A Markov Chain Theory Approach to Characterizing the Minimax Optimality of Stochastic Gradient Descent (for Least Squares)

Prateek Jain[1], Sham M. Kakade*[2], Rahul Kidambi[3], Praneeth Netrapalli[4], Venkata Krishna Pillutla[5], and Aaron Sidford[6]

1 **Microsoft Research, Bangalore, India**
praneeth@microsoft.com
2 **University of Washington, Seattle, WA, USA**
sham@cs.washington.edu
3 **University of Washington, Seattle, WA, USA**
rkidambi@uw.edu
4 **Microsoft Research, Bangalore, India**
praneeth@microsoft.com
5 **University of Washington, Seattle, WA, USA**
pillutla@cs.washington.edu
6 **Stanford University, Palo Alto, CA, USA**
sidford@stanford.edu

Abstract

This work provides a simplified proof of the statistical minimax optimality of (iterate averaged) stochastic gradient descent (SGD), for the special case of least squares. This result is obtained by analyzing SGD as a stochastic process and by sharply characterizing the stationary covariance matrix of this process. The finite rate optimality characterization captures the constant factors and addresses model mis-specification.

1998 ACM Subject Classification G.1.6 Optimization

Keywords and phrases Stochastic Gradient Descent, Minimax Optimality, Least Squares Regression

Digital Object Identifier 10.4230/LIPIcs.FSTTCS.2017.2

Category Invited Paper

1 Introduction

Stochastic gradient descent is among the most commonly used practical algorithms for large scale stochastic optimization. The seminal result of [9, 8] formalized this effectiveness, showing that for certain (locally quadric) problems, asymptotically, stochastic gradient descent is statistically minimax optimal (provided the iterates are averaged). There are a number of more modern proofs [1, 3, 2, 5] of this fact, which provide finite rates of

* Sham Kakade acknowledges funding from the Washington Research Foundation Fund for Innovation in Data-Intensive Discovery.

37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2017).
Editors: Satya Lokam and R. Ramanujam; Article No. 2; pp. 2:1–2:10
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

convergence. Other recent algorithms also achieve the statistically optimal minimax rate, with finite convergence rates [4].

This work provides a short proof of this minimax optimality for SGD for the special case of least squares through a characterization of SGD as a stochastic process. The proof builds on ideas developed in [2, 5].

SGD for least squares. The expected square loss for $w \in \mathbb{R}^d$ over input-output pairs (x, y), where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ are sampled from a distribution $\mathcal{D}$, is:

$$L(w) = \frac{1}{2}\,\mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - w \cdot x)^2]$$

The optimal weight is denoted by:

$$w^* := \operatorname*{argmin}_w L(w)\,.$$

Assume the argmin in unique.

Stochastic gradient descent proceeds as follows: at each iteration t, using an i.i.d. sample $(x_t, y_t) \sim \mathcal{D}$, the update of w_t is:

$$w_t = w_{t-1} + \gamma(y_t - w_{t-1} \cdot x_t)x_t$$

where γ is a fixed stepsize.

Notation. For a symmetric positive definite matrix A and a vector x, define:

$$\|x\|_A^2 := x^\top Ax.$$

For a symmetric matrix M, define the induced matrix norm under A as:

$$\|M\|_A := \max_{\|v\|=1} \frac{v^\top Mv}{v^\top Av} = \|A^{-1/2}MA^{-1/2}\|\,.$$

The statistically optimal rate. Using n samples (and for large enough n), the minimax optimal rate is achieved by the maximum likelihood estimator (the MLE), or, equivalently, the empirical risk minimizer. Given n i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$, define

$$\widehat{w}_n^{\mathrm{MLE}} := \arg\min_w \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\,(y_i - w \cdot x_i)^2$$

where $\widehat{w}_n^{\mathrm{MLE}}$ denotes the MLE estimator over the n samples.

This rate can be characterized as follows: define

$$\sigma^2_{\mathrm{MLE}} := \frac{1}{2}\mathbb{E}\left[(y - w^*x)^2\|x\|^2_{H^{-1}}\right]\,,$$

and the (asymptotic) rate of the MLE is $\sigma^2_{\mathrm{MLE}}/n$ [7, 10]. Precisely,

$$\lim_{n\to\infty} \frac{\mathbb{E}[L(\widehat{w}_n^{\mathrm{MLE}})] - L(w^*)}{\sigma^2_{\mathrm{MLE}}/n} = 1,$$

The works of [9, 8] proved that a certain averaged stochastic gradient method achieves this minimax rate, in the limit.

For the case of additive noise models (i.e. the "well-specified" case), the assumption is that $y = w^* \cdot x + \eta$, with η being independent of x). Here, it is straightforward to see that:

$$\frac{\sigma^2_{\text{MLE}}}{n} = \frac{1}{2}\frac{d\sigma^2}{n}.$$

The rate of σ^2_{MLE}/n is still minimax optimal even among mis-specified models, where the additive noise assumption may not hold [6, 7, 10].

Assumptions. Assume the fourth moment of x is finite. Denote the second moment matrix of x as

$$H := \mathbb{E}[xx^\top],$$

and suppose H is strictly positive definite with minimal eigenvalue:

$$\mu := \sigma_{\min}(H).$$

Define R^2 as the smallest value which satisfies:

$$\mathbb{E}[\|x\|^2 xx^\top] \preceq R^2 \mathbb{E}[xx^\top].$$

This implies $\text{Tr}(H) = \mathbb{E}\|x\|^2 \leq R^2$.

2 Statistical Risk Bounds

Define:

$$\Sigma := \mathbb{E}[(y - w^* x)^2 xx^\top],$$

and so the optimal constant in the rate can be written as:

$$\sigma^2_{\text{MLE}} = \frac{1}{2}\text{Tr}(H^{-1}\Sigma) = \frac{1}{2}\mathbb{E}\left[(y - w^* x)^2 \|x\|^2_{H^{-1}}\right],$$

For the mis-specified case, it is helpful to define:

$$\rho_{\text{misspec}} := \frac{d\|\Sigma\|_H}{\text{Tr}(H^{-1}\Sigma)},$$

which can be viewed as a measure of how mis-specified the model is. Note if the model is well-specified, then $\rho_{\text{misspec}} = 1$.

Denote the average iterate, averaged from iteration t to T, by:

$$\overline{w}_{t:T} := \frac{1}{T-t}\sum_{t'=t}^{T-1} w_{t'}.$$

▶ **Theorem 1.** *Suppose $\gamma < \frac{1}{R^2}$. The risk is bounded as:*

$$\mathbb{E}[L(\overline{w}_{t:T})] - L(w^*)$$
$$\leq \left(\sqrt{\frac{1}{2}\exp\left(-\gamma\mu t\right) R^2 \|w_0 - w^*\|^2} + \sqrt{\left(1 + \frac{\gamma R^2}{1-\gamma R^2}\rho_{\text{misspec}}\right)\frac{\sigma^2_{\text{MLE}}}{T-t}}\right)^2.$$

The bias term (the first term) decays at a geometric rate (one can set $t = T/2$ or maintain multiple running averages if T is not known in advance). If $\gamma = 1/(2R^2)$ and the model is well-specified ($\rho_{\text{misspec}} = 1$), then the variance term is $2\sigma_{\text{MLE}}/\sqrt{T-t}$, and the rate of the bias contraction is μ/R^2. If the model is not well specified, then using a smaller stepsize of $\gamma = 1/(2\rho_{\text{misspec}}R^2)$, leads to the same minimax optimal rate (up to a constant factor of 2), albeit at a slower bias contraction rate. In the mis-specified case, an example in [5] shows that such a smaller stepsize is required in order to be within a constant factor of the minimax rate. An even smaller stepsize leads to a constant even closer to that of the optimal rate.

3 Analysis

The analysis first characterizes a bias/variance decomposition, where the variance is bounded in terms of properties of the stationary covariance of w_t. Then this asymptotic covariance matrix is analyzed.

Throughout assume:

$$\gamma < \frac{1}{R^2} .$$

3.1 The Bias-Variance Decomposition

The gradient at w^* in iteration t is:

$$\xi_t := -(y_t - w^* \cdot x_t)x_t ,$$

which is a mean 0 quantity. Also define:

$$B_t := \mathrm{I} - x_t x_t^\top .$$

The update rule can be written as:

$$\begin{aligned} w_t - w^* &= w_{t-1} - w^* + \gamma(y_t - w_{t-1} \cdot x_t)x_t \\ &= (\mathrm{I} - \gamma x_t x_t^\top)(w_{t-1} - w^*) - \gamma\xi_t \\ &= B_t(w_{t-1} - w^*) - \gamma\xi_t . \end{aligned}$$

Roughly speaking, the above shows how the process on $w_t - w^*$ consists of a contraction along with an addition of a zero mean quantity.

From recursion,

$$w_t - w^* = B_t \cdots B_1(w_0 - w^*) - \gamma \left(\xi_t + B_t\xi_{t-1} + \cdots + B_t \cdots B_2\xi_1\right) .$$

This immediately leads to the following lemma.

▶ **Lemma 2.** *The error is bounded as:*

$$\mathbb{E}[L(\overline{w}_{t:T})] - L(w^*) \leq \frac{1}{2}\left(\sqrt{\mathbb{E}[\|\overline{w}_{t:T} - w^*\|_H^2 | \xi_0 = \cdots = \xi_T = 0]} + \sqrt{\mathbb{E}[\|\overline{w}_{t:T} - w^*\|_H^2 | w_0 = w^*]}\right)^2 ,$$

where

$$\begin{aligned} \mathbb{E}[\|\overline{w}_{t:T} - w^*\|_H^2 | \xi_0 = \cdots = \xi_T = 0] &= \mathbb{E}\|B_t \cdots B_1(w_0 - w^*)\|_H^2 , \\ \mathbb{E}[\|\overline{w}_{t:T} - w^*\|_H^2 | w_0 = w^*] &= \gamma^2\mathbb{E}\|\xi_t + B_t\xi_{t-1} + \cdots + B_t \cdots B_2\xi_1\|_H^2 . \end{aligned}$$

The first term can be interpreted as the bias. $\mathbb{E}[\|\overline{w}_{t:T} - w^*\|_H^2 | \xi_0 = \cdots = \xi_T = 0]$ is the risk in a process without additive noise; the conditioning is a little misleading and is meant to denote the error in a process without additive noise. The second term, when squared, gives rise to the variance; it is the error under a process driven solely by noise where $w_0 = w^*$.

Proof. First, for vector valued random variables u and v, the fact that $(\mathbb{E}u^\top H v)^2 \leq \mathbb{E}[\|u\|_H^2]\mathbb{E}[\|v\|_H^2]$ implies

$$\mathbb{E}\|u+v\|_H^2 \leq \left(\sqrt{\mathbb{E}\|u\|_H^2} + \sqrt{\mathbb{E}\|v\|_H^2}\right)^2 .$$

To complete the proof of the lemma, note $\mathbb{E}L(w) - L(w^*) = \frac{1}{2}\mathbb{E}\|w - w^*\|_H^2$. ◀

Bias. The bias term is characterized as follows:

▶ **Lemma 3.** *For all t,*

$$\mathbb{E}[\|\overline{w}_{t:T} - w^*\|_H^2 | \xi_0 = \cdots = \xi_T = 0] \leq \exp(-\gamma\mu t)\|w_0 - w^*\|^2 .$$

Proof. Assume $\xi_t = 0$ for all t. Observe:

$$\begin{aligned} \mathbb{E}\|w_t - w^*\|^2 &= \mathbb{E}\|w_{t-1} - w^*\|^2 - 2\gamma(w_{t-1} - w^*)^\top \mathbb{E}[xx^\top](w_{t-1} - w^*) \\ &\quad + \gamma^2 (w_{t-1} - w^*)^\top \mathbb{E}[\|x\|^2 xx^\top](w_{t-1} - w^*) \\ &\leq \mathbb{E}\|w_{t-1} - w^*\|^2 - 2\gamma(w_{t-1} - w^*)^\top H (w_{t-1} - w^*) \\ &\quad + \gamma^2 R^2 (w_{t-1} - w^*)^\top H (w_{t-1} - w^*) \\ &\leq \mathbb{E}\|w_{t-1} - w^*\|^2 - \gamma \mathbb{E}\|w_{t-1} - w^*\|_H^2 \\ &\leq (1 - \gamma\mu)\mathbb{E}\|w_{t-1} - w^*\|^2 , \end{aligned}$$

which completes the proof. ◀

Variance. Now suppose $w_0 = w^*$. Define the covariance matrix:

$$C_t := \mathbb{E}[(w_t - w^*)(w_t - w^*)^\top | w_0 = w^*]$$

Using the recursion, $w_t - w^* = B_t(w_{t-1} - w^*) + \gamma\xi_t$,

$$C_{t+1} = C_t - \gamma H C_t - \gamma C_t H + \gamma^2 \mathbb{E}[(x^\top C_t x) xx^\top] + \gamma^2 \Sigma \tag{1}$$

which follows from:

$$\mathbb{E}[(w_t - w^*)\xi_{t+1}^\top] = 0 , \text{ and } \mathbb{E}[(x_{t+1}x_{t+1}^\top)(w_t - w^*)\xi_{t+1}^\top] = 0$$

(these hold since $w_t - w^*$ is mean 0 and both x_{t+1} and ξ_{t+1} are independent of $w_t - w^*$).

▶ **Lemma 4.** *Suppose $w_0 = w^*$. There exists a unique C_∞ such that:*

$$0 = C_0 \preceq C_1 \preceq \cdots \preceq C_\infty$$

where C_∞ satisfies:

$$C_\infty = C_\infty - \gamma H C_\infty - \gamma C_\infty H + \gamma^2 \mathbb{E}[(x^\top C_\infty x) xx^\top] + \gamma^2 \Sigma . \tag{2}$$

Proof. By recursion,

$$\begin{aligned} w_t - w^* &= B_t(w_{t-1} - w^*) + \gamma\xi_t \\ &= \gamma\left(\xi_t + B_t\xi_{t-1} + \cdots + B_t \cdots B_2\xi_1\right) . \end{aligned}$$

Using that ξ_t is mean zero and independent of $B_{t'}$ and $\xi_{t'}$ for $t < t'$,

$$C_t = \gamma^2\left(\mathbb{E}[\xi_t\xi_t^\top] + \mathbb{E}[B_t\xi_{t-1}\xi_{t-1}^\top B_t] + \cdots + \mathbb{E}[B_t \cdots B_2\xi_1\xi_1^\top B_2^\top \cdots B_t^\top]\right)$$

Now using that $\mathbb{E}[\xi_1\xi_1^\top] = \Sigma$ and that ξ_t and $B_{t'}$ are independent (for $t \neq t'$),

$$\begin{aligned} C_t &= \gamma^2\left(\Sigma + \mathbb{E}[B_2\Sigma B_2] + \cdots + \mathbb{E}[B_t \cdots B_2\Sigma B_2^\top \cdots B_t^\top]\right) \\ &= C_{t-1} + \gamma^2\mathbb{E}[B_t \cdots B_2\Sigma B_2^\top \cdots B_t^\top] \end{aligned}$$

which proves $C_{t-1} \preceq C_t$.

To prove the limit exists, it suffices to first argue the trace of C_t is uniformly bounded from above, for all t. By taking the trace of update rule, Equation 1, for C_t,

$$\mathrm{Tr}(C_{t+1}) = \mathrm{Tr}(C_t) - 2\gamma\mathrm{Tr}(HC_t) + \gamma^2\mathrm{Tr}(\mathbb{E}[(x^\top C_t x)xx^\top]) + \gamma^2\mathrm{Tr}(\Sigma) .$$

Observe:

$$\mathrm{Tr}(\mathbb{E}[(x^\top C_t x)xx^\top]) = \mathrm{Tr}(\mathbb{E}[(x^\top C_t x)\|x\|^2]) = \mathrm{Tr}(C_t\mathbb{E}[\|x\|^2 xx^\top]) \leq R^2\mathrm{Tr}(C_t H) \tag{3}$$

and, using $\gamma \leq 1/R^2$,

$$\mathrm{Tr}(C_{t+1}) \leq \mathrm{Tr}(C_t) - \gamma\mathrm{Tr}(HC_t) + \gamma^2\mathrm{Tr}(\Sigma) \leq (1-\gamma\mu)\mathrm{Tr}(C_t) + \gamma^2\mathrm{Tr}(\Sigma) \leq \frac{\gamma\mathrm{Tr}(\Sigma)}{\mu} .$$

proving the uniform boundedness of the trace of C_t. Now, for any fixed v, the limit of $v^\top C_t v$ exists, by the monotone convergence theorem. From this, it follows that every entry of the matrix C_t converges. ◀

▶ **Lemma 5.** *Define:*

$$\overline{w}_T := \frac{1}{T}\sum_{t=0}^{T-1} w_t .$$

and so:

$$\frac{1}{2}\mathbb{E}[\|\overline{w}_T - w^*\|_H^2 | w_0 = w^*] \leq \frac{\mathrm{Tr}(C_\infty)}{\gamma T}$$

Proof. Note

$$\begin{aligned} &\mathbb{E}[(\overline{w}_T - w^*)(\overline{w}_T - w^*)^\top | w_0 = w^*] \\ &\quad = \frac{1}{T^2}\sum_{t=0}^{T-1}\sum_{t'=0}^{T-1}\mathbb{E}[(w_t - w^*)(w_{t'} - w^*)^\top | w_0 = w^*] \\ &\quad \preceq \frac{1}{T^2}\sum_{t=0}^{T-1}\sum_{t'=t}^{T-1}\Big(\mathbb{E}[(w_t - w^*)(w_{t'} - w^*)^\top | w_0 = w^*] + \\ &\qquad\qquad \mathbb{E}[(w_{t'} - w^*)(w_t - w^*)^\top | w_0 = w^*]\Big) , \end{aligned}$$

double counting the diagonal terms $\mathbb{E}[(w_t - w^*)(w_t - w^*)^\top | w_0 = w^*] \succeq 0$. For $t \le t'$, $\mathbb{E}[(w_{t'} - w^*)|w_0 = w^*] = (\mathrm{I} - \gamma H)^{t'-t}\mathbb{E}[(w_t - w^*)|w_0 = w^*]$. To see why, consider the recursion $w_t - w^* = (\mathrm{I} - \gamma x_t x_t^\top)(w_{t-1} - w^*) - \gamma\xi_t$ and take expectations to get $\mathbb{E}[w_t - w^*|w_0 = w^*] = (\mathrm{I} - \gamma H)\mathbb{E}[w_{t-1} - w^*|w_0 = w^*]$ since the sample x_t is is independent of the w_{t-1}. From this,

$$\mathbb{E}[(\overline{w}_T - w^*)(\overline{w}_T - w^*)^\top | w_0 = w^*] \preceq \frac{1}{T^2}\sum_{t=0}^{T-1}\sum_{\tau=0}^{T-t-1}(I - \gamma H)^\tau C_t + C_t(\mathrm{I} - \gamma H)^\tau ,$$

and so,

$$\begin{aligned}\mathbb{E}[\|\overline{w}_T - w^*\|_H^2 | w_0 = w^*] &= \mathrm{Tr}\big(H\mathbb{E}[(\overline{w}_T - w^*)(\overline{w}_T - w^*)^\top | w_0 = w^*]\big) \\ &\le \frac{1}{T^2}\sum_{t=0}^{T-1}\sum_{\tau=0}^{T-t-1}\mathrm{Tr}\big(H(\mathrm{I} - \gamma H)^\tau C_t\big) + \mathrm{Tr}\big(C_t(\mathrm{I} - \gamma H)^\tau H\big) .\end{aligned}$$

Notice that $H(\mathrm{I} - \gamma H)^\tau = (\mathrm{I} - \gamma H)^\tau H$ for any non-negative integer τ. Since $H \succ 0$ and $I - \gamma H \succeq 0$, $H(\mathrm{I} - \gamma H)^\tau \succeq 0$ because the product of two commuting PSD matrices is PSD. Also note that for PSD matrices A, B, $\mathrm{Tr}AB \ge 0$. Hence,

$$\begin{aligned}\mathbb{E}[\|\overline{w}_T - w^*\|_H^2 | w_0 = w^*] &\le \frac{2}{T^2}\sum_{t=0}^{T-1}\sum_{\tau=0}^{\infty}\mathrm{Tr}\big(H(\mathrm{I} - \gamma H)^\tau C_t\big) \\ &= \frac{2}{T^2}\sum_{t=0}^{T-1}\mathrm{Tr}\big(H(\sum_{\tau=0}^{\infty}(\mathrm{I} - \gamma H)^\tau)C_t\big) \\ &= \frac{2}{T^2}\sum_{t=0}^{T-1}\mathrm{Tr}\big(H(\gamma H)^{-1}C_t\big) \qquad (*) \\ &= \frac{2}{\gamma T^2}\sum_{t=0}^{T-1}\mathrm{Tr}(C_t) \\ &\le \frac{2}{\gamma T}\cdot\mathrm{Tr}(C_\infty) ,\end{aligned}$$

from lemma 4 where $(*)$ followed from

$$(\gamma H)^{-1} = (\mathrm{I} - (\mathrm{I} - \gamma H))^{-1} = \sum_{\tau=0}^{\infty}(\mathrm{I} - \gamma H)^\tau ,$$

and the series converges because $\mathrm{I} - \gamma H \prec \mathrm{I}$. ◀

3.2 Stationary Distribution Analysis

Define two linear operators on symmetric matrices, $\mathcal{S}$ and $\mathcal{T}$ — where $\mathcal{S}$ and $\mathcal{T}$ can be viewed as matrices acting on $\binom{d+1}{2}$ dimensions — as follows:

$$\mathcal{S} \circ M := \mathbb{E}[(x^\top M x)xx^\top] , \qquad \mathcal{T} \circ M := HM + MH .$$

With this, C_∞ is the solution to:

$$\mathcal{T} \circ C_\infty = \gamma\mathcal{S} \circ C_\infty + \gamma\Sigma \qquad (4)$$

(due to Equation 2).

▶ **Lemma 6.** *(Crude C_∞ bound) C_∞ is bounded as:*

$$C_\infty \preceq \frac{\gamma\|\Sigma\|_H}{1-\gamma R^2}\,\mathrm{I}\,.$$

Proof. Define one more linear operator as follows:

$$\widetilde{\mathcal{T}} \circ M := \mathcal{T} \circ M - \gamma HMH = HM + MH - \gamma HMH\,.$$

The inverse of this operator can be written as:

$$\widetilde{\mathcal{T}}^{-1} \circ M = \gamma \sum_{t=0}^{\infty} (\mathrm{I} - \gamma \widetilde{\mathcal{T}})^t \circ M = \gamma \sum_{t=0}^{\infty} (\mathrm{I} - \gamma H)^t M (\mathrm{I} - \gamma H)^t\,.$$

which exists since the sum converges due to that $0 \preceq \mathrm{I} - \gamma H \preceq \mathrm{I}$.

A few inequalities are helpful: If $0 \preceq M \preceq M'$, then

$$0 \preceq \widetilde{\mathcal{T}}^{-1} \circ M \preceq \widetilde{\mathcal{T}}^{-1} \circ M'\,, \tag{5}$$

since

$$\widetilde{\mathcal{T}}^{-1} \circ M = \gamma \sum_{t=0}^{\infty} (\mathrm{I} - \gamma H)^t M (\mathrm{I} - \gamma H)^t \preceq \gamma \sum_{t=0}^{\infty} (\mathrm{I} - \gamma H)^t M' (\mathrm{I} - \gamma H)^t = \widetilde{\mathcal{T}}^{-1} \circ M'\,,$$

(which follows since $0 \preceq \mathrm{I} - \gamma H$). Also, if $0 \preceq M \preceq M'$, then

$$0 \preceq \mathcal{S} \circ M \preceq \mathcal{S} \circ M'\,, \tag{6}$$

which implies:

$$0 \preceq \widetilde{\mathcal{T}}^{-1} \circ \mathcal{S} \circ M \preceq \widetilde{\mathcal{T}}^{-1} \circ \mathcal{S} \circ M'\,. \tag{7}$$

The following inequality is also of use:

$$\Sigma \preceq \|H^{-1/2} \Sigma H^{-1/2}\| H = \|\Sigma\|_H H\,.$$

By definition of $\widetilde{\mathcal{T}}$,

$$\widetilde{\mathcal{T}} \circ C_\infty = \gamma \mathcal{S} \circ C_\infty + \gamma \Sigma - \gamma H C_\infty H\,.$$

Using this and Equation 5,

$$\begin{aligned} C_\infty &= \gamma \widetilde{\mathcal{T}}^{-1} \circ \mathcal{S} \circ C_\infty + \gamma \widetilde{\mathcal{T}}^{-1} \circ \Sigma - \gamma \widetilde{\mathcal{T}}^{-1} \circ (H C_\infty H) \\ &\preceq \gamma \widetilde{\mathcal{T}}^{-1} \circ \mathcal{S} \circ C_\infty + \gamma \widetilde{\mathcal{T}}^{-1} \circ \Sigma \\ &\preceq \gamma \widetilde{\mathcal{T}}^{-1} \circ \mathcal{S} \circ C_\infty + \gamma \|\Sigma\|_H \widetilde{\mathcal{T}}^{-1} \circ H\,. \end{aligned}$$

Proceeding recursively by using Equation 7,

$$\begin{aligned} C_\infty &\preceq (\gamma \widetilde{\mathcal{T}}^{-1} \circ \mathcal{S})^2 \circ C_\infty + \gamma \|\Sigma\|_H (\gamma \widetilde{\mathcal{T}}^{-1} \circ \mathcal{S}) \circ \widetilde{\mathcal{T}}^{-1} \circ H + \gamma \|\Sigma\|_H \widetilde{\mathcal{T}}^{-1} \circ H \\ &\preceq \gamma \|\Sigma\|_H \sum_{t=0}^{\infty} (\gamma \widetilde{\mathcal{T}}^{-1} \circ \mathcal{S})^t \circ \widetilde{\mathcal{T}}^{-1} \circ H\,. \end{aligned}$$

Using

$$\mathcal{S} \circ \mathrm{I} \preceq R^2 H$$

and

$$\widetilde{\mathcal{T}}^{-1} \circ H$$
$$= \gamma \sum_{t=0}^{\infty} (\mathrm{I} - \gamma H)^{2t} H = \gamma \sum_{t=0}^{\infty} (\mathrm{I} - \gamma 2H + \gamma^2 H)^t H \preceq \gamma \sum_{t=0}^{\infty} (\mathrm{I} - \gamma H)^t H = \gamma (\gamma H)^{-1} H = \mathrm{I}$$

leads to

$$C_\infty \preceq \gamma \|\Sigma\|_H \sum_{t=0}^{\infty} (\gamma R^2)^t \mathrm{I} = \frac{\gamma \|\Sigma\|_H}{1 - \gamma R^2} \mathrm{I}\,,$$

which completes the proof. ◀

▶ **Lemma 7.** *(Refined C_∞ bound) The* $\mathrm{Tr}(C_\infty)$ *is bounded as:*

$$\mathrm{Tr}(C_\infty) \leq \frac{\gamma}{2} \mathrm{Tr}(H^{-1}\Sigma) + \frac{1}{2} \frac{\gamma^2 R^2}{1 - \gamma R^2} d \|\Sigma\|_H$$

Proof. From Lemma 6 and Equation 6,

$$\mathcal{S} \circ C_\infty \preceq \frac{\gamma \|\Sigma\|_H}{1 - \gamma R^2} \mathcal{S} \circ \mathrm{I} \preceq \frac{\gamma R^2 \|\Sigma\|_H}{1 - \gamma R^2} H\,.$$

Also, from Equation 2, C_∞ satisfies:

$$H C_\infty + C_\infty H = \gamma \mathcal{S} \circ C_\infty + \gamma \Sigma\,.$$

Multiplying this by H^{-1} and taking the trace leads to:

$$\begin{aligned} \mathrm{Tr}(C_\infty) &= \frac{\gamma}{2} \mathrm{Tr}(H^{-1} \cdot (\mathcal{S} \circ C_\infty)) + \frac{\gamma}{2} \mathrm{Tr}(H^{-1}\Sigma) \\ &\leq \frac{1}{2} \frac{\gamma^2 R^2}{1 - \gamma R^2} \|\Sigma\|_H \,\mathrm{Tr}(H^{-1} H) + \frac{\gamma}{2} \mathrm{Tr}(H^{-1}\Sigma) \\ &= \frac{1}{2} \frac{\gamma^2 R^2}{1 - \gamma R^2} d \|\Sigma\|_H + \frac{\gamma}{2} \mathrm{Tr}(H^{-1}\Sigma) \end{aligned}$$

which completes the proof. ◀

3.3 Completing the proof of Theorem 1

Proof. The proof of the theorem is completed by applying the developed lemmas. For the bias term, using convexity leads to:

$$\begin{aligned} \frac{1}{2}\mathbb{E}[\|\overline{w}_{t:T} - w^*\|_H^2 | \xi_0 = \cdots \xi_T = 0] &\leq \frac{1}{2} R^2 \mathbb{E}[\|\overline{w}_{t:T} - w^*\|^2 | \xi_0 = \cdots \xi_T = 0] \\ &\leq \frac{1}{2} \frac{R^2}{T - t} \sum_{t'=t}^{T-1} \mathbb{E}[\|w_{t'} - w^*\|^2 | \xi_0 = \cdots \xi_T = 0] \\ &\leq \frac{1}{2} \exp(-\gamma \mu t) R^2 \|w_0 - w^*\|^2\,. \end{aligned}$$

For the variance term, observe that

$$\frac{1}{2}\mathbb{E}[\|\overline{w}_{t:T} - w^*\|_H^2 | w_0 = w^*] \leq \frac{\mathrm{Tr}(C_\infty)}{\gamma(T - t)} \leq \frac{1}{T - t} \left(\frac{1}{2} \mathrm{Tr}(H^{-1}\Sigma) + \frac{1}{2} \frac{\gamma R^2}{1 - \gamma R^2} d \|\Sigma\|_H \right)\,,$$

which completes the proof. ◀

References

1 Francis R. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research (JMLR)*, volume 15, 2014.

2 Alexandre Défossez and Francis R. Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *AISTATS*, volume 38, 2015.

3 Aymeric Dieuleveut and Francis R. Bach. Non-parametric stochastic approximation with large step sizes. *The Annals of Statistics*, 2015.

4 Roy Frostig, Rong Ge, Sham M. Kakade, and Aaron Sidford. Competing with the empirical risk minimizer in a single pass. In *COLT*, 2015.

5 Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic approximation through mini-batching and tail-averaging. *CoRR*, abs/1610.03774, 2016.

6 Harold J. Kushner and Dean S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, 1978.

7 Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, 1998.

8 Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, volume 30, 1992.

9 David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. *Tech. Report, ORIE, Cornell University*, 1988.

10 Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Publishers, 2000.