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

Prateek Jain\textsuperscript{1}, Sham M. Kakade\textsuperscript{2}, Rahul Kidambi\textsuperscript{2}, Praneeth Netrapalli\textsuperscript{1}, Venkata Krishna Pillutla\textsuperscript{2}, and Aaron Sidford\textsuperscript{3}

\textsuperscript{1}Microsoft Research, Bangalore, India, \{prajain,praneeth\}@microsoft.com
\textsuperscript{2}University of Washington, Seattle, WA, USA, \{sham,pillutla\}@cs.washington.edu
\textsuperscript{3}Stanford University, Palo Alto, CA, USA, sidford@stanford.edu

July 24, 2018

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.

1 Introduction

Stochastic gradient descent is among the most commonly used practical algorithms for large scale stochastic optimization. The seminal result of \cite{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 \cite{1, 3, 2, 5} of this fact, which provide finite rates of convergence. Other recent algorithms also achieve the statistically optimal minimax rate, with finite convergence rates \cite{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 \cite{2, 5}.

\textbf{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:

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

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 $\gamma$ is a fixed stepsize.
Notation. For a symmetric positive definite matrix $A$ and a vector $x$, define:
$$\|x\|_A^2 := x^\top A x.$$ 
For a symmetric matrix $M$, define the induced matrix norm under $A$ as:
$$\|M\|_A := \max_{\|v\|=1} v^\top M v / \|A v\| = \|A^{-1/2} M A^{-1/2}\|.$$ 

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
$$\hat{w}_{\text{MLE}}^n := \arg \min_w \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (y_i - w \cdot x_i)^2,$$
where $\hat{w}_{\text{MLE}}^n$ denotes the MLE estimator over the $n$ samples.

This rate can be characterized as follows: define
$$\sigma_{\text{MLE}}^2 := \frac{1}{2} \mathbb{E} \left[ (y - w^* x)^2 \|x\|_{H^{-1}}^2 \right],$$
and the (asymptotic) rate of the MLE is $\sigma_{\text{MLE}}^2 / n$ [7, 10]. Precisely,
$$\lim_{n \to \infty} \frac{\mathbb{E}[L(\hat{w}_{\text{MLE}}^n)] - L(w^*)}{\sigma_{\text{MLE}}^2 / n} = 1,$$
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 $\eta$ being independent of $x$. Here, it is straightforward to see that:
$$\frac{\sigma_{\text{MLE}}^2}{n} = \frac{1}{2} \frac{d \sigma^2}{n}.$$
The rate of $\sigma_{\text{MLE}}^2 / n$ is still minimax optimal even among mis-specified models, where the additive noise assumption may not hold [3, 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:

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

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:

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

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 $$2\sigma_{\text{MLE}}^2 / \sqrt{T-t}$$, and the rate of the bias contraction is $$\mu / R^2$$. 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:

$$B_t := I - x_t x_t^\top.$$

The update rule can be written as:

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

$$= (I - \gamma x_t x_t^\top)(w_{t-1} - w^*) - \gamma \xi_t$$

$$= B_t(w_{t-1} - w^*) - \gamma \xi_t.$$
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,
\[
w_t - w^* = B_t \cdots B_1 (w_0 - w^*) - \gamma (\xi_t + B_t \xi_{t-1} + \cdots + B_t \cdots B_2 \xi_1). \tag{1}
\]

It is helpful to consider a certain bias and variance decomposition. Let us write:
\[
E[\| w_{t:T} - w^* \|^2_H | \xi_0 = \cdots = \xi_T = 0] := \frac{1}{(T-t)^2} E \left[ \left\| \sum_{\tau=t}^{T-1} B_\tau \cdots B_1 (w_0 - w^*) \right\|^2_H \right].
\]

and
\[
E[\| w_{t:T} - w^* \|^2_H | w_0 = w^*] = \left( \frac{\gamma}{T-t} \right)^2 \cdot E \left[ \left\| \sum_{\tau=t}^{T-1} (\xi_\tau + B_\tau \xi_{\tau-1} + \cdots + B_\tau \cdots B_2 \xi_1) \right\|^2_H \right].
\]

(The first conditional expectation notation slightly abuses notation, and should be taken as a definition\(^\dagger\)).

**Lemma 1.** The error is bounded as:
\[
E[L(w_{t:T})] - L(w^*) \leq \frac{1}{2} \left( \sqrt{E[\| w_{t:T} - w^* \|^2_H | \xi_0 = \cdots = \xi_T = 0]} + \sqrt{E[\| w_{t:T} - w^* \|^2_H | w_0 = w^*]} \right)^2.
\]

**Proof.** Equation\(^\dagger\) implies that:
\[
\overline{w}_{t:T} - w^* = \frac{1}{T-t} \sum_{\tau=t}^{T-1} B_\tau \cdots B_1 (w_0 - w^*) - \frac{\gamma}{T-t} \sum_{\tau=t}^{T-1} (\xi_\tau + B_\tau \xi_{\tau-1} + \cdots + B_\tau \cdots B_2 \xi_1).
\]

Now observe that for vector valued random variables \( u \) and \( v \), \((Eu^T H v)^2 \leq E[\| u \|^2_H | E[\| v \|^2_H]]\) implies
\[
E[\| u + v \|^2_H] \leq \left( \sqrt{E[\| u \|^2_H]} + \sqrt{E[\| v \|^2_H]} \right)^2,
\]
the proof of the lemma follows by noting that \( E[L(\overline{w}_{t:T})] - L(w^*) = \frac{1}{2} E[\| w_{t:T} - w^* \|^2_H] \).

**Bias.** The bias term is characterized as follows:

**Lemma 2.** For all \( t \),
\[
E[\| w_t - w^* \|^2 | \xi_0 = \cdots = \xi_T = 0] \leq \exp(-\gamma \mu t) \| w_0 - w^* \|^2.
\]

**Proof.** Assume \( \xi_t = 0 \) for all \( t \). Observe:
\[
E[\| w_t - w^* \|^2] = E[\| w_{t-1} - w^* \|^2 - 2\gamma (w_{t-1} - w^*)^\top E[xx^\top] (w_{t-1} - w^*) + \gamma^2 (w_{t-1} - w^*)^\top E[\| x \|^2 xx^\top] (w_{t-1} - w^*)
\]
\[
\leq E[\| w_{t-1} - w^* \|^2 - 2\gamma (w_{t-1} - w^*)^\top H (w_{t-1} - w^*) + \gamma^2 R^2 (w_{t-1} - w^*)^\top H (w_{t-1} - w^*)
\]
\[
\leq E[\| w_{t-1} - w^* \|^2 - \gamma E[\| w_{t-1} - w^* \|^2_H]
\]
\[
\leq (1 - \gamma \mu) E[\| w_{t-1} - w^* \|^2_H],
\]
which completes the proof.\(\square\)

\(^\dagger\)The abuse is due that the right hand side drops the conditioning.
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, \( w_t - w^* = B_t(w_{t-1} - w^*) + \gamma \xi_t \),

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

which follows from:

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

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

**Lemma 3.** Suppose \( w_0 = w^* \). There exists a unique \( C_\infty \) such that:

\[
0 = C_0 \preceq C_1 \preceq \cdots \preceq C_\infty
\]

where \( C_\infty \) satisfies:

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

**Proof.** By recursion,

\[
w_t - w^* = B_t(w_{t-1} - w^*) + \gamma \xi_t = \gamma (\xi_t + B_t \xi_{t-1} + \cdots + B_t \cdots B_2 \xi_1).
\]

Using that \( \xi_t \) is mean zero and independent of \( B_t \) and \( \xi_t \) for \( t < t' \),

\[
C_t = \gamma^2 (\mathbb{E}[\xi_t \xi_t^\top] + \mathbb{E}[B_t \xi_{t-1} \xi_{t-1}^\top B_t] + \cdots + \mathbb{E}[B_t \cdots B_2 \xi_1 \xi_1^\top B_2^\top \cdots B_1^\top])
\]

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

\[
C_t = \gamma^2 (\Sigma + \mathbb{E}[B_2 \Sigma B_2] + \cdots + \mathbb{E}[B_t \cdots B_2 \Sigma B_2]) = C_{t-1} + \gamma^2 \mathbb{E}[B_t \cdots B_2 \Sigma B_2^\top \cdots B_1^\top]
\]

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 2 for \( C_t \),

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

Observe:

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

and, using \( \gamma \leq 1/R^2 \),

\[
\text{Tr}(C_{t+1}) \leq \text{Tr}(C_t) - \gamma \text{Tr}(HC_t) + \gamma^2 \text{Tr} = (1 - \gamma \mu) \text{Tr}(C_t) + \gamma^2 \text{Tr}(\Sigma) \leq \frac{\gamma \text{Tr}(\Sigma)}{\mu}
\]

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. \( \square \)

**Lemma 4.** Define:

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

and so:

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

5
Proof. Note

\[ \mathbb{E}[(w_T - w^*)(w_T - w^*)^T | w_0 = w^*] = \frac{1}{T^2} \sum_{t=0}^{T-1} \sum_{t'=0}^{T-1} \mathbb{E}[(w_t - w^*)(w_{t'} - w^*)^T | w_0 = w^*] \]

\[ \leq \frac{1}{T^2} \sum_{t=0}^{T-1} \sum_{t'=1}^{T-1} \left( \mathbb{E}[(w_t - w^*)(w_{t'} - w^*)^T | w_0 = w^*] + \mathbb{E}[(w_{t'} - w^*)(w_t - w^*)^T | w_0 = w^*] \right) , \]

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

\[ \mathbb{E}[(w_T - w^*)(w_T - w^*)^T | w_0 = w^*] \leq \frac{1}{T^2} \sum_{t=0}^{T-1} \sum_{t=0}^{T-1} (I - \gamma H)^T C_t + C_t (I - \gamma H)^T , \]

and so,

\[ \mathbb{E}[\|w_T - w^*\|^2 | w_0 = w^*] = \text{Tr} \left( H \mathbb{E}[(w_T - w^*)(w_T - w^*)^T | w_0 = w^*] \right) \]

\[ \leq \frac{1}{T^2} \sum_{t=0}^{T-1} \sum_{t=0}^{T-1} \text{Tr}(H(I - \gamma H)^T C_t) + \text{Tr}(C_t (I - \gamma H)^T H) . \]

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

\[ \mathbb{E}[\|w_T - w^*\|^2 | w_0 = w^*] \leq \frac{2}{T^2} \sum_{t=0}^{T-1} \sum_{t=0}^{\infty} \text{Tr}(H(I - \gamma H)^T C_t) \]

\[ = \frac{2}{T^2} \sum_{t=0}^{T-1} \text{Tr}(H(\sum_{\tau=0}^{\infty} (I - \gamma H)^\tau) C_t) \]

\[ = \frac{2}{T^2} \sum_{t=0}^{T-1} \text{Tr}(H(\gamma H)^{-1} C_t) \]

\[ = \frac{2}{T^2} \sum_{t=0}^{T-1} \text{Tr}(C_t) \]

\[ \leq \frac{2}{\gamma T} \cdot \text{Tr}(C_\infty) , \]

from lemma where \((*)\) followed from

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

and the series converges because \( I - \gamma H \prec I \). \( \square \)
3.2 Stationary Distribution Analysis

Define two linear operators on symmetric matrices, $S$ and $T$ — where $S$ and $T$ can be viewed as matrices acting on $(d+1)/2$ dimensions — as follows:

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

With this, $C_\infty$ is the solution to:

$$ T \circ C_\infty = \gamma S \circ C_\infty + \gamma \Sigma $$

(due to Equation 3).

**Lemma 5.** (Crude $C_\infty$ bound) $C_\infty$ is bounded as:

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

**Proof.** Define one more linear operator as follows:

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

The inverse of this operator can be written as:

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

which exists since the sum converges due to fact that $0 \preceq I - \gamma H \prec I$.

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

$$ 0 \preceq \tilde{T}^{-1} \circ M \preceq \tilde{T}^{-1} \circ M', $$

since

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

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

$$ 0 \preceq S \circ M \preceq S \circ M', $$

which implies:

$$ 0 \preceq \tilde{T}^{-1} \circ S \circ M \preceq \tilde{T}^{-1} \circ S \circ M'. $$

The following inequality is also of use:

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

By definition of $\tilde{T}$,

$$ \tilde{T} \circ C_\infty = \gamma S \circ C_\infty + \gamma \Sigma - \gamma HC_\infty H. $$

Using this and Equation 6

$$ C_\infty = \gamma \tilde{T}^{-1} \circ S \circ C_\infty + \gamma \tilde{T}^{-1} \circ \Sigma - \gamma \tilde{T}^{-1} \circ (HC_\infty H) \preceq \gamma \tilde{T}^{-1} \circ S \circ C_\infty + \gamma \tilde{T}^{-1} \circ \Sigma \preceq \gamma \tilde{T}^{-1} \circ S \circ C_\infty + \gamma \|\Sigma\|_H \tilde{T}^{-1} \circ H. $$
Proceeding recursively by using Equation 8,
\[ C_\infty \preceq (\gamma \bar{T}^{-1} \circ S)^2 \circ C_\infty + \gamma \|\Sigma\|_H (\gamma \bar{T}^{-1} \circ S) \circ \bar{T}^{-1} \circ H + \gamma \|\Sigma\|_H \bar{T}^{-1} \circ H. \]

Using
\[ S \circ I \preceq R^2 H \]
and
\[ \bar{T}^{-1} \circ H = \gamma \sum_{t=0}^\infty (1-\gamma H)^t H = \gamma \sum_{t=0}^\infty (1-\gamma 2H + \gamma^2 H)^t H \leq \gamma \sum_{t=0}^\infty (1-\gamma H)^t H = \gamma (\gamma H)^{-1} H = I \]
leads to
\[ C_\infty \preceq \gamma \|\Sigma\|_H \sum_{t=0}^\infty (\gamma R^2)^t I = \frac{\gamma\|\Sigma\|_H}{1-\gamma R^2} I, \]
which completes the proof.

Lemma 6. (Refined \( C_\infty \) bound) The \( \text{Tr}(C_\infty) \) is bounded as:
\[ \text{Tr}(C_\infty) \leq \frac{\gamma}{2} \text{Tr}(H^{-1} \cdot (S \circ C_\infty)) + \frac{1}{2} \frac{\gamma^2 R^2}{1-\gamma^2 R^2} d \|\Sigma\|_H. \]

Proof. From Lemma 5 and Equation 8,
\[ S \circ C_\infty \preceq \frac{\gamma\|\Sigma\|_H}{1-\gamma R^2} S \circ I \leq \frac{\gamma R^2\|\Sigma\|_H}{1-\gamma R^2} H. \]
Also, from Equation 3 \( C_\infty \) satisfies:
\[ H C_\infty + C_\infty H = \gamma S \circ C_\infty + \gamma \Sigma. \]
Multiplying this by \( H^{-1} \) and taking the trace leads to:
\[ \text{Tr}(C_\infty) = \frac{\gamma}{2} \text{Tr}(H^{-1} \cdot (S \circ C_\infty)) + \frac{\gamma}{2} \text{Tr}(H^{-1} \Sigma) \leq \frac{1}{2} \frac{\gamma^2 R^2}{1-\gamma R^2} \|\Sigma\|_H \text{Tr}(H^{-1} H) + \frac{\gamma}{2} \text{Tr}(H^{-1} \Sigma) \]
\[ = \frac{1}{2} \frac{\gamma^2 R^2}{1-\gamma R^2} d \|\Sigma\|_H + \frac{\gamma}{2} \text{Tr}(H^{-1} \Sigma) \]
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:
\[ \frac{1}{2} \mathbb{E}[\|\mathbf{w}_{t:T} - \mathbf{w}^*\|_H^2 | \xi_0 = \cdots \xi_T = 0] \leq \frac{1}{2} \frac{R^2 \mathbb{E}[\|\mathbf{w}_{t:T} - \mathbf{w}^*\|^2 | \xi_0 = \cdots \xi_T = 0]}{T - t} \leq \frac{1}{2} \exp(-\gamma \mu t) R^2 \|w_0 - w^*\|^2. \]
For the variance term, observe that
\[
\frac{1}{2} \mathbb{E}[\|w_{T} - w^*\|^2_H | w_0 = w^*] \leq \frac{\text{Tr}(C_\infty)}{\gamma(T-t)} \leq \frac{1}{T-t} \left( \frac{1}{2} \text{Tr}(H^{-1}\Sigma) + \frac{1}{2} \frac{\gamma R^2}{1 - \gamma R^2} d\|\Sigma\|_H \right),
\]
which completes the proof.

**Acknowledgements.** Sham Kakade acknowledges funding from the Washington Research Foundation Fund for Innovation in Data-Intensive Discovery. The authors also thank Zaid Harchaoui for helpful discussions.

**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.