha' = (1-\alpha)^{-1}$ to push this singularity at infinity. We decide to keep our parameterization for the clarity of the result. The same type of assumption can be stated for the optimum.

Assumption 6 (source condition: β-regularity of the optimum)) *The covariance matrix at optimum $M_0 = \theta_*\theta_*^\top$ is such that there exists $\beta > -1$ such that $\mathrm{Tr}(M_0 H^{-\beta}) < +\infty$. Here, we define*

$$C_\beta := \mathrm{Tr}\big(H^{-\beta}M_0\big) = \sum_i \lambda_i^{-\beta} m_i^0.$$

Here are important remarks on this assumption. First note that

$$\mathrm{Tr}(M_0 H^{-\beta}) = \mathrm{Tr}(\theta_*\theta_*^\top H^{-\beta}) = \theta_*^\top H^{-\beta}\theta_* = \|H^{-\beta/2}\theta_*\|_{\mathcal{H}}^2.$$

Hence, $\beta = 0$ corresponds to Assumption 2 which is the attainable case. It is worth noting that for $\beta \in (-1, 0)$, this assumption takes into account the fact that θ_* does not necessarily belong to $\mathcal{H}$. In such a case, it is still possible to define properly θ_* as the infimum over $\mathcal{H}$ of the risk R defined in Eq.(1). However, for sake of clarity and to avoid cumbersome technicalities, we refer to [9, p.1395] for the extension of the analysis is this case. As before, for $\beta > 0$, the larger the β the stricter the assumption is. Finally as a limiting case, $\beta \to +\infty$ corresponds to the fact that θ_* belongs to a finite dimensional space. This hypothesis is often called the *source condition* in the literature because it quantifies the complexity of the optimum [see 8, 29, 20, for further details]. Once again, parameterization has not yet been fixed by the literature and one often uses the parameter r that corresponds to $r = \frac{\beta+1}{2}$. We have the following theorem under these capacity and source conditions.

Theorem 3 *Assume Assumptions 5, 6 with constants $\alpha \in (0, 1)$ and $\beta > -1$ respectively. Then, for $T \geqslant 3$, we have the following bound for the expected risk of the estimator given by the Tth iterate of SGD:*

$$\mathbb{E}\mathcal{R}(\theta_T) \leqslant 2C_\beta \left(\frac{1+\beta}{\gamma}\right)^{1+\beta} \frac{1}{T^{1+\alpha\wedge\beta}}, \tag{7}$$

where $\gamma^{1-\alpha} \leqslant (32\kappa\xi_\alpha R_\alpha)^{-1}$ *and* $\xi_\alpha = \sum_{n\geqslant 1} \frac{1}{n^{1+\alpha}}$.

Let us comment this result. Its proof can be found in Appendix C.3.

Discussion on the limit cases for α and β. Let us note a few observations on the two parameters α and β. Firstly, it is observed in [6] that the rate $1/T^{1+\alpha\wedge\beta}$ is optimal in terms of power law for SGD. Second, the convergence rate depends on the minimum of the two. It implies that either the regularity of the features, or of the optimum, one or the other is a bottleneck for the convergence of SGD. More precisely if $\alpha < \beta$, the features are not regular enough to counter the multiplicative noise of SGD. Conversely, when the optimum is the bottleneck, there is no difference between SGD and Gradient Descent (GD), as $T^{1+\beta}$ is the rate of convergence of GD. Note that, $\alpha \to 0$, corresponds to the setting of Theorem 1 where a $\log(T)$ appears. Hence, it is normal that our bound does not hold in this limit: indeed, we remark that the step-size shrinks to 0 as ξ_α goes to infinity. Another interesting limit is the case where $\alpha \to 1$: as said before this cannot occur in infinite (or large) dimension as $R_\alpha \to$"d". Therefore the bound blows up in this limit also. The fact that the step-size depends on α is a weakness of our result and is due to the mixing power of the covariance eigenspaces. This dependence could be eliminated with an extra assumption on a lower bound on the decrease rate of the spectrum of the covariance matrix as made by [8]. Finally note that the rate of convergence is always strictly that $1/T$ for $\beta > 0$ and is remarkably adaptive to the possible misspecification of the problem $\beta \in (-1, 0)$.

Comparison with the literature. We can first compare to the result of the noisy version of the setting studied by [9]. Naturally, when some additive noise is assumed, averaging is necessary, and the results are weaker: γ is not adaptive to the problem, depends on the time-horizon and the rates are always slower than $O(1/T)$. However, when the noise is 0, the averaged iterates achieves the exact same convergence rate as in our theorem. The closest results to our theorem is by [6], where the same rates are shown. The important difference is that their results are given for the $\min$ of the function value and not the final iterate: our theorem solves an open problem stated in the aforementioned paper. Remark however that one superiority of [6] is that the step-size does not depend on α showing some adaptivity with this parameter. As noted in the previous paragraph, we could fix this difference with an extra assumption on the spectrum of the covariance matrix. Finally, rates of convergence for the noiseless setting have been address by [16] with some truncated version of the kernel ridge estimator. It was an open problem stated by the author to understand whether SGD could achieve those non-parametric rates. As for the $\min$, this problem seems to be properly solved now. A main difference is that the bounds of [16] suffer from some saturation in the well specified setting, when $\beta > 0$, whereas our estimator does not. When $\alpha - 1 \leqslant \beta \leqslant 0$, then the bounds match, but when the problem is really misspecified, for $\beta \leqslant \alpha - 1$, the bound of [16] is strictly better than ours. As we know that our result is optimal for SGD, it raises the interesting question whether some form of acceleration [10] or multiple passes over the data [21] could reach these rates.

Link with kernel regression. These capacity and source conditions are classical in the reproducing kernel Hilbert spaces community. For sake of clearness, we did not mention any specificity on the features. However, as it has been done in [30, 9], SGD can be "kernelized" and the descent becomes an optimization algorithm in the RKHS space of functions. In this context, the *capacity and source* conditions take on their full meaning: when RKHS are Sobolev spaces, they represent a smoothness condition on the features and the optimum respectively. The coefficient α would be the degree of regularity that we choose for the kernel, and β the prior we have on the optimum solution. For more details, we refer to section 4 of [21] or section 3.1 of [6].

4 Development of the SGD recursions and proof outline

In this section we provide an overview of the arguments that comprise the proof of our results.

Recursions for multiplicative noise. We can rewrite the SGD recursion Eq. (3) for the deviation to the optimum $\eta_t := \theta_t - \theta_*$ as a descent on the *risk* plus a multiplicative noise term. For $t \geqslant 1$,

$$\eta_{t+1} = \left(I - \gamma x_t x_t^\top\right)\eta_t = \left(I - \gamma H\right)\eta_t + \gamma\left(H - x_t x_t^\top\right)\eta_t. \tag{8}$$

We can also write the recursion satisfied by the covariance of the iterates, $M_t := \mathbb{E}\left[\eta_t \eta_t^\top\right]$:

$$M_{t+1} = (I - \gamma H)\, M_t\, (I - \gamma H) + \gamma^2 \mathbb{E}\left[\left(H - x_t x_t^\top\right) M_t \left(H - x_t x_t^\top\right)\right]. \tag{9}$$

Note a main difference between the two recursions, one is a deterministic recursion over operators whereas the other one is a stochastic recursion on vectors.

Deviations sequence and initial conditions. We emphasis that η_t and M_t represent *deviations* to the optimum θ_*. This is why proving that θ_t goes to θ_* corresponds to η_t and M_t going to 0. As the initial point of the SGD recursion is $\theta_0 = 0$, the initial conditions $\eta_{t=0}$ and $M_{t=0}$ represent, in fact, respectively the optimum and the covariance at optimum.

Eigenspaces of the covariance. The core of the analysis rests on a reformulation of our problem in the eigenspaces of the covariance operator. Recall that $(\lambda_i, v_i)_i$ are the sorted eigenelements of H. Define also the decomposition of M_t in the basis of $v_i v_j^\top$, $M_t := \sum_i m_i^t v_i v_i^\top + \sum_{i \neq j} m_{ij}^t v_i v_j^\top$. We can write the expected risk of the estimator given by the t-th iterate of SGD, that we denote by f_t:

$$\mathsf{f}_t := \mathbb{E}\mathcal{R}(\theta_t) = \frac{1}{2}\mathrm{Tr}\left(M_t H\right) = \frac{1}{2}\sum_i \lambda_i m_i^t. \tag{10}$$

It is quite remarkable that the $(m_{ij})_{i \neq j}$ do not play any role in the recursions. The aim now is to write a recursion for the m_i^t that leads to a recursion for the function value f_t through Eq. (10). This is the purpose of the following and central lemma, whose proof can be found in Appendix B.1.

Lemma 4 (Recursion on the covariance of iterates) *Define* $\mathsf{f}_i^t := \mathbb{E}\left[\langle v_i, X\rangle^2 X^\top M_t X\right]$. *For $t \geqslant 1$, we have the following recursion on the covariance of the iterates of SGD,*

$$m_i^{t+1} = m_i^t - 2\gamma\lambda_i m_i^t + \gamma^2 \mathsf{f}_i^t, \tag{11}$$

$$m_i^{t+1} = \left(1 - 2\gamma\lambda_i\right)^t m_i^0 + \gamma^2 \sum_{k=0}^{t} \left(1 - 2\gamma\lambda_i\right)^{t-k} \mathsf{f}_i^k, . \tag{12}$$

Let us make comments on this important lemma. Firstly, note that the only difference with the deterministic recursion (gradient descent) is the presence of the "mixing terms" $(\mathsf{f}_i^t)_i$. In fact, these terms affect dramatically the dynamics. Indeed, there is no reason that the iterates decrease along the iterations: this is what makes the analysis more tedious for the final iterate. On the contrary, the usual convergence rate for the *averaged iterates* is easily obtained by summing Eq. (11) and comparing the terms. Note finally that the strength of Eq. (11) lies in the fact that the different eigenspaces (*e.g.* m_i^t and m_j^t for $i \neq j$) interact only through f_i^t and f_j^t. By summing Eq. (12), and appropriately controlling sums of f_i^t's with Assumptions 1, 3 or 5, we can get recursive inequalities on f_t. These show that f_t is in fact a Lyapunov function ; going further, we use discrete versions of Gronwall-type inequalities to finally upper-bound it. To exemplify this reasoning, we present, in the following lemma, a version of Lyapunov control on f_t using Assumptions 1. Its proof can be found in Appendix B.2.

Lemma 5 (Recursion on f_t) *Under Assumption 1, assume $\gamma \leqslant (4\lambda_{\max})^{-1}$. For $t \geqslant 1$ we have the following recursion on the function value f_t. For all i,*

$$\mathsf{f}_t \leqslant \frac{\mathrm{Tr}(M_0)}{4\gamma t} + \gamma\, R \sum_{k=0}^{t-1} \frac{\mathsf{f}_k}{t-k} \tag{13}$$

The decrease of the function value f_t is controlled by the sum of a bias term –characterizing how fast the initial conditions m_i^0 are forgotten–, and a variance term –characterizing how the noise reverberates through the iterates. All the theorems are proven using the same technique: the aim is to use the inequality recursively to control the variance term. And the different Assumptions 1, 3, 5 lead to different variance term. Trying to factorize the proofs in one general bound does not allow for a clear presentation of the results. This is why, despite the apparent redundancy of the proof technique, for the sake of clarity and easy reading, we preferred to split the different proofs of the theorem and factorize only certain technical lemmas. All the proofs of Theorems 1, 2, 3 can be found in Appendix C, respectively in Sections C.1, C.2, C.3.

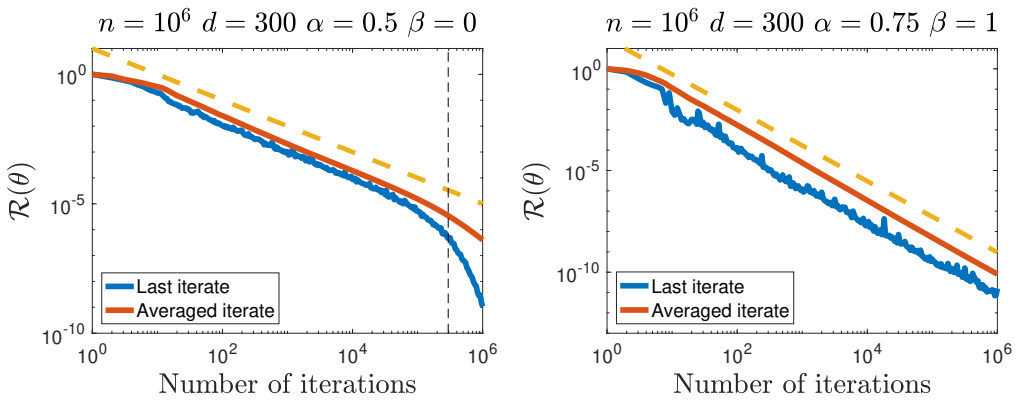

Figure 1: Least-squares regression. Left: $\alpha = 0.5$ and $\beta = 0$. The vertical dashed line marks the transition to the linear convergence regime. Right: $\alpha = 0.75$ and $\beta = 1$. The orange dashed line represents the curve $1/T^{1+\alpha\wedge\beta}$ predicted by Theorem 3.

5 Experiments

We illustrate our theoretical results on a synthetic least-squares problem using the SGD algorithm defined in Eq. (3). For $d = 300$ we consider a stream of normally distributed inputs $(x_n)_n$ whose covariance matrix H has random eigenvectors v_i and eigenvalues $1/i^{1/(1-\alpha)}$ for $i = 1, \dots, d$. The optimum is chosen randomly: $\theta_* = \sum 1/i^{\frac{1-\beta/(1-\alpha)}{2}} v_i$. This allows to reproduce the setting where the coefficient α and β of the capacity and source conditions are perfectly controlled. The outputs $(y_n)_n$ are generated through $y_n = \langle \theta_*, x_n \rangle$. We take a step-size $\gamma = \frac{1}{2\mathrm{Tr}H}$. All results are averaged over ten repetitions. We compare the performance of the last iterate and the averaged iterate of SGD on two different problems: one corresponding to $\alpha = 0.5$ and $\beta = 0$ and the other to $\alpha = 0.75$ and $\beta = 1$. For those problems, the predicted convergence is given by Theorem 3: $O(1/T^{1+\alpha\wedge\beta})$. Hence we expect a $1/T$ convergence in the first example and a fast $1/T^{1.75}$ rates for the second.

Firstly note that the bound of Theorem 3 is perfectly matched by these two examples. Secondly, we can see that, the averaged SGD and the final iterate show the same behavior quite accordingly to the theory (see comments on Theorem 3) even if we can notice a slightly -but real- better performance (result are shown in log-scale) from the final iterate. However, the main difference is that, as the averaged iterates show some saturation, the final iterate meets a point where it changes to a linear convergence regime (left plot in Figure 1). This adaptivity of the final iterate appearing after time scale $\tau \sim 1/(\gamma\lambda_{\min})$ represents a true asset in choosing the final iterate versus the averaged one. Finally notice that the linear rate regime cannot be seen in the right plot as the maximum number of iterations shown, $n = 10^6$, is negligible when compared to $\tau \sim d^{1/(1-\alpha)} = 300^{1/(1-0.75)} \sim 10^{10}$.

6 Conclusion and Perspectives

In this paper, we proceed to a detailed analysis of the convergence rates of SGD in the noiseless least-squares setting. We derive a systematic study of the SGD last iterate that leverages a sequence of fine-grained assumptions allowing us to exhibit fast polynomial rates. The absence of additive noise changes dramatically the behavior of SGD: no decaying step sizes or averaging are needed for convergence as constant step-size SGD naturally adapts to the problem. The development of the SGD recursion as an interacting particle systems open many perspectives: can it be used to analyze SGD accelerations? multiple passes over the data? Another significant result is that, from an optimization perspective, we give a prototypal example where the last iterate of constant step-size SGD provably converges in the non-strongly convex case. This raises the following fundamental question: can we extend our method to show the convergence of SGD last iterate in the non-strongly convex case?

References

[1] D. Aldous and D. Lanoue. A lecture on the averaging process. *Probability Surveys*, 9(none):90 – 102, 2012.

[2] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

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

[4] R. Bassily, M. Belkin, and S. Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

[5] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116 (32):15849–15854, 2019.

[6] R. Berthier, F. Bach, and P. Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. In *Advances in Neural Information Processing Systems*, 2020.

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

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

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

[10] 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.

[11] N. Flammarion and F. Bach. Stochastic composite least-squares regression with convergence rate $o(1/n)$. *Conference on Learning Theory*, pages 831–875, 2017.

[12] R. Ge, S. M. Kakade, R. Kidambi, and P. Netrapalli. The step decay schedule: a near optimal, geometrically decaying learning rate procedure for least squares. *Advances in Neural Information Processing Systems*, 32:14977–14988, 2019.

[13] N. J. A. Harvey, C. Liaw, Y. Plan, and S. Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

[14] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pages 545–604, 2018.

[15] P. Jain, D. Nagaraj, and P. Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory*, pages 1752–1755. PMLR, 2019.

[16] K. Jun, A. Cutkosky, and F. Orabona. Kernel truncated randomized ridge regression: optimal rates and low noise acceleration. *Advances in neural information processing systems*, 32, 2019.

[17] C. Liu and M. Belkin. Accelerating sgd with momentum for over-parameterized learning. In *International Conference on Learning Representations*, 2020.

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

[19] S. Y. Meng, S. Vaswani, I. H. Laradji, M. Schmidt, and S. Lacoste-Julien. Fast and furious convergence: stochastic second order methods under interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 1375–1386. PMLR, 2020.

[20] N. Mücke and E. Reiss. Stochastic gradient descent in Hilbert scales: smoothness, preconditioning and earlier stopping. *CoRR*, abs/2006.10840, 2020.

[21] L. Pillaud-Vivien, A. Rudi, and F. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31:8114–8124, 2018.

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

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

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

[25] O. Shamir. Open problem: is averaging needed for strongly convex stochastic gradient descent? In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 47.1–47.3. JMLR Workshop and Conference Proceedings, 25–27 Jun 2012.

[26] O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

[27] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[28] M. V. Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11(1):23–35, 1998.

[29] I. Steinwart, D. R. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Conference On Learning Theory*. PMLR, 2009.

[30] P. Tarres and Y. Yao. Online learning as stochastic approximation of regularization paths: optimality and almost-sure convergence. *IEEE Trans. Inf. Theory*, 60(9):5716–5735, 2014.

[31] P. Tseng. An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.

[32] S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

[33] S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien. Painless stochastic gradient: interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems*, pages 3727–3740, 2019.

[34] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *International Conference on Machine Learning*, 2004.

[35] D. Zou, J. Wu, V. Braverman, Q. Gu, and S. M. Kakade. Benign overfitting of constant-stepsize sgd for linear regression. *COLT*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] , In our contributions list in the introduction, for point one check Theorems 1, 2, Theorem 3 for the second point and Lemmas 4, 5 for the remaining.
   (b) Did you describe the limitations of your work? [Yes] , see comments in 'comparison with literature' after Theorem 3.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] , our work is theoretical in nature and has neither direct or indirect negative societal impacts stemming from either methodology or application.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] .
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] , see assumptions from 1 to 6.
   (b) Did you include complete proofs of all theoretical results? [Yes] , we have included them in Appendix A,B,C.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] , we include all the necessary material to reproduce our experiment results.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] , we specify all the details, see 5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] , our training is not computationally intensive and can be run on any laptop or PC.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] , we use synthetic date and run SGD on it. Hence, this is not applicable in our case.
   (b) Did you mention the license of the assets? [N/A] , we haven't used any assets.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] , we haven't used any assets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] , irrelevant in our case as we haven't used or curated any data.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] , we use a synthetic data. Hence, it is not applicable for this work.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] , it is not applicable for us as we haven't crowd-sourced any data.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

Organization of the Appendix

In Appendix A we present a detailed summary of our results and of the different assumptions. In Appendix B we prove Lemma 4 and 5 which enable to obtain the main SGD recursions. Then in Appendix C we prove our main results: Theorems 1, 2 and 3.

A Detailed summary of the results and of the different assumptions

A.1 Capacity and Source Conditions and Results

Summary on the assumptions. All the assumptions we discussed in the article are summarized in Tables 2 and 3. It clearly shows that the assumptions go stricter and stricter revealing a finer description of the problem. Note also that they can be paired two by two, one sequence corresponding to the features and the other one to the optimum: Assumption 1 with 2, Assumption 3 with 4 and Assumption 5 with 6. The case $\beta \in (-1, 0)$, when the optimum is non-attainable, is left aside in the panels for the sake of clarity.

Assumption	Name	Condition
Assumption **1**	Fourth Moment Condition	$\mathbb{E}\left[\|X\|^2 \; XX^\top\right] \preccurlyeq RH$
$\Uparrow$		$\Uparrow$
Assumption **3**	Features $\log$-regularity	$\mathbb{E}\left[\left\langle X, \ln\left(H^{-1}\right)X\right\rangle \; XX^\top\right] \preccurlyeq R_{\ln} H$
$\Uparrow$		$\Uparrow$
Assumption **5**	Capacity condition	$\mathbb{E}\left[\left\|H^{-\alpha/2}X\right\|^2 \; XX^\top\right] \preccurlyeq R_\alpha H, \alpha \in (0,1)$

Table 2: Table showing conditions and implications for each Assumption on the features.

Assumption	Name	Condition	Finite constant
Assumption **2**	Attainable case	$\|\theta_*\|_{\mathcal{H}} < +\infty$	$\|\theta_*\|_{\mathcal{H}}$
$\Uparrow$		$\Uparrow$	
Assumption **4**	Optimum $\log$-regularity	$\mathrm{Tr}(M_0 \ln(H^{-1})) < +\infty$	$C_{\ln}$
$\Uparrow$		$\Uparrow$	
Assumption **6**	Source condition	$\mathrm{Tr}(M_0 H^{-\beta}) < +\infty, \; \beta > 0$	C_β

Table 3: Table showing conditions and implications for each Assumption on the optimum.

Summary on the results. Each pair of assumptions (on the features and on the optimum) corresponds to a given rate of convergence. This is what is summarized in the Table 4. Stricter assumptions naturally lead to enhanced convergence rates or bigger step-sizes. Going from a theorem to the next one, the only difference is that we strengthen the assumptions to improve either the convergence rate or the step-size. Every theorem is a bound on the *expected risk* given by the *last iterate* of the SGD recursion with *constant* step size γ: $\mathbb{E}\mathcal{R}(\theta_T) = \frac{1}{2}\mathbb{E}\,\mathrm{Tr}\left[(\theta_T - \theta_*)(\theta_T - \theta_*)^\top H\right]$, where $\mathbb{E}$ stands for the expectation with respect to the stream of data used by the SGD algorithm until time T. Only for the Theorem 1, for which the assumptions are the weakest, we allow the step-size to depends on the finite horizon T (through its logarithm). All the theorems are stated with respect to finite constants defined thanks to the different assumptions. The reader can refer to Tables 2 and 3 for a concise summary of them. All the proofs have been written in Appendix C.

Details on $\log$-regularity Note that we have stated Assumption 3 with a reference λ_{o} and also a condition $\lambda_{\mathrm{o}} \geq 7R_{\ln}$ for better presentation of results. Here we try to describe this better and also the choice of λ_0. Lets make an assumption with out any dependence on λ_{o} that there exist a constant $\tilde{R}_{\ln} \geq 0$ as follows:

$$\mathbb{E}\left[\left\langle X, \ln\left(H^{-1}\right)X\right\rangle XX^\top\right] \preccurlyeq \tilde{R}_{\ln} H. \tag{14}$$

Assumptions → Theorem	**Rate**	**Step-size**
Assumption **1**, **2** → Theorem **1**	$\ln(T)/T$	$O(1/\ln(T))$
$\Uparrow$		
Assumption **3**, **4** → Theorem **2**	$1/T$	$O(1)$
$\Uparrow$		
Assumption **5**, **6** → Theorem **3**	$1/T^{1+\alpha}$	$O(1)$

Table 4: Table showing different upper-bounds for the convergence of the SGD final iterate given different assumptions.

Note that this is a stricter than 1. Hence we assumme that it holds with finite constant $\tilde{R}$ i.e.

$$\begin{aligned}
\mathbb{E}\left[\|X\|^2 XX^\top\right] &\preccurlyeq \tilde{R}H \\
\mathbb{E}\left[\left\langle X, \ln\left(\lambda_{\mathrm{o}} H^{-1}\right) X\right\rangle XX^\top\right] &= \ln\lambda_{\mathrm{o}}\mathbb{E}\left[\left\langle X, \ln\left(H^{-1}\right) X\right\rangle XX^\top\right] + \mathbb{E}\left[\|X\|^2 XX^\top\right] \\
&\preccurlyeq \left(\tilde{R}\lambda_{\mathrm{o}} + \tilde{R}_{\ln}\right) H.
\end{aligned}$$

Hence with $R_{\ln} = \tilde{R}\ln\lambda_{\mathrm{o}} + \tilde{R}_{\ln}$, it satisfies Assumption 3. Now coming to the condition, we need

$$7\left(\tilde{R}\ln\lambda_{\mathrm{o}} + \tilde{R}_{\ln}\right) \leq \lambda_{\mathrm{o}}.$$

Without loss of generality assume $\tilde{R}_{\ln} \geq \tilde{R}$, $\ln\left(\ln\tilde{R}_{\ln}\right) \leq \ln\tilde{R}_{\ln}$ and $\ln\tilde{R}_{\ln} \geq 1$. Let $\lambda_o = 50\tilde{R}_{\ln}\ln\tilde{R}_{\ln}$, then

$$\begin{aligned}
7\left(\tilde{R}\ln\lambda_{\mathrm{o}} + \tilde{R}_{\ln}\right) &\leqslant 7\left(\tilde{R}_{\ln}\ln\left(50\tilde{R}_{\ln}\ln\tilde{R}_{\ln}\right)\right) + 7\tilde{R}_{\ln} \\
&\leq 7\ln 50\tilde{R}_{\ln} + 7\tilde{R}_{\ln}\ln\tilde{R}_{\ln} + 7\tilde{R}_{\ln}\ln\left(\ln\tilde{R}_{\ln}\right) + 7\tilde{R}_{\ln} \\
&\leq (7\ln 50 + 21)\tilde{R}_{\ln}\ln\tilde{R}_{\ln} \leq 50\tilde{R}_{\ln}\ln\tilde{R}_{\ln} \leq \lambda_{\mathrm{o}}.
\end{aligned}$$

which satisfies the required condition. Hence we can always find such a λ_{o} if $\tilde{R}_{\ln}$ is finite.

Capacity condition and Eigen decay Here we try to further describe the link between our capacity conditions and the eigen value decay. The capacity condition we had in Assumption.5,

$$\begin{aligned}
\mathbb{E}\left[\left\langle X, H^{-\alpha}X\right\rangle XX^\top\right] &\preccurlyeq R_\alpha H, \\
\mathbb{E}\left[\left\langle X, H^{-\alpha}X\right\rangle XX^\top H^{-\alpha}\right] &\preccurlyeq R_\alpha H^{1-\alpha}, \qquad \because H \text{ is PSD} \\
\mathrm{Tr}\big(\mathbb{E}\left[\left\langle X, H^{-\alpha}X\right\rangle XX^\top H^{-\alpha}\right]\big) &\leqslant \mathrm{Tr}\big(R_\alpha H^{1-\alpha}\big), \\
\mathbb{E}\left[\left\langle X, H^{-\alpha}X\right\rangle \mathrm{Tr}XX^\top H^{-\alpha}\right] &\leqslant R_\alpha \mathrm{Tr}\big(H^{1-\alpha}\big) \qquad \text{Using } \mathrm{Tr}XX^\top H^{-\alpha} = \left\langle X, H^{-\alpha}X\right\rangle, \\
\mathbb{E}\left[\left\langle X, H^{-\alpha}X\right\rangle^2\right] &\leq R_\alpha \mathrm{Tr}\big(H^{1-\alpha}\big).
\end{aligned}$$

Now using Cauchy-Schwarz we have,

$$\left(\mathrm{Tr}H^{1-\alpha}\right)^2 = \mathrm{Tr}\big(\mathbb{E}\left[XX^\top\right]H^{-\alpha}\big)^2 = \mathbb{E}\left[\left\langle X, H^{-\alpha}X\right\rangle\right]^2 \leq \mathbb{E}\left[\left\langle X, H^{-\alpha}X\right\rangle^2\right] \leq R_\alpha \mathrm{Tr}\big(H^{1-\alpha}\big).$$

Hence, $\mathrm{Tr}H^{1-\alpha} \leq R_\alpha$.

So from Assumption 5 we have $\mathrm{Tr}H^{1-\alpha} \leqslant R_\alpha$. Using the fact that the eigen values are enumerated in sorted non-increasing order, we have

$$i\lambda_i^{1-\alpha} \leq \sum_{j\leqslant i}\lambda_j^{1-\alpha} \leqslant \mathrm{Tr}H^{1-\alpha} \leqslant R_\alpha$$

which gives that λ_i is of order $O(1/i^{\frac{1}{1-\alpha}})$. Similarly if we do the analysis for Assumption 3, it implies $\mathrm{Tr}\big(H\ln(\lambda_0 H^{-1})\big) \leq R_{\ln}$.

B Proofs of the SGD recursions: Lemma 4 and 5

B.1 Proof of Lemma 4

First recall that for $t \geqslant 1$, Eq. (9) gives that

$$M_{t+1} = (I - \gamma H)\, M_t\, (I - \gamma H) + \gamma^2 \mathbb{E}\left[\left(H - x_t x_t^\top\right) M_t \left(H - x_t x_t^\top\right)\right].$$

Lets compute $\mathbb{E}\left[\left(H - x_t x_t^\top\right) M_t \left(H - x_t x_t^\top\right)\right]$. Using that $\mathbb{E}[x_t x_t^\top] = H$, we first write,

$$\begin{aligned}
\mathbb{E}\left[\left(H - x_t x_t^\top\right) M_t \left(H - x_t x_t^\top\right)\right] &= \mathbb{E}\left[H M_t H\right] - \mathbb{E}\left[x_t x_t^\top M_t H\right] - \mathbb{E}\left[H M_t x_t\, x_t^\top\right] + \mathbb{E}\left[x_t x_t^\top M_t\, x_t x_t^\top\right] \\
&= H M_t H - H M_t H - H M_t H + \mathbb{E}\left[x_t^\top M_t x_t\;\; x_t x_t^\top\right] \\
&= \mathbb{E}\left[x_t^\top M_t x_t\, x_t x_t^\top\right] - H M_t H.
\end{aligned}$$

Now using this in Eq. (9) we have

$$\begin{aligned}
M_{t+1} &= (I - \gamma H)\, M_t\, (I - \gamma H) + \gamma^2 \left[\mathbb{E}\left[x_t^\top M_t x_t\, x_t x_t^\top\right] - H M_t H\right] \\
&= M_t - \gamma H M_t - \gamma M_t H + \gamma^2 \mathbb{E}\left[x_t^\top M_t x_t\, x_t x_t^\top\right].
\end{aligned}$$

Now we project the above term on $v_i v_i^\top$

$$\begin{aligned}
v_i^\top M_{t+1} v_i &= v_i^\top M_t v_i - \gamma v_i^\top H M_t v_i - \gamma v_i^\top M_t H v_i + \gamma^2 v_i^\top \mathbb{E}\left[x_t^\top M_t x_t\, x_t x_t^\top\right] v_i \\
&= v_i^\top M_t v_i - 2\gamma \lambda_i v_i^\top M_t v_i + \gamma^2 \mathbb{E}\left[\langle v_i, x_t\rangle^2\, x_t^\top M_t x_t\right] \qquad \text{Using } H v_i = \lambda_i v_i
\end{aligned}$$

Note from the definition of $M_i^t = \sum_i m_i^t v_i v_i^\top + \sum_{ij} m_{ij}^t v_i v_j^\top$, we can see $v_i^\top M_t v_i^\top = m_i^t$, $v_i^\top M_{t+1} v_i^\top = m_i^{t+1}$, $v_j^\top M_t v_j^\top = m_j^t$. Now we have the following recursion,

$$m_i^{t+1} = m_i^t - 2\gamma \lambda_i m_i^t + \gamma^2 \mathbb{E}\left[\langle v_i, x_t\rangle^2\, x_t^\top M_t x_t\right]$$

Using notation $\mathsf{f}_i^t = \mathbb{E}\left[\langle v_i, x_t\rangle^2\, x_t^\top M_t x_t\right]$, we prove the Eq.(11). This proves a part of Lemma.(4). The second part we prove using recursion Eq. (12) by induction. For base case $t = 1$ we can see that lemma holds directly from Eq. (10). Assume the lemma holds for $k = t - 1$. We know from Eq. (11) that

$$m_i^t = (1 - \gamma \lambda_i)\, m_i^{t-1} + \gamma^2 \mathsf{f}_i^t,$$

and from the induction hypothesis, we have

$$m_i^{t-1} = (1 - \gamma \lambda_i)^{t-1}\, m_i^0 + \gamma^2 \sum_{k=0}^{t-2} \lambda_i\, (1 - \gamma \lambda_i)^{t-k-2}\, \mathsf{f}_i^k.$$

Merging these two inequalities we have,

$$\begin{aligned}
m_i^t &= (1 - \gamma \lambda_i) \left((1 - \gamma \lambda_i)^{t-1}\, m_i^0 + \gamma^2 \sum_{k=0}^{t-2} \lambda_i\, (1 - \gamma \lambda_i)^{t-k-2}\, \mathsf{f}_i^k \right) + \gamma^2 \lambda_i \mathsf{f}_i^k \\
&= (1 - \gamma \lambda_i)^t\, m_i^0 + \gamma^2 \sum_{k=0}^{t-1} \lambda_i\, (1 - \gamma \lambda_i)^{t-k-1}\, \mathsf{f}_i^k.
\end{aligned}$$

This completes the proof of Lemma 4.

B.2 Proof of Lemma 5

Here we prove the Lemma 5. First note from Eq. (12) of Lemma 4 we have the following recursion,

$$m_i^t = (1-2\gamma\lambda_i)^t\, m_i^0 + \gamma^2 \sum_{k=0}^{t-1} (1-2\gamma\lambda_i)^{t-k-1}\, \mathsf{f}_i^k. \tag{15}$$

Again for $\gamma \leq (4\lambda_{\max})^{-1}$, we have

$$m_i^t \leq (1-2\gamma\lambda_i)^t\, m_i^0 + 2\gamma^2 \sum_{k=0}^{t-1} (1-2\gamma\lambda_i)^{t-k}\, \mathsf{f}_i^k \tag{16}$$

$$\lambda_i m_i^t \leq \lambda_i\, (1-2\gamma\lambda_i)^t\, m_i^0 + 2\gamma^2 \sum_{k=0}^{t-1} \lambda_i\, (1-2\gamma\lambda_i)^{t-k}\, \mathsf{f}_i^k \tag{17}$$

For $x \in (0,1)$, we have $x(1-x)^t \leq xe^{-xt} \leq 1/t$. Now using the following natural bounds, we get

$$\lambda_i(1-2\gamma\lambda_i)^{t-k} \leqslant \frac{1}{2\gamma(t-k)}$$

$$\lambda_i m_i^t \leq \frac{m_i^0}{2\gamma t} + \gamma \sum_{k=0}^{t-1} \frac{\mathsf{f}_i^k}{t-k}$$

$$\sum_i \lambda_i m_i^t \leqslant \sum_i \frac{m_i^0}{2\gamma t} + \gamma \sum_{k=0}^{t-1} \left(\sum_i \mathsf{f}_i^k\right) \frac{1}{t-k} \tag{18}$$

Now lets compute the sum below. Using $\mathsf{f}_i^t = \mathbb{E}\left[\langle v_i, x_t\rangle^2\, x_t^\top M_t x_t\right]$, we have

$$\begin{aligned}
\sum_i \mathsf{f}_i^t &= \sum_i \mathbb{E}\left[\langle v_i, x_t\rangle^2\, x_t^\top M_t x_t\right] \\
&= \mathbb{E}\left[\left(\sum_i \langle v_i, x_t\rangle^2\right) x_t^\top M_t x_t\right] \\
&= \mathbb{E}\left[\|x_t\|^2\, \mathrm{Tr}\big(x_t x_t^\top M_t\big)\right] \\
&= \mathrm{Tr}\Big[\mathbb{E}\left[\|x_t\|^2\, x_t x_t^\top\right] M_t\Big]
\end{aligned}$$

Note for any two PSD matrices A, B if $A \preccurlyeq B$ then $\mathrm{Tr}(AM) \leq \mathrm{Tr}(BM)$, for any PSD matrix M. From Assumption 1, we have

$$\begin{aligned}
\mathbb{E}\left[\|x_t\|^2\, x_t x_t^\top\right] &\preccurlyeq RH \\
\mathrm{Tr}\Big[\mathbb{E}\left[\|x_t\|^2\, x_t x_t^\top\right] M_t\Big] &\leqslant R\mathrm{Tr}(M_t H) \\
\sum_i \mathsf{f}_i^t &\leqslant R\mathrm{Tr}(M_t H)
\end{aligned}$$

Now replacing the sum in Eq.(18), we get

$$\sum_i \lambda_i m_i^t \leqslant \sum_i \frac{m_i^0}{2\gamma t} + \gamma\, R \sum_{k=0}^{t-1} \frac{\mathrm{Tr}(M_k H)}{t-k}$$

Using Eq.(10), we have the recursion

$$2\mathsf{f}_t \leq \frac{\mathrm{Tr}(M_0)}{2\gamma t} + 2\gamma\, R \sum_{k=0}^{t-1} \frac{\mathsf{f}_k}{t-k}$$
$$\mathsf{f}_t \leq \frac{\mathrm{Tr}(M_0)}{4\gamma t} + \gamma\, R \sum_{k=0}^{t-1} \frac{\mathsf{f}_k}{t-k}$$

This proves Lemma 5.

C Proofs of the main results

C.1 Proof of Theorem 1

From Lemma 5, we have for all $t \geqslant 1$,

$$\mathsf{f}_t \leqslant \frac{\mathrm{Tr}(M_0)}{4\gamma t} + \gamma R \sum_{k=0}^{t-1} \frac{\mathsf{f}_k}{t-k}.$$

And then, the technique we use throughout all the proofs of this section rest on a control of the second term $\sum_{k=0}^{t-1} \frac{\mathsf{f}_k}{(t-k)}$ using the recursion. Indeed, by summing,

$$\begin{aligned}
\sum_{t=0}^{T-1} \frac{\mathsf{f}_t}{T-t} &\leqslant \frac{\mathsf{f}_0}{T} + \frac{\mathrm{Tr}(M_0)}{4\gamma} \sum_{t=1}^{T-1} \frac{1}{t(T-t)} + \gamma R \sum_{t=1}^{T-1} \sum_{k=0}^{t-1} \frac{\mathsf{f}_k}{(t-k)(T-t)} \\
&\leqslant \frac{\mathsf{f}_0}{T} + \frac{\mathrm{Tr}(M_0)}{4\gamma} \sum_{t=1}^{T-1} \frac{1}{t(T-t)} + \gamma R \sum_{t=1}^{T-1} \sum_{k=0}^{t-1} \frac{\mathsf{f}_k}{(t-k)(T-t)} \\
&= \frac{\mathsf{f}_0}{T} + \frac{\mathrm{Tr}(M_0)}{4\gamma T} \sum_{t=1}^{T-1} \left[\frac{1}{t} + \frac{1}{T-t}\right] + \gamma R \sum_{k=0}^{T-2} \mathsf{f}_k \sum_{t=k+1}^{T-1} \frac{1}{(t-k)(T-t)} \\
&\leqslant \frac{\mathsf{f}_0}{T} + \frac{\mathrm{Tr}(M_0)\ln(T)}{2\gamma T} + 2\gamma R \ln(T) \sum_{k=0}^{T-1} \frac{\mathsf{f}_k}{T-k}.
\end{aligned}$$

Noting that $\mathsf{f}_0 = \sum_i \lambda_i m_i^0 \leqslant \lambda_{\max}\mathrm{Tr}(M_0) \leqslant \frac{\mathrm{Tr}(M_0)}{2\gamma}$, for $T \geqslant 2$,

$$\begin{aligned}
\sum_{t=0}^{T-1} \frac{\mathsf{f}_t}{T-t} &\leqslant \frac{\mathrm{Tr}(M_0)\ln(T)}{\gamma T} + 2\gamma R \ln(T) \sum_{k=0}^{T-1} \frac{\mathsf{f}_k}{T-k} \\
(1 - 2\gamma R \ln(T)) \sum_{t=0}^{T-1} \frac{\mathsf{f}_t}{T-t} &\leqslant \frac{\mathrm{Tr}(M_0)\ln(T)}{\gamma T}.
\end{aligned}$$

Hence, for $\gamma = (4R\ln(T))^{-1}$, we have

$$\sum_{t=0}^{T-1} \frac{\mathsf{f}_t}{T-t} \leqslant \frac{2\mathrm{Tr}(M_0)\ln(T)}{\gamma T},$$

and, hence, for all $T \geqslant 1$,

$$\begin{aligned}
\mathsf{f}_T &\leqslant \frac{\mathrm{Tr}(M_0)}{4\gamma T} + \gamma R \sum_{k=0}^{T-1} \frac{\mathsf{f}_k}{T-k} \\
&\leqslant \frac{R\mathrm{Tr}(M_0)\ln(T)}{T} + \frac{2R\mathrm{Tr}(M_0)\ln(T)}{T} \\
&\leqslant 3R\mathrm{Tr}(M_0)\frac{\ln(T)}{T}.
\end{aligned}$$

That concludes the proof of Theorem 1.

C.2 Proof of Theorem 2

The proof of this theorem follows the same principle but uses slightly better estimations. Indeed, we replace the previous $1/n$ bound by a finer bound. This is the statement of the following lemma. For this, we define the following summation $S_n(x)$ for some $x \in (0, 1/4]$ and $n \geq 2$,

$$S_n(x) := \sum_{k=0}^{n-1} \frac{(1-x)^k}{n-k}. \tag{19}$$

Lemma 6 *For any $x \in (0, 1/4]$, $n \geq 1$, we have the following bound*

$$xS(x) \leqslant \frac{7\ln(1/x)}{n}. \tag{20}$$

Proof

In the first step, we slightly reformulate this expression, then try to bound that expression by a continuous integral which gives the desired result. From Eq. (19), we have

$$\begin{aligned} S_n(x) &= \sum_{k=0}^{n-1} \frac{(1-x)^k}{n-k} = (1-x)^n \sum_{k=0}^{n-1} \frac{(1-x)^{k-n}}{n-k} \\ (1-x)^{-n} S_n(x) &= \sum_{k=1}^{n} \frac{(1-x)^{-k}}{k} \end{aligned}$$

We compute $(1-x)^{-n} S_n(x)$. Note that $g(y) = (1-x)^{-y}/y$ is convex for $y > 0$. For any convex function g, we have that for any integer $k \geq 2$, $y \in [k, k+1]$ we have that $g(k) \leq g(y) + g(y-1)$. Indeed, if g is in the increasing phase then $g(y)$ dominates $g(k)$, else $g(y-1)$ dominates in the decreasing phase. Using this we can see that for $k = 2 \cdots n$,

$$\begin{aligned} g(k) &\leq \int_k^{k+1} (g(y) + g(y-1))\, dy \\ \sum_{k=2}^{n-1} \frac{(1-x)^{-k}}{k} &\leq \int_2^n (g(y) + g(y-1))\, dy. \end{aligned}$$

Hence,

$$(1-x)^{-n} S_n(x) = \sum_{k=1}^{n} \frac{(1-x)^{-k}}{k} \leq (1-x)^{-1} + 2\int_1^n \frac{(1-x)^{-y}}{y} + \frac{(1-x)^{-n}}{n}\, dy. \tag{21}$$

This leads us to try to bound the integral $\int_1^{n+1} \frac{(1-x)^{-y}}{y}$. Using the change of variable $(1-x)^{-y} = e^t$. We can rewrite the above integral as follows:

$$\int_1^n \frac{(1-x)^{-y}}{y}\, dy = \int_{-\ln(1-x)}^{-n\ln(1-x)} \frac{e^t}{t}\, dt.$$

For the sake of clearer notations, let us define $a := -\ln(1-x)$ such that as $0 < x < 1$, we have $a > 0$. The equation simplifies to the following

$$\begin{aligned} \int_1^n \frac{(1-x)^{-y}}{y} &\leq \int_a^{na} \frac{e^t}{t}\, dt \\ &\leq \int_a^1 \frac{e^t}{t}\, dt + \int_1^{na} \frac{e^t}{t}\, dt \end{aligned}$$

For $t \leq 1$, we can see that $e^t/t \leq e/t$, where as for $t \geq 2$ we use the bound $e^t/t \leq 2e^t(t-1)/t^2$. Using these bounds,

$$\begin{aligned}
\int_1^n \frac{(1-x)^{-y}}{y} &\leq e\int_a^1 \frac{1}{t}\,dt + \int_1^2 \frac{e^t}{t} + 2\int_2^{na} \frac{(t-1)e^t}{t^2}\,dt \\
&\leq e\ln\frac{1}{a} + e^2 - e + 2\left[\frac{e^x}{x}\right]_2^{na} \\
&\leq e\ln\frac{1}{a} + e^2 - e + 2\left[\frac{e^{na}}{na} - \frac{e^2}{2}\right] \\
&\leq e\ln\frac{1}{a} + 2\frac{e^{na}}{na}
\end{aligned}$$

Re-substituting, $a = -\ln(1-x)$, we can see the $a \geqslant x$ and hence,

$$\int_1^n \frac{(1-x)^{-y}}{y} \leq e\ln\frac{1}{x} + 2\frac{(1-x)^{-n}}{nx},$$

such that we can write

$$\begin{aligned}
(1-x)^{-n}S_n(x) &\leq (1-x)^{-1} + 2\left[e\ln\frac{1}{x} + 2\frac{(1-x)^{-n}}{nx}\right] + \frac{(1-x)^{-n}}{n} \\
S_n(x) &\leq (1-x)^{n-1} + 2e\ln\frac{1}{x}(1-x)^{-n} + \frac{4}{n} + \frac{1}{n} \\
xS_n(x) &\leq x(1-x)^{n-1} + 2e\,x(1-x)^{-n}\,\ln\frac{1}{x} + \frac{4}{n} + \frac{x}{n}.
\end{aligned}$$

Now we use the fact that $x(1-x)^n \leq xe^{-nx} \leq 1/(en)$. Leveraging also that $x \leq 1/4$, so that $\ln\frac{1}{x} > 1$ and $(1-x)^{-1} \leq 4/3$. We can finally bound,

$$xS_n(x) \leq \frac{4}{3e}\frac{1}{n} + \frac{1}{4n} + \frac{4}{n} + 2\frac{1}{n}\ln\frac{1}{x} \leq \frac{7}{n}\ln\frac{1}{x}$$

This completes the proof.

Using the above we prove Theorem 2. Note that from 5, we need to bound $\sum_{k=0}^{t-1}\frac{\mathsf{f}_k}{t-k}$. To tightly bound this we first bound individual $\sum_{k=0}^{t-1}\frac{m_i^k}{t-k}$ and then uses these to bound the summation of function values.

Recall that for all $t \geqslant 1$, from Eq.(12) and $\gamma \leq (4\lambda_{max})^{-1}$

$$\begin{aligned}
m_i^t &= (1-2\gamma\lambda_i)^t\, m_i^0 + \gamma^2\sum_{k=0}^{t-1}(1-2\gamma\lambda_i)^{t-k-1}\,\mathsf{f}_i^k \\
m_i^t &= (1-2\gamma\lambda_i)^t\, m_i^0 + 2\gamma^2\sum_{k=0}^{t-1}(1-2\gamma\lambda_i)^{t-k}\,\mathsf{f}_i^k
\end{aligned}$$

And then, the technique we use in this section rests on a control of the second term $\sum_{k=0}^{t-1}\frac{\mathsf{f}_k}{t-k}$ using the recursion. Indeed, by summing, Then we will use carefully the bound in Lemma 6 for the second term. As said before, the only difference with the proof of the previous theorem is the special care in estimations to avoid the logarithm at the price of a slightly more stringent assumption.

$$\begin{aligned}
\sum_{t=0}^{T-1}\frac{m_i^t}{T-t} &= \sum_{t=0}^{T-1}\frac{(1-2\gamma\lambda_i)^t}{T-t}m_i^0 + 2\gamma^2\sum_{t=0}^{T-1}\sum_{k=0}^{t-1}(1-2\gamma\lambda_i)^{t-k}\,\mathsf{f}_i^k \\
&\leqslant \frac{m_i^0}{T} + \sum_{t=1}^{T-1}\frac{(1-2\gamma\lambda_i)^t}{T-t}m_i^0 + 2\gamma^2\sum_{k=0}^{T-2}\mathsf{f}_i^k\sum_{t=k+1}^{T-1}\frac{(1-2\gamma\lambda_i)^{t-k}}{T-t} \\
\sum_{t=0}^{T-1}\frac{\lambda_i m_i^t}{T-t} &\leqslant \frac{\lambda_i m_i^0}{T} + \sum_{t=1}^{T-1}\frac{(1-2\gamma\lambda_i)^t}{T-t}m_i^0 + \gamma\sum_{k=0}^{T-2}\mathsf{f}_i^k\sum_{t=1}^{T-k-1}(2\gamma\lambda_i)\frac{(1-2\gamma\lambda_i)^t}{T-k-t} \\
&\leqslant \frac{\lambda_i m_i^0}{T} + \frac{m_i^0}{2\gamma}(2\gamma\lambda_i)S_T(2\gamma\lambda_i) + \gamma\sum_{k=0}^{T-1}\mathsf{f}_i^k\,(2\gamma\lambda_i)S_{T-k}(2\gamma\lambda_i).
\end{aligned}$$

And then, by applying Lemma 6 :

$$\sum_{t=0}^{T-1} \frac{\lambda_i m_i^t}{T-t} \leqslant \frac{m_i^0}{T} + \frac{7m_i^0}{2\gamma} \ln\left(\frac{1}{2\gamma\lambda_i}\right)\frac{1}{T} + 7\gamma \sum_{k=0}^{T-1} \mathsf{f}_i^k \ \ln\left(\frac{1}{2\gamma\lambda_i}\right)\frac{1}{T-k}.$$

For now lets assume that λ_{o} be such that $2\gamma\lambda_{\mathrm{o}} \geqslant 1$, we will check this later at the end. We have,

$$\sum_{t=0}^{T-1} \frac{\lambda_i m_i^t}{T-t} \leqslant \frac{m_i^0}{T} + \frac{7m_i^0}{4\gamma} \ln\left(\frac{\lambda_{\mathrm{o}}}{\lambda_i}\right)\frac{1}{T} + 7\gamma \sum_{k=0}^{T-1} \mathsf{f}_i^k \ \ln\left(\frac{\lambda_{\mathrm{o}}}{\lambda_i}\right)\frac{1}{T-k}.$$

Now summing over i, we get

$$\sum_{t=0}^{T-1}\sum_i \frac{\lambda_i m_i^t}{T-t} \leqslant \sum_i \frac{m_i^0}{T} + \sum_i \frac{7m_i^0}{4\gamma} \ln\left(\frac{\lambda_{\mathrm{o}}}{\lambda_i}\right)\frac{1}{T} + 7\gamma \sum_{k=0}^{T-1} \left(\sum_i \mathsf{f}_i^k \ln\left(\frac{\lambda_{\mathrm{o}}}{\lambda_i}\right)\right)\frac{1}{T-k}. \tag{22}$$

Note from Eq.(10), we have $\sum_i \lambda_i m_i^t = \mathsf{f}_t$. Lets calculate the remaining terms

$$\begin{aligned}
\sum_i \mathsf{f}_i^k \ \ln\left(\frac{\lambda_{\mathrm{o}}}{\lambda_i}\right) &= \sum_i \ln\left(\frac{\lambda_{\mathrm{o}}}{\lambda_i}\right)\mathbb{E}\left[\left(\langle v_i, x_t\rangle^2\right) x_t^\top M_t x_t\right] \\
&= \mathbb{E}\left[\left(\sum_i \ln\left(\frac{\lambda_{\mathrm{o}}}{\lambda_i}\right)\langle v_i, x_t\rangle^2\right) x_t^\top M_t x_t\right] \\
&= \mathbb{E}\left[\left\langle x_t, \ln\left(\lambda_{\mathrm{o}} H^{-1}\right) x_t\right\rangle \ \mathrm{Tr}\big(x_t x_t^\top M_t\big)\right] \\
&= \mathrm{Tr}\big[\mathbb{E}\left[\left\langle x_t, \ln\left(\lambda_{\mathrm{o}} H^{-1}\right) x_t\right\rangle \ x_t x_t^\top\right] M_t\big]
\end{aligned}$$

Note for any two PSD matrices A, B if $A \preccurlyeq B$ then $\mathrm{Tr}(AM) \leq \mathrm{Tr}(BM)$, for any PSD matrix M. From Assumption 3, we have

$$\begin{aligned}
\mathbb{E}\left[\left\langle x_t, \ln\left(\lambda_{\mathrm{o}} H^{-1}\right) x_t\right\rangle \ x_t x_t^\top\right] &\preccurlyeq R_{\ln} H \\
\mathrm{Tr}\big[\mathbb{E}\left[\left\langle x_t, \ln\left(\lambda_{\mathrm{o}} H^{-1}\right) x_t\right\rangle \ x_t x_t^\top\right] M_t\big] &\leqslant R_{\ln}\mathrm{Tr}(M_t H) \\
\sum_i \mathsf{f}_i^t \ln\left(\frac{\lambda_{\mathrm{o}}}{\lambda_i}\right) &\leqslant R_{\ln}\mathrm{Tr}(M_t H)
\end{aligned}$$

From Eq.(10), we get

$$\sum_i \mathsf{f}_i^t \ln\left(\frac{\lambda_{\mathrm{o}}}{\lambda_i}\right) \leqslant 2R_{\ln}\mathsf{f}_t \tag{23}$$

The other term we get from Assumption 4

$$\sum_i m_i^0 \ln\left(\frac{\lambda_{\mathrm{o}}}{\lambda_i}\right) = \mathrm{Tr}\big(M_0 \ln\left(\Lambda_0 H^{-1}\right)\big) = C_{\ln}$$

Substituting the above term, Eq.(23) in Eq.(22) we get

$$\begin{aligned}
\sum_{t=0}^{T-1} 2\frac{\mathsf{f}_t}{T-t} &\leq 2\mathsf{f}_0 + \frac{7C_{\ln}}{2T} + 14\gamma R_{\ln} \sum\sum_{t=0}^{T-1} \frac{\mathsf{f}_t}{T-t} \\
\sum_{t=0}^{T-1} \frac{\mathsf{f}_t}{T-t} &\leq \mathsf{f}_0 + \frac{7C_{\ln}}{4T} + 7\gamma R_{\ln} \sum_{t=0}^{T-1} \frac{\mathsf{f}_t}{T-t}
\end{aligned}$$

And like in the previous proof, $\mathsf{f}_0 = \sum_i \lambda_i m_i^0 \leqslant \lambda_{\max}\sum_i m_i^0 \leqslant \frac{1}{\gamma}\sum_i m_i^0 \ln(1/(\gamma\lambda_i)) \leq C_{\ln}/\gamma$.

$$\sum_{t=0}^{T-1} \frac{\mathsf{f}_t}{T-t} \leqslant \frac{3C_{\ln}}{\gamma T} + 7\gamma R_{\ln} \sum_{k=0}^{T-1} \frac{\mathsf{f}_k}{T-k}.$$

Hence, for $\gamma = (14R_{\ln})^{-1}$, we have

$$\sum_{t=0}^{T-1} \frac{\mathsf{f}_t}{T-t} \leqslant \frac{6C_{\ln}}{\gamma T},$$

and, hence, we conclude like for the previous theorems. We know that Assumption 3 is stricter than Assumption 1 i.e there exists a constant $R^{'} = \left(\ln\left(\frac{\lambda_{\mathrm{o}}}{\lambda_{max}}\right)\right)^{-1} R_{\ln}$ such that Assumption 1 holds with this. Indeed, this allows us to use Lemma 5, for all $T \geqslant 1$,

$$\begin{aligned}
\mathsf{f}_t &\leqslant \frac{\mathrm{Tr}(M_0)}{4\gamma T} + \gamma R^{'} \sum_{k=0}^{T-1} \frac{\mathsf{f}_k}{T-k} \\
&\leqslant \frac{14R_{\ln}\mathrm{Tr}(M_0)}{4T} + \frac{6R^{'}C_{\ln}}{T} \\
&\leqslant \frac{4R_{\ln}\mathrm{Tr}(M_0)}{T} + \frac{6R^{'}C_{\ln}}{T}.
\end{aligned}$$

We can always choose λ_{o} large enough such that $\left(\ln\left(\frac{\lambda_{\mathrm{o}}}{\lambda_{max}}\right)\right) > 1$. With this we note that $\mathrm{Tr}(M_0) \leqslant C_{\ln}$ and $\mathrm{Tr}(H) \leqslant R_{\ln}$. Hence,

$$\mathsf{f}_t \leqslant \frac{10R_{\ln}C_{\ln}}{T}$$

That concludes the proof of Theorem 2.

Note that we have $2\lambda_{\mathrm{o}}\gamma \geqslant 1$, as we chose λ_{o} such that $7R_{\ln} \leqslant \lambda_{\mathrm{o}}$.

C.3 Proof of Theorem 3

Once again we proceed with the same technique as above. The aim here is to tighten the estimation for both the first and the second term with the capacity and source assumptions of the problem (Assumptions 5 and 6).

This estimation rests on the inequality stated in the following Lemma.

Lemma 7 *For $x \in (0,1)$ and $t \geqslant 1$, for $r > 0$ we have the following inequality*

$$x^r(1-x)^t \leq \frac{r^r}{t^r} \tag{24}$$

Proof Let $x \in (0,1)$, $t \geqslant 1$ and $r > 0$. It is standard to note that $(1-x)^t \leqslant e^{-tx}$. Hence,

$$x^r(1-x)^t \leqslant x^r e^{-tx}.$$

Now, a rapid look at the maximum of the function $x \to x^r e^{-tx}$ gives us that it attains its maximum for $x = r/t$. Hence

$$x^r(1-x)^t \leqslant x^r e^{-tx} \leqslant \left(\frac{r}{t}\right)^r e^{-r} \leqslant \frac{r^r}{t^r},$$

and this proves the Lemma.

Again, recall that for all $t \geqslant 1$, from Eq.(12) and $\gamma \leq (4\lambda_{max})^{-1}$

$$\begin{aligned}
m_i^t &= (1-2\gamma\lambda_i)^t\, m_i^0 + \gamma^2 \sum_{k=0}^{t-1} (1-2\gamma\lambda_i)^{t-k-1}\, \mathsf{f}_i^k \\
m_i^t &= (1-2\gamma\lambda_i)^t\, m_i^0 + 2\gamma^2 \sum_{k=0}^{t-1} (1-2\gamma\lambda_i)^{t-k}\, \mathsf{f}_i^k \\
\lambda_i m_i^t &= \lambda_i^{1+\beta}\, (1-2\gamma\lambda_i)^t\, \lambda_i^{-\beta} m_i^0 + 2\gamma^2 \lambda_i^{-\alpha} \sum_{k=0}^{t-1} \lambda^{1+\alpha}\, (1-2\gamma\lambda_i)^{t-k}\, \mathsf{f}_i^k
\end{aligned}$$

Thanks to Lemma (7), we can bound the above expression as,

$$m_i^t = (1-2\gamma\lambda_i)^t\, m_i^0 + 2\gamma^2 \sum_{k=0}^{t-1} (1-2\gamma\lambda_i)^{t-k}\, \mathsf{f}_i^k$$
$$\lambda_i m_i^t = \left(\frac{1+\beta}{\gamma t}\right)^{1+\beta} \lambda_i^{-\beta} m_i^0 + 2\gamma^2 \sum_{k=0}^{t-1} \left(\frac{1+\alpha}{\gamma t}\right)^{1+\alpha} \lambda_i^{-\alpha} \mathsf{f}_i^k$$

Summing across i we get

$$\sum_i \lambda_i m_i^t = \left(\frac{1+\beta}{\gamma t}\right)^{1+\beta} \sum_i \lambda_i^{-\beta} m_i^0 + 2\gamma^{1-\alpha}\,(1+\alpha)^{1+\alpha} \sum_{k=0}^{t-1} \frac{1}{t^{1+\alpha}} \sum_i \lambda_i^{-\alpha} \mathsf{f}_i^k$$

Note from Eq.(10), we have $\sum_i \lambda_i m_i^t = \mathsf{f}_t$. Lets calculate the remaining terms

$$\begin{aligned}
\sum_i \mathsf{f}_i^t\, {\lambda_i}^{-\alpha} &= \sum_i {\lambda_i}^{-\alpha} \mathbb{E}\left[\left(\langle v_i, x_t\rangle^2\right) x_t^\top M_t x_t\right] \\
&= \mathbb{E}\left[\left(\sum_i {\lambda_i}^{-\alpha} \langle v_i, x_t\rangle^2\right) x_t^\top M_t x_t\right] \\
&= \mathbb{E}\left[\left\langle x_t, H^{1-\alpha} x_t\right\rangle\ \mathrm{Tr}\big(x_t x_t^\top M_t\big)\right] \\
&= \mathrm{Tr}\big[\mathbb{E}\left[\left\langle x_t, H^{1-\alpha} x_t\right\rangle\ x_t x_t^\top\right] M_t\big]
\end{aligned}$$

Note for any two PSD matrices A, B if $A \preccurlyeq B$ then $\mathrm{Tr}(AM) \leq \mathrm{Tr}(BM)$, for any PSD matrix M. From Assumption 3, we have

$$\begin{aligned}
\mathbb{E}\left[\left\langle x_t, H^{1-\alpha} x_t\right\rangle\ x_t x_t^\top\right] &\preccurlyeq R_\alpha H \\
\mathrm{Tr}\big[\mathbb{E}\left[\left\langle x_t, H^{1-\alpha} x_t\right\rangle\ x_t x_t^\top\right] M_t\big] &\leqslant R_\alpha \mathrm{Tr}(M_t H) \\
\sum_i \mathsf{f}_i^t {\lambda_i}^{-\alpha} &\leqslant R_\alpha \mathrm{Tr}(M_t H)
\end{aligned}$$

From Eq.(10), we get

$$\sum_i \mathsf{f}_i^t {\lambda_i}^{-\alpha} \leqslant 2R_\alpha \mathsf{f}_t \tag{25}$$

The other term we get from Assumption 4

$$\sum_i m_i^0 {\lambda_i}^{-\alpha} = \mathrm{Tr}\big(M_0 H^{-\beta}\big) = C_\beta$$

Substituting the above computed terms we get the above

$$\mathsf{f}_t \leqslant \frac{C_\beta}{2}\left(\frac{1+\beta}{\gamma t}\right)^{1+\beta} + 2\,(1+\alpha)^{1+\alpha}\,\gamma^{1-\alpha} R_\alpha \sum_{k=0}^{t-1} \frac{\mathsf{f}_k}{(t-k)^{1+\alpha}}. \tag{26}$$

And from this, we use the same technique as for the previous theorems to bound the second term of the right side of the inequality. To accomplish this we need a bound on the following sum. For $\beta > -1$, $\alpha \in (0,1)$, $T \geqslant 2$,

$$\mathsf{S}_T(\alpha,\beta) := \sum_{t=1}^{T-1} \frac{1}{t^{1+\beta}(T-t)^{1+\alpha}}$$

This is the aim of the Lemma below.

Lemma 8 *For $\beta > -1$, $\alpha \in (0,1)$, $T \geqslant 2$, we have the following upper-bound,*

$$\mathsf{S}_T(\alpha,\beta) \leqslant \frac{2^{2+\alpha\wedge\beta}\xi_{\alpha\vee\beta}}{T^{1+\alpha\wedge\beta}},$$

where, for $u > 0$, $\xi u := \sum_{k\geqslant 1}\frac{1}{k^{1+u}}$ and we use the following classical notations: $\alpha\wedge\beta = \min(\alpha,\beta)$ and $\alpha\vee\beta = \max(\alpha,\beta) > 0$.

Proof Let us assume that $\beta \leqslant \alpha$, we have,

$$\begin{aligned}
\mathsf{S}_T(\alpha,\beta) = \sum_{t=1}^{T-1}\frac{1}{t^{1+\beta}(T-t)^{1+\alpha}} &= \sum_{t=1}^{T-1}\frac{1}{(t(T-t))^{1+\beta}}\frac{1}{(T-t)^{\alpha-\beta}} \\
&= \frac{1}{T^{1+\beta}}\left[\sum_{t=1}^{T-1}\frac{1}{(T-t)^{\alpha-\beta}}\left(\frac{1}{t}+\frac{1}{T-t}\right)^{1+\beta}\right] \\
&= \frac{2^{1+\beta}}{T^{1+\beta}}\left[\sum_{t=1}^{T-1}\frac{1}{(T-t)^{\alpha-\beta}t^{1+\beta}}+\sum_{t=1}^{T-1}\frac{1}{(T-t)^{\alpha+1}}\right].
\end{aligned}$$

Now the second term is trivially upper bounded by ξ_α. And for the first term, we use Young's inequality with coefficient $(p,q) = (\frac{1+\alpha}{\alpha-\beta}, \frac{1+\alpha}{1+\beta})$:

$$\begin{aligned}
\sum_{t=1}^{T-1}\frac{1}{(T-t)^{\alpha-\beta}t^{1+\beta}} &\leqslant \frac{1}{p}\sum_{t=1}^{T-1}\frac{1}{(T-t)^{(\alpha-\beta)p}}+\frac{1}{q}\sum_{t=1}^{T-1}\frac{1}{t^{(1+\beta)q}} \\
&= \frac{1}{p}\sum_{t=1}^{T-1}\frac{1}{t^{1+\alpha}}+\frac{1}{q}\sum_{t=1}^{T-1}\frac{1}{t^{1+\alpha}} \\
&= \sum_{t=1}^{T-1}\frac{1}{t^{1+\alpha}} \leqslant \xi_\alpha.
\end{aligned}$$

This concludes the proof is the case where $\beta \leqslant \alpha$.

Symmetrically, if $\alpha \leqslant \beta$, by a change of variable $t \to T-t$,

$$\mathsf{S}_T(\alpha,\beta) = \sum_{t=1}^{T-1}\frac{1}{t^{1+\beta}(T-t)^{1+\alpha}} = \sum_{t=1}^{T-1}\frac{1}{t^{1+\alpha}(T-t)^{1+\beta}} = \mathsf{S}_T(\beta,\alpha),$$

and the proof follows.

Thanks to the Lemma we continue the proof of Theorem 3. Recall Eq. (26): for $T \geqslant 2$,

$$\mathsf{f}_t \leqslant \frac{C_\beta}{2}\left(\frac{1+\beta}{\gamma t}\right)^{1+\beta} + 2\,(1+\alpha)^{1+\alpha}\gamma^{1-\alpha}R_\alpha\sum_{k=0}^{t-1}\frac{\mathsf{f}_k}{(t-k)^{1+\alpha}}.$$

Now, we proceed like previously and try to second term in the above inequality,

$$\begin{aligned}
\sum_{t=0}^{T-1}\frac{\mathsf{f}_t}{(T-t)^{1+\alpha}} &\leqslant \frac{\mathsf{f}_0}{T^{1+\alpha}} + \frac{C_\beta}{2}\left(\frac{1+\beta}{\gamma}\right)^{1+\beta}\sum_{t=1}^{T-1}\frac{1}{t^{1+\beta}(T-t)^{1+\alpha}} \\
&\qquad + 2(1+\alpha)^{1+\alpha}\kappa\gamma^{1-\alpha}R_\alpha\sum_{t=1}^{T-1}\sum_{k=0}^{t-1}\frac{\mathsf{f}_k}{(t-k)^{1+\alpha}(T-t)^{1+\alpha}} \\
&\leqslant \frac{\mathsf{f}_0}{T^{1+\alpha}} + \frac{C_\beta}{2}\left(\frac{1+\beta}{\gamma}\right)^{1+\beta} S_T(\alpha,\beta) \\
&\qquad + 2(1+\alpha)^{1+\alpha}\kappa\gamma^{1-\alpha}R_\alpha\sum_{k=1}^{T-1}\mathsf{f}_k\sum_{t=k+1}^{T-1}\frac{1}{(t-k)^{1+\alpha}(T-t)^{1+\alpha}} \\
&\leqslant \frac{\mathsf{f}_0}{T^{1+\alpha}} + \frac{C_\beta}{2}\left(\frac{1+\beta}{\gamma}\right)^{1+\beta} S_T(\alpha,\beta) + 2(1+\alpha)^{1+\alpha}\kappa\gamma^{1-\alpha}R_\alpha\sum_{k=1}^{T-1}\mathsf{f}_k S_{T-k}(\alpha,\alpha).
\end{aligned}$$

And applying Lemma 8, and the fact that $\frac{\mathsf{f}_0}{T^{1+\alpha}} \leqslant \frac{C_\beta}{\gamma^\beta T^{1+\alpha\wedge\beta}}$,

$$\begin{aligned}
\sum_{t=0}^{T-1} \frac{\mathsf{f}_t}{(T-t)^{1+\alpha}} &\leqslant \frac{\mathsf{f}_0}{T^{1+\alpha}} + \frac{C_\beta}{2}\left(\frac{1+\beta}{\gamma}\right)^{1+\beta} \frac{2^{2+\alpha\wedge\beta}\xi_{\alpha\vee\beta}}{T^{1+\alpha\wedge\beta}} \\
&\qquad + 2(1+\alpha)^{1+\alpha}\kappa\gamma^{1-\alpha}R_\alpha 2^{2+\alpha}\xi_\alpha \sum_{k=1}^{T-1} \frac{\mathsf{f}_k}{(T-t)^{1+\alpha}} \\
&\leqslant 2^{2+\alpha\wedge\beta} C_\beta \left(\frac{1+\beta}{\gamma}\right)^{1+\beta} \frac{\xi_{\alpha\vee\beta}}{T^{1+\alpha\wedge\beta}} \\
&\qquad + 2^{3+\alpha}(1+\alpha)^{1+\alpha}\kappa\xi_\alpha\gamma^{1-\alpha}R_\alpha \sum_{k=1}^{T-1} \frac{\mathsf{f}_k}{(T-t)^{1+\alpha}}.
\end{aligned}$$

Now, for γ such that $2^{3+\alpha}(1+\alpha)^{1+\alpha}\xi_\alpha\gamma^{1-\alpha}R_\alpha \leqslant 1/2$, i.e., for simplicity,

$$\gamma^{1-\alpha} \leqslant (128\kappa\xi_\alpha R_\alpha)^{-1},$$

then, we have,

$$\sum_{t=0}^{T-1} \frac{\mathsf{f}_t}{(T-t)^{1+\alpha}} \leqslant 8C_\beta \left(\frac{1+\beta}{\gamma}\right)^{1+\beta} \frac{\xi_{\alpha\vee\beta}}{T^{1+\alpha\wedge\beta}},$$

and, hence, we conclude like for the previous theorems. Indeed, for all $T \geqslant 1$, recalling Eq. (26): for $T \geqslant 2$,

$$\begin{aligned}
\mathsf{f}_T &\leqslant \frac{C_\beta}{2}\left(\frac{1+\beta}{\gamma T}\right)^{1+\beta} + 2\kappa\,(1+\alpha)^{1+\alpha}\,\gamma^{1-\alpha}R_\alpha \sum_{k=0}^{T-1} \frac{\mathsf{f}_t}{(T-t)^{1+\alpha}} \\
&\leqslant \frac{C_\beta}{2}\left(\frac{1+\beta}{\gamma T}\right)^{1+\beta} + \frac{1}{4\xi_\alpha}\sum_{t=0}^{T-1} \frac{\mathsf{f}_t}{(T-t)^{1+\alpha}} \\
&\leqslant \frac{C_\beta}{2}\left(\frac{1+\beta}{\gamma T}\right)^{1+\beta} + \frac{2}{\xi_\alpha}C_\beta\left(\frac{1+\beta}{\gamma}\right)^{1+\beta} \frac{\xi_{\alpha\vee\beta}}{T^{1+\alpha\wedge\beta}} \\
&\leqslant \frac{C_\beta}{2}\left(\frac{1+\beta}{\gamma T}\right)^{1+\beta} + 2C_\beta\left(\frac{1+\beta}{\gamma}\right)^{1+\beta} \frac{1}{T^{1+\alpha\wedge\beta}} \\
&\leqslant 3C_\beta\left(\frac{1+\beta}{\gamma}\right)^{1+\beta} \frac{1}{T^{1+\alpha\wedge\beta}}.
\end{aligned}$$

That concludes the proof of Theorem 3.