Polynomial convergence of iterations of certain random operators in Hilbert space

Soumyadip Ghosh, Yingdong Lu, Tomasz Nowicki
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

Abstract
We study the convergence of a random iterative sequence of a family of operators on infinite dimensional Hilbert spaces, inspired by the Stochastic Gradient Descent (SGD) algorithm in the case of the noiseless regression, as studied in [1]. We identify conditions that are strictly broader than previously known for polynomial convergence rate in various norms, and characterize the roles the randomness plays in determining the best multiplicative constants. Additionally, we prove almost sure convergence of the sequence.

Keywords: polynomial convergence, random operators, Stochastic Gradient Descent algorithm

2000 MSC: 46N10, 47B80, 60B10

1. Introduction

On a real Hilbert space $\mathbb{H}$ with inner product $\langle \cdot | \cdot \rangle$, define a family of rank 1 operators $S_\gamma$ for $x \in \mathbb{H}$, and for given $\gamma \in [0, 1)$ a family of operators $T_\gamma$ acting on $\mathbb{H}$ by

$$
S_{\gamma} : \mathbb{H} \ni \theta \mapsto \langle \theta | x \rangle x \in \mathbb{H}, \quad T_{\gamma} : \mathbb{H} \ni \theta \mapsto \theta - \gamma S_{\gamma} \theta \in \mathbb{H}.
$$

(1)

The operator $T_\gamma$ is motivated by the steps of the stochastic gradient descent (SGD) algorithm for a noiseless linear regression problem in infinite dimension, see e.g. [1]. Assume that there exists an optimal parameter $\theta^* \in \mathbb{H}$ such that the data $y \in \mathbb{R}$ and $x \in \mathbb{H}$ always satisfy $y = \langle \theta^* | x \rangle$. In SGD applications, the task of determining $\theta^*$ with respect to the independent sampling $(x(1), y(1)), \ldots, (x(n), y(n)), \ldots$ using the cost function $L(\vartheta | x) = (y - \langle \vartheta | x \rangle)^2 = \langle \vartheta - \vartheta^* | x \rangle^2$ (derived from the assumption) is carried by the following iterative scheme: given initial $\vartheta_0 \in \mathbb{H}$ (usually for practical reasons $\vartheta_0 = 0$, but the convergence should not depend on it) we set

$$
\vartheta(n + 1) = \vartheta(n) - \gamma \frac{\partial L}{\partial \vartheta}(\vartheta(n)) = \vartheta(n) - \gamma (\vartheta(n) - \vartheta^*) x(n) \cdot x(n).
$$

(2)

The parameter $\gamma > 0$ is a small step size along the negative gradient of the cost function. In Equation (1), $\theta = \vartheta - \vartheta^*$ represents the difference between the output of the algorithm and the optimum. In terms of $\theta$ the cost function is equal to $L(\theta | x) = \langle \theta | x \rangle^2$. To prove that $\vartheta(n) \to \vartheta^*$ is now equivalent to prove that $\theta(n) \to 0$. We note that the properties of this model are invariant under a normalization, i.e. a rescaling of the variables $x$ and $y$ by some (same) constant and the parameter $\gamma$ by the square of its inverse.

Conceptually, $S_\gamma$ projects $\theta$ to the $x$ direction (with the factor $\|x\|^2$), and $T_\gamma$ takes a proportion $\gamma$ of the image of the projection away from the original $\theta$. When $T_\gamma$ is iterated for randomly selected $x$, and for $\gamma$ small enough, one would expect that the image, hence the error of the algorithm, eventually vanishes.

We prove the polynomial convergence rate of the average of the sequence which is explicitly determined only by the regularity of the initial state, Theorem 1 and 2. For convergence of the second moment, under a condition on the regularity of the random distribution (Assumption (A)), the convergence rate remains the same, Theorem 3. In another words, under (A), the regularity of the random sequence only affects the coefficient not the order of the polynomial convergence. Additionally we demonstrate almost sure convergence of the sequence, Theorem 4.
The rest of the paper will be organized as follow: in Section 2, we present our main results and their implications; in Section 3, we discuss the basic properties of the key operators and some key assumptions of the papers; the proofs the convergence rates are presented in Section 4, while the proof of the almost sure convergence is presented in Section 5, with proofs of technical lemmata collected in Section 6.

2. Main results

For iid random variables $x(1), \ldots, x(n), \cdots \in \mathbb{H}$ and given $\theta(0)$, the recursive definition (2) becomes

$$\theta(n + 1) = \theta(n) - \gamma \langle \theta(n), x(n) \rangle = T_{x(n)}(\theta(n)).$$

Furthermore, define the average operators $S$ and $T$ of $S_x$ and $T_x$ by

$$S = \mathbb{E}[S_x] : \mathbb{H} \to \mathbb{H}, \quad T = \mathbb{E}[T_x] : \mathbb{H} \to \mathbb{H},$$

where the symbol $\mathbb{E}[\cdot]$ denotes the expected value w.r.t. the distribution of the vector $x$, but also the expected value w.r.t. the product distribution of the samples. We assume that $S$ and $T$ are bounded and well defined on $\mathbb{H}$, for which it is enough to assume that $\mathbb{E}[(\|x\|^2) < \infty$. We note that $S$ (as we shall see being symmetric), when defined on all $\mathbb{H}$, is bounded by Hellinger–Toeplitz Theorem, (for basic materials and theorems of functional analysis used in this paper, see, e.g. [2]) even without the condition on $\mathbb{E}[(\|x\|^2)$. The operators have finite norms, in particular $\|S_x\|^2 < \infty$. Because $S$ is also non-negative, the powers $S^n$ are well defined for (some) real values of $\beta$, certainly for all $\beta \geq 0$, $S^0 = 1d$ and $S^1 = S$.

Example 1. The basic example illustrating the variable $x$ to keep in mind is related to the Gaussian Free Field [3]. Let $(e_i)_{i=1}^{\infty}$ be any orthonormal basis in $\mathbb{H}$. Define the random variable $x = \sum_{i=0}^{\infty} x_i e_i$, where $x_i$ are independent variables with mean 0 and variances $\mathbb{E}[x_i^2] = \lambda_i$, note that for $i \neq j$, $\mathbb{E}[x_i x_j] = \mathbb{E}[x_i] \mathbb{E}[x_j] = 0$. In this setting, with $\theta = \sum_i \theta_i e_i$ we have $\langle \theta | x \rangle = \left( \sum_{i=0}^{\infty} \theta_i | x_i e_i \rangle = \sum_i \theta_i \mathbb{E}[x_i] e_i \right)$

$$S\theta = \mathbb{E}[S_x \theta] = \mathbb{E}[\langle \theta | x \rangle | x] = \mathbb{E} \left[ \sum_i \theta_i (x_i - \mathbb{E}[x_i]) \cdot \sum_k (x_k e_k) \right] = \sum_k \sum_i (\theta_i \mathbb{E}[x_i x_k] e_k) = \sum_i \theta_i \mathbb{E}[x_i^2] e_i = \sum_{i=0}^{\infty} \lambda_i \theta_i e_i.$$

We conclude that $S \theta \in \mathbb{H}$ for every $\theta \in \mathbb{H}$ iff $\lambda_i$ are uniformly bounded.

We shall investigate the rate of convergence by using the "norms"

$$\varphi_\beta : \mathbb{H} \to \mathbb{R}, \quad \mathbb{H} \ni \theta \mapsto \varphi_\beta(\theta) = \langle \theta | S^\beta \theta \rangle = : \|\theta\|_\beta^\beta,$$

given $\theta(0)$, $\phi_\beta : \mathbb{R} \to \mathbb{R}$, $\mathbb{R} \ni \beta \mapsto \phi_\beta(\beta) = \mathbb{E}[\varphi_\beta(\theta(n))] = \mathbb{E}[(\|\theta(n)\|_\beta^\beta)].$

The numbers $\phi$ depend on the starting $\theta(0)$ but, due to the expected value, not on the choice of the samples $(x)$. We introduce the limits of applicable $\beta$, for $\theta, x \in \mathbb{H}$, as,

$$\alpha(\theta) = \sup(\beta : \varphi_\beta(\beta) < \infty) \quad \text{and} \quad \alpha = \sup(\beta : \mathbb{E}[\varphi_\beta(\beta)] < \infty).$$

We have $\alpha(\theta) \geq 0$ and, as we shall see (Lemma 3), $\alpha \leq 1$.

First two Theorems bound $\phi_\beta(\beta)$ for averages $\mathbb{E}[\theta(n)] = T^n \theta(0)$, depending only on $\varphi_\beta(\theta(0))$, where $\theta(0) = \theta(0) - \theta^* (= -\theta^*)$.

**Theorem 1** (Upper bound for the average $\theta(n)$). Given $\theta(0) = \theta$ and $T^n \theta = \mathbb{E}[\theta(n)]$ we have,

$$\text{for every } n, \quad \|T^n \theta\|_\beta^\beta \leq \exp(-\beta) (\frac{B^\beta}{n})^\beta \cdot \|\theta\|_\beta^\beta.$$

**Theorem 2** (Lower bound for the average $\theta(n)$). Given $\theta(0) = \theta$ and $T^n \theta = \mathbb{E}[\theta(n)]$, for any sequence $(t_n) > 0$ such that $\sum_n 1/(nt_n) < \infty$, we have,

$$\text{if } \|T^n \theta\|_\beta^\beta \leq \frac{1}{nt_n} \text{ for every } n, \text{ then } \|\theta\|_\beta^\beta < \infty.$$
Examples of slow increasing sequences $t_n$ with $\sum_n 1/(nt_n) < \infty$ are $n^\epsilon$, $(\ln n)^{1+\epsilon}$ or $\ln n \cdot (\ln \ln n)^{1+\epsilon}$ etc. with any $\epsilon > 0$.

In order to produce the upper bound of the square of $\theta(n)$, we need an additional assumption. First let’s define a family of inequalities:

$$
E \left[ (\theta | x)^2 (x | S^{-\beta} x) \right] \leq C_{\beta \kappa} \left( \theta | S^{-\beta} \theta \right) ( = C_{\beta \kappa} \varphi_{\kappa-1}(\theta)) \tag{C_{\beta \kappa}}.
$$

Assumption (A). There is an $\alpha > 0$ such that the distribution of $x$ satisfies (C_{\beta \kappa}) with $\kappa = \beta < \alpha$. For every $\beta < \alpha$ there is a $C_{\beta}$ (:= $C_{\beta \beta}$) such that for every $\theta \in \mathbb{H}$:

$$
E \left[ (\theta | x)^2 \varphi_{\beta}(x) \right] \leq C_{\beta} \left( \theta | S^{-\beta} \theta \right) = C_{\beta \kappa} \varphi_{\kappa-1}(\theta).
$$

Theorem 3 (Upper bound for the average $\|\theta(n)\|^2$). Assuming (A), for any $0 \leq \beta < \alpha(\theta)$ if we take $\gamma < 2/C_{\beta}$ then $E \left( \|\theta(n)\|^2 \right) \leq O(1)n^{-\beta}$.

The average of $\|\theta(n)\|^2$ is lower bounded by $E \left( \|x\|^2 \right)$. Theorem 2 applied to $\beta > \alpha(\theta)$ allows us to take $t_n = Cn^{\beta-\kappa}$, $\alpha(\theta) < \kappa < \beta$. Thus $\|T^{\alpha(\theta)}\|^2$ cannot be bounded by $Cn^{\kappa} = Cn^{\kappa}n_t$, as it would imply $\|\theta\|_2^2 < \infty$ a contradiction to $\kappa > \alpha(\theta)$.

Convergence in norms, including the average, implies that $\theta_n$ converge to zero in probability as a sequence of random variables in $\mathbb{H}$, see, e.g. [4]. It is natural to examine the almost sure convergence of the $\theta_n$.

Theorem 4 (Almost sure convergence). If $E \left( \|x\|^2 \right) < \infty$ and $\delta := \inf_{\|x\| = 1} E \left( (\theta | x)^2 \right) > 0$ then the sequence $\theta_n$ converges to zero almost surely for $\gamma < \delta/E \left( \|x\|^2 \right)$. If for some $M > 0$, $\|x\|^2 \leq M$ almost surely then such convergence occurs for $\gamma \leq 2/M$.

We shall see in Proposition 1 that the condition in Theorem 4, $E \left( \|x\|^2 \right) < \infty$, is satisfied under Assumption (A).

For proofs of Theorems 1, 2, and 3, see Section 4, Propositions 2, 3, and 4 with $\kappa = 0$. For proof of Theorem 4 see Section 5. Example 2. Let the distribution of $x$ in Example 1 be even with the property that $y_i = x_i^2$ has a $\Gamma$-density $t^{ri-1} \exp(-t) / \Gamma(\lambda_i)$ for $t \in [0, \infty)$. Therefore, $E \left[ x_i^2 \right] = E \left[ y_i \right] = \int_0^\infty t^{ri} \exp(-t) dt / \Gamma(\lambda_i) = \Gamma(\lambda_i + 1) / \Gamma(\lambda_i) = \lambda_i$. Similarly $E \left[ x_i^2 \right] = E \left[ y_i^2 \right] = \lambda_i(1 + \lambda_i)$. Given $\alpha \in (0, 1)$ let $0 < \lambda_i \searrow 0$ be such that $\sum_{i=0}^n \lambda_i^{-1-\beta} =: K_{\beta} < \infty$ for any $\beta \in (0, \alpha)$.

Some elementary calculation (see Section 6.1) using which we can make moments provide that LHS of (A) is equal to

$$
E \left[ (\theta | x)^2 \varphi_{\beta}(x) \right] = E \left[ \left( \sum_i \theta_i x_i \right) \cdot \left( \sum_i \theta_i x_i \right) \cdot \left( \sum_j \lambda_j^{-\beta} x_j^2 \right) \right] \sum_i \theta_i^2 \lambda_i \cdot K_{\beta} + \sum_i \theta_i^2 \lambda_i^{-1-\beta}.
$$

Now we can use Assumption (B) below (which is not essential, as without it we would just have a less pleasant constant) and get an upper bound for the LHS by $(K_{\beta} + 1) \left( \theta | S^{-\beta} \theta \right)$, which is the RHS of (A) with $C_{\beta} = K_{\beta} + 1$. That proves that our example satisfies Assumption (A).

Remark 1. This example also shows that the bound in (A) is accurate. The collection of $\theta$’s satisfying the inequality (C_{\beta \kappa}) with $\kappa = \beta$, that is Assumption (A), is larger than the collection satisfying a stronger assumption, the inequality (C_{\beta \kappa}) with $\kappa < \beta$. In particular this applies to $\kappa = 0$, which is the condition used in [1]. Example 2 provides a family of distributions that satisfy Assumption (A) but not (C_{\beta \kappa}) with $\kappa = 0$. Indeed, as $\theta$ is arbitrary we can take in our example $\theta = e_1$. Then the LHS will be equal (as in the last expression above) to $K_{\beta} \lambda_i + \lambda_i^{-\beta}$ which, for any given $C_{\beta \kappa}$ and for sufficiently large $i$ (and therefore small $\lambda_i$) is larger than the RHS equal to $C_{\beta \kappa} \lambda_i^{-1-\beta}$ due to $1 - \kappa > 1 - \beta$ and $\lambda_i \searrow 0$. 

3
3. Properties of the operators

In this section we present basic properties of the operators, which can be easily deduced directly from the definitions. Recall the definitions of $S_{x}$ and $T_{x}$ in (1) and their averages $S$ and $T$ in (4). We assume that both $S_{x}$ and $S$ are bounded and well defined for all $\theta \in \mathbb{H}$.

**Property 1** ($S_{x}$ and the average $S$ are symmetric and non-negative).

- $\langle \eta | S_{x} \theta \rangle = \langle \eta | x \rangle \langle \theta | x \rangle = \langle \theta | S_{x} \eta \rangle$;
- $\langle \eta | S \theta \rangle = \mathbb{E}[(\eta | S_{x} \theta \rangle] = \mathbb{E}[(\theta | S_{x} \eta \rangle] = \langle \theta | S \eta \rangle$;
- Non-negativity: $\langle \theta | S_{x} \theta \rangle = \langle \theta | x \rangle^{2}$.

**Property 2** ($S$ admits an orthonormal (ON) basis of eigen-vectors).

- As the operator $S$ is symmetric, non-negative and defined on all $\mathbb{H}$, it has an ON basis $(\mathbf{e}_{i})$ of eigen-vectors, with corresponding bounded non-negative eigenvalues $(\lambda_{i})$.
- If in this basis $\theta = \sum \theta_{i} \mathbf{e}_{i}$, then $S \theta = \sum \lambda_{i} \theta_{i} \mathbf{e}_{i}$.

**Property 3** (The moments of $x$).

- Each feature coordinate $x_{i}$ of $x$ in the ON basis $(\mathbf{e})$ has finite second moment: $\mathbb{E}[x_{i}^{2}] = \lambda_{i}$.
- Using $(\mathbf{e}_{i}|\mathbf{e}_{j}) = 1$ and $(\mathbf{e}_{i}|x) = x_{i}$ for the features vector $x = \sum_{i} \mathbf{e}_{i}$, we obtain:
  $$\lambda_{i} = (\mathbf{e}_{i}|\lambda \mathbf{e}_{i}) = (\mathbf{e}_{i}|S \mathbf{e}_{i}) = (\mathbf{e}_{i}|\mathbb{E}[S \mathbf{e}_{i}]) = \mathbb{E}[(\mathbf{e}_{i}|x \mathbf{e}_{i})] = \mathbb{E}[x_{i}^{2}]$$.

- The coordinates of $x$ in the ON basis $(\mathbf{e})$ are de-correlated: $\mathbb{E}[x_{i}x_{j}] = 0$ (are uncorrelated if $\mathbb{E}[x] = 0$).
- Using $\mathbf{e}_{i} + \mathbf{e}_{j}$ and the orthonormality we get
  $$\lambda_{i} + \lambda_{j} = (\mathbf{e}_{i} + \mathbf{e}_{j}|\lambda \mathbf{e}_{i} + \lambda \mathbf{e}_{j}) = (\mathbf{e}_{i} + \mathbf{e}_{j}|S \mathbf{e}_{i} + \mathbf{e}_{j}) = (\mathbf{e}_{i} + \mathbf{e}_{j}|\mathbb{E}[S \mathbf{e}_{i} + \mathbf{e}_{j}]) = \mathbb{E}[(\mathbf{e}_{i} + \mathbf{e}_{j}|x \mathbf{e}_{i} + \mathbf{e}_{j})] = \mathbb{E}[(x_{i} + x_{j})^{2}] = \lambda_{i} + 2\mathbb{E}[x_{i}x_{j}] + \lambda_{j}$$

- Special form of $S$ in the ON basis.
  $\mathbb{E}[(\theta | S_{x} \theta \rangle] = \mathbb{E}[(\theta | x \rangle^{2}] = \mathbb{E}[|x_{i}|^{2}] = \mathbb{E}[\sum_{i} \theta_{i}^{2}] = \mathbb{E}[\sum_{i} \theta_{i}^{2} \mathbb{E}[x_{i}^{2}]] = \sum_{i} \lambda_{i} \theta_{i}^{2}$.

We note that when $\lambda_{i} = 0$ we have $\mathbb{E}[x_{i}^{2}] = 0$, so that $x_{i} = 0$ a.s. and we may restrict ourselves to the closure of the subspace $\mathcal{H} = \sum_{i=0}^{\infty} h_{i} \mathbf{e}_{i} \subset \mathbb{H}$, where $S \theta = 0$ only when $\theta = 0$. We have $\mathbb{E}[\varphi(\mathbf{x})] = \mathbb{E}[\left(\sum x_{i}^{2}\right)^{\frac{1}{2}}] = \mathbb{E}[\left(\sum_{i=0}^{\infty} \lambda_{i}^{1-\beta} \mathbb{E}[x_{i}^{2}] \right)^{\frac{1}{2}}] = \sum_{i=0}^{\infty} \lambda_{i}^{1-\beta} \mathbb{E}[x_{i}^{2}] = \sum_{i=0}^{\infty} \lambda_{i}^{1-\beta}$. In particular the sum is infinite for $\beta > 0$ as $\lambda_{i}$ are bounded so $\alpha < 1$. Also $\mathbb{E}[\mathbb{E}[x^{2}]] = \mathbb{E}[\sum x_{i}^{2}] = \sum \lambda_{i}$. From now on we shall use

**Assumption (B).** For any eigenvalue $\lambda$ in the spectrum of $S$ we have $0 < \lambda < \frac{1}{\beta} < 1$.

This is not a loss of generality. The operator is continuous, hence bounded and its spectrum is compact. It is positive and symmetric. Let $\lambda_{0} = \sup \lambda$. As we are interested in the iterations of $T_{x} = 1 - \gamma S_{x}$ for small $\gamma$ we may assume that $\gamma < 1/2\lambda_{0}$ by changing either $x$ (and $y$) to $x/2\lambda_{0}$ (and to $y/2\lambda_{0}$) or changing $S_{x}$ to $\theta \mapsto \langle \theta | x \rangle \cdot x/2\lambda_{0}$, effectively using $\gamma' = \gamma \cdot 2\lambda_{0}$.

Using the ON basis the operators $S^{\beta}: \mathbb{H} \rightarrow \mathbb{H}$ can be now defined by $S^{\beta} \theta = \sum \lambda_{i}^{\beta} \theta \mathbf{e}_{i}$.

With the definitions (5) of $\alpha(\theta)$ and $\alpha$ from Section 1 we have,

**Property 4** (Bounds on the powers $S^{\beta}$). (1) Given $\theta$, $\|\theta\|_{0}^{2}$ is an increasing function of $\beta$; (2) $\alpha(\theta) \geq 0$; and (3) $\alpha \leq 1$, independently of the distribution of data $x$. If $\|x\|_{2} < \infty$ then $\alpha \geq 0$ and $\sum \lambda_{i} < \infty$. 

4
**Proposition 1 (Bounds on moments).** Suppose that Assumption (A) is satisfied with \( \beta = 0 \), which means that there exists a \( C_0 \) such that for all \( \theta \) we have \( \mathbb{E}[(\theta|x)^2]|x|^2] \leq C_0(\theta|S\theta) \), then
\[
\mathbb{E}[\|x\|^2]^2 \leq \mathbb{E}[|x|^2]^2 \leq C_0^2.
\]

For proof see Section 6.2. Proposition 1 implies that both the second and the fourth moments of \( x \) are finite.

**Lemma 2 (Series and function \( \Gamma \)).** For any \( \alpha > 0 \) there exists a constant \( K \) such that for every \( 0 < \mu < 1/2 \) and \( 0 < \kappa < \alpha \) we have \( K\Gamma(\kappa) < \sum_n(1-\mu)^n(n\mu)^\kappa/n \leq K^{-1}\Gamma(\kappa) \), where, for \( \Re(z) > 0 \), \( \Gamma(z) = \int_0^\infty e^{-t}t^{z-1}dt \). (Proof: see Section 6.4.)

**Proposition 3 (Lower bound).** Let the sequence \( (\kappa_n) \) be such that \( \sum_n 1/(n\kappa_n) < \infty \).

if for some \( 0 \leq \kappa < \beta \), \( \phi_0(\kappa) = \|T^n\|_2^2 \leq \frac{1}{n^{\beta-x}/\Gamma_n} \) for all \( n \), then \( \|\theta\|_p^2 < \infty \).
The arbitrary sequence $t_n$ in Proposition 3 is mostly interesting in case $||\theta||_{\alpha(\theta)} = \infty$.

**Proof.** We use again the convention $q_i = -\ln(1 - \gamma\lambda_i) \in (0, \ln 4)$

\[
\|T^n\|_c^2 = O(1)\gamma^{n-k}\sum_i \exp(-nq_i)(q_i)^{\beta-k}\lambda_i^{-\beta}\theta_i^2,
\]

\[
\infty > \sum_n \frac{1}{m_n} \cong \sum_n \frac{x^{\beta-k}}{n^{\beta-k}}||T^n\|_c^2 = O(1)\sum_n \left[ \sum_i \exp(-nq_i)(q_i)^{\beta-k-1}q_i \cdot \lambda_i^{-\beta}\theta_i^2 \right]
\]

\[
= O(1)\sum_i \left[ \sum_i \exp(-nq_i)(nq_i)^{\beta-k-1}q_i \right] \cdot \lambda_i^{-\beta}\theta_i^2 = O(1)\sum_i \Gamma(\beta - k) \cdot \lambda_i^{-\beta}\theta_i^2
\]

\[
= O(1)\Gamma(\beta - k) \cdot \sum_i \lambda_i^{-\beta}\theta_i^2 = O(1)\Gamma(\beta - k)||\theta||_\beta^2,
\]

where we approximated the series by the integral as in Lemma 2 and changed the variables in the integral. \(\square\)

**Lemma 3.** Let $0 < a_n < 1$ satisfies $a_{n+1} \leq a_n - a_1^{1+w}$ for some $w > 0$. Then $a_n \leq a_0(1 + nw\alpha_0)^{-1/w}$. If $c_{n+1} \leq c_n - Kc_n^{1+w}$ then $c_n \leq c_0(1 + kwK_0\alpha_0)^{-1/w}$. (Proof: see Section 6.5.)

**Lemma 4** (Hölder inequality for $\varphi$, see [1]). Let $\beta < \kappa < \alpha$ and $p = \frac{\alpha - \beta}{\beta - \kappa}$. Then

\[
\varphi_\kappa \leq \varphi_\beta^{1-p}\varphi_\alpha^{1-p} = \varphi_\beta^{1-p}\frac{\varphi_\beta^{1+p}}{\varphi_\alpha^{1+p}} \quad \text{and} \quad \varphi_\beta \geq \varphi_\kappa^{1-\frac{1}{p}}\varphi_\alpha^{1-\frac{1}{p}}.
\]

**Proof.** We have $\kappa = p\beta + (1 - p)\alpha$ and $\varphi_\kappa(\theta) = \sum \lambda_i^2 \theta_i^2 = \sum \lambda_i^{p(1+p)(1-\beta)} \theta_i^{2(1-\beta)} = \sum (\lambda_i^p \theta_i^2)^{p(1-p)} \leq (\sum \lambda_i^p \theta_i^2)^{p(1-p)}$.

**Lemma 5** (Main recursion formula, see [1]).

\[
E[\varphi_\beta(Tx\theta)] = \varphi_\beta(\theta) - 2\gamma\varphi_{\beta-1}(\theta) + \gamma^2E[(\theta|x)^2 \langle x|S^\beta x \rangle]. \tag{6}
\]

(Proof: see Section 6.6.)

Another form of the last term of (6) is $E[(\theta|x)^2 \langle x|S^\beta x \rangle] = E[(\theta|Sx\theta) \varphi_\beta(x)]$.

**Corollary 1.** Under (A), if $\gamma < 2/C_0$ then for any $\theta$, the sequence $E[(\theta|S^\theta\theta)]$ is decreasing in $n$ and thus bounded from above by $M_\beta := E[(\theta|\theta)S^\theta\theta]$ uniformly in $n$.

(Proof: see Section 6.7.)

**Proposition 4** (Upper bound for the convergence of $\theta(n)$). Under (A), for any $0 < \kappa < \beta < \alpha(\theta)$, if $\gamma < 2/C_0$ then we have

\[
E[\langle \theta(n)|S^\kappa \theta(n) \rangle] \leq O(1)n^{-\beta-k}.
\]

**Proof.** For any $\kappa < \beta < \alpha$ we have with $p = \frac{\beta - \kappa}{\beta - \alpha} \in (0, 1)$ the convex combination $\kappa = p(\alpha - 1) + (1 - p)\beta$.

By Lemma 4 (Hölder inequality) we get $E(\theta|S_{=\kappa} \theta) \leq E(\theta|S_{=\gamma} \theta)^p E(\theta|S_{=\beta} \theta)^{1-p}$, from which it follows that $E(\theta|S_{=\kappa} \theta) \geq E(\theta|S_{=\gamma} \theta)^{1/p} E(\theta|S_{=\beta} \theta)^{1-1/p}$. We apply this to the sequence $\theta(n)$ and get

\[
E(\theta(n)|S_{=\kappa} \theta(n)) \geq E(\theta(n)|S_{=\gamma} \theta(n))^{1/p} E(\theta(n)|S_{=\beta} \theta(n))^{1-1/p} \geq E(\theta(n)|S_{=\kappa} \theta(n))^{1/p} M_\beta^{1-1/p},
\]

where in the last inequality we used Corollary 1. Setting $E(\theta(n)|S_{=\gamma} \theta(n)) = \phi(\theta), w = \frac{\beta - \alpha}{\beta - \gamma}$ and $K = M_\beta^{1-w}$ we get the recursion $\phi_{n+1}(k) \leq \phi_n(k) - K\phi_n(k)^{1+w}$. Now apply Lemma 3 and get $\phi_n(k) \leq O(1)n^{-1/w}$, where the constant $O(1)$ may depend on $\kappa$ and $\beta$, but not on $n$. \(\square\)
5. Almost sure convergence

Denote,

\[ h(z) := \mathbb{E}[(z|x)^2], \quad M_n := \left( \frac{\theta_n}{||\theta_n||} |x_n+1\right)^2 - h\left( \frac{\theta_n}{||\theta_n||} \right). \]

**Lemma 6 (Martingale).** Under the condition of \( \mathbb{E} [||x||^4] < \infty \), \( M_n \) is a martingale sequence.

**Proof.** By the definition of \( h(\cdot) \), we have,

\[ \mathbb{E} [M_n z(x_1, x_2, \ldots, x_{n-1})] = \mathbb{E} \left[ \left( \frac{\theta_{n-1}}{||\theta_{n-1}||} |x_n\right)^2 - h\left( \frac{\theta_{n-1}}{||\theta_{n-1}||} \right) z(x_1, x_2, \ldots, x_{n-1}) \right] = 0. \]

\[ \square \]

**Proof of Theorem 4 on almost sure convergence.** Recall that by (1), \( \theta_{n+1} = T_{x_{n+1}} \theta_n = \theta_n - \gamma S_{x_{n+1}} \theta_n \) and by Lemma 5 (use recursion formula (6) with \( \beta = 0 \)) we have:

\[ ||\theta_{n+1}||^2 = ||\theta_n||^2 - \gamma (2 - \gamma ||\theta_n||^2) (\theta_n |x_{n+1}||^2). \]

Let us consider first the boundedness condition, \( ||x||^2 \leq M \) a.s. Then for \( \gamma < 2/M \) the sequence \( ||\theta_n||^2 \) is decreasing. Hence, \( ||\theta_n|| \) converges almost surely, and as \( \theta_n \) converges to zero in probability, the result follows.

Now let us look at the more general case when \( \mathbb{E} [||x||^4] < \infty \). From the recursion formula we further have,

\[ ||\theta_{n+1}||^2 = ||\theta_n||^2 \left[ 1 - 2\gamma \left( \frac{\theta_n}{||\theta_n||} |x_{n+1}||^2 + \gamma^2 \left( \frac{\theta_n}{||\theta_n||} |x_{n+1}||^2 \right) \right]. \]

Hence, we can write,

\[ ||\theta_{N+1}||^2 = ||\theta_0||^2 \prod_{n=0}^{N} \left[ 1 - 2\gamma \left( \frac{\theta_n}{||\theta_n||} |x_{n+1}||^2 + \gamma^2 \left( \frac{\theta_n}{||\theta_n||} |x_{n+1}||^2 \right) \right]. \]

In order to prove that \( ||\theta(n)||^2 \to 0 \) almost surely, using

\[ \log \left\{ \prod_{n=0}^{N} \left[ 1 - 2\gamma \left( \frac{\theta_n}{||\theta_n||} |x_{n+1}||^2 + \gamma^2 \left( \frac{\theta_n}{||\theta_n||} |x_{n+1}||^2 \right) \right] \right\} \]

\[ \leq - \sum_{n=0}^{N} 2\gamma \left( \frac{\theta_n}{||\theta_n||} |x_{n+1}||^2 - \gamma^2 \left( \frac{\theta_n}{||\theta_n||} |x_{n+1}||^2 \right) \right), \]

it is enough to prove that

\[ \sum_{n=0}^{N} M_n + \sum_{n=0}^{N} \log \left( \frac{\theta_n}{||\theta_n||} \right) \geq \frac{\gamma}{2} \sum_{n=0}^{N} \left( \frac{\theta_n}{||\theta_n||} |x_{n+1}||^2 \right), \]

\[ \text{(7)} \]

tends to +\( \infty \) almost surely.

By Lemma 6, we know that \( \sum_{n=0}^{N} M_n \) is a martingale, and by Doob’s convergence theorem, it converges almost surely to a random variable with finite mean, see e.g. [5]. The second term in (7) is lower bounded by \( (N + 1)\delta \), where \( \delta = \inf_{||x|| = 1} h(z) > 0 \) by assumption. The third term is lower bounded by

\[ \frac{\gamma}{2} \sum_{n=0}^{N} ||x_{n+1}||^4 = -(N + 1)\frac{\gamma}{2} \frac{1}{N + 1} \sum_{n=0}^{N} ||x_{n+1}||^4. \]

By the assumption of the finiteness of the fourth moment, we know that, \( \frac{1}{N + 1} \sum_{n=0}^{N} ||x_{n+1}||^4 \) converges almost surely to \( \mathbb{E} [||x||^4] \). Hence, under the assumed condition that \( \gamma < \delta / \mathbb{E} [||x||^4] \), the desired result follows. \[ \square \]
6. Proofs of Technical Lemmata

Proof 6.1 (for the equality in Example 2).

\[ E \left[ (\theta x)^2 \varphi(x) \right] = E \left[ \left( \sum \theta_i x_i \right) \cdot \left( \sum \theta_i x_i \right) \cdot \left( \sum \lambda_j^{\beta} x_j^2 \right) \right] \]

\[ = \sum_{i,j} E \left[ \theta_i \theta_j \lambda_j^{\beta} x_i x_j x_i x_j \right] = \left( \sum_i + \sum_j + \sum_i + \sum_j \right) E \left[ \lambda_j^{\beta} x_j^2 \right] \]

which using independence, 0 mean and calculated moments is equal to

\[ = 0 + 0 + 0 + \sum_{j=1} E \left[ \theta_j^2 \lambda_j^{\beta} x_j^2 \right] + \sum_{j} E \left[ \lambda_j^{\beta} x_j^2 \right] \]

\[ = \sum_{i,j} E \left[ \theta_i^2 \lambda_i^{\beta} x_i^2 x_j^2 \right] = - \sum_{i} E \left[ \theta_i^2 x_i^2 \right] E \left[ \lambda_i^{\beta} x_i^2 \right] + \sum_{i} E \left[ \theta_i^2 \lambda_i^{\beta} x_i^2 \right] \]

\[ = \left( \sum_i \theta_i^2 \lambda_i \right) \cdot \left( \sum_i \lambda_i^{\beta} \right) - \sum_i \left( \theta_i^2 \lambda_i \right) \left( \lambda_i^{\beta} \right) + \sum_i \theta_i^2 \lambda_i^{\beta} (1 + \lambda_i) \lambda_i \]

\[ = \sum_i \theta_i^2 \lambda_i \cdot K + \sum_i \theta_i^2 \lambda_i^{\beta} \cdot . \]

Proof 6.2 (of Proposition 1). Apply \( \theta = e \) from the ON basis of \( S \) and get \( E [x_i^2] = \leq C_0 \lambda_i \). After summing up for \( i \leq N \) we get

\[ E \left[ \sum_{i \leq N} x_i^2 \right]^2 \leq C_0 \sum_{i \leq N} \lambda_i \]

Thus for any \( N \) we have \( E \left[ \sum_{i \leq N} x_i^2 \right] \leq C_0 \) and the Proposition follows by taking the limit \( \lim N \to \infty \).

Proof 6.3 (Lemma 1). The function \( f \) is continuous, and for \( \lambda > 0 \), \( \lambda \neq 1 \) we have: \( f'(\lambda) = f(\lambda) \cdot \frac{1}{1 - \lambda} \cdot (-m\lambda + \tau(1 - \lambda)) \). Then, as \( f(0) = f(1) = 0 \) and \( f > 0 \), and the only local maximum is possible at \( \lambda_\ast \) where the value is

\[ f(\lambda_\ast) = \left( 1 - \frac{\tau}{m + \tau} \right)^{\frac{m-\tau}{m + \tau}} \cdot \left( 1 - \frac{\tau}{m + \tau} \right)^{\frac{1}{\tau}} \cdot \left( \frac{\tau}{m + \tau} \right)^{\frac{1}{\tau}} \]

As \( 1 - x \leq e^{-x} \leq 1 - (1 - e^{-1})x \) for \( 0 \leq x \leq 1 \), we have, \( 1 - y \leq e^{(e-1)\tau} \) with \( y = (1 - e^{-1})x \). With \( z = \frac{m - \tau}{m + \tau} \) in place of \( x \) on one side we get \( (1 - z)^{\frac{m - \tau}{m + \tau}} \leq e^{-\tau} \) and with the same \( z \) in place of \( y \) on the other side we get \( (1 - z)^{\frac{m - \tau}{m + \tau}} \leq e^{-\tau} \). For \( 1 \leq \lambda < 2 - \epsilon \) we observe that \( f \) is increasing there and \( f(\lambda) \leq f(1 - \epsilon) \leq (1 - \epsilon)^{m-\tau} \) which, as \( m \to \infty \), decreases to 0 faster than \( m^{-\tau} \).

Proof 6.4 (Lemma 2). For \( q = -\ln(1 - \mu) \) with \( 0 < \mu < \frac{1}{4} \) we have \( q < (2 \ln 2) \mu \) so for the term of the series we have \( e^{\mu q}(\ln 4)^{-x} \leq \exp(-\ln(1 - \mu))(\ln 4)^{-x} < e^{\mu q}(\ln 4)^{-x} \), where the bounds can be tightened if we know the sign of \( \kappa - 1 \). Now we can estimate the series \( \sum_{n} e^{\mu q}(\ln 4)^{-x} \) by the integral \( \int_{0}^{\infty} e^{\mu q}(\ln 4)^{-x} \, d(qn) = \Gamma(x) \) (use the variable \( t = qn \)). If \( \kappa \leq 1 \) then the function to integrate is monotone and the comparison is standard. For \( \kappa > 1 \) the function has a maximum at \( \kappa - 1 \), and some care needs to be taken around this point. Luckily the values of the function for neighboring \( n \)'s are comparable:

\[ e^{\frac{\mu q(\kappa + 1)}{\ln 4}} \frac{e^{-n q}}{e^{-n q} - 1} = e^{-q} \left( 1 + \frac{1}{n} \right)^{\kappa - 1} e^{-q} \]

which is bounded from above and below for bounded \( q \) and \( \kappa \), even near the maximum \( qn \approx \kappa - 1 \). So that there exists \( K > 0 \), such that, for every \( n > 0 \),

\[ K \leq \frac{e^{-q n}(\ln 4)^{-x} q}{\int_{n-1}^{n} e^{q n}(\ln 4)^{-x} q} \leq \frac{1}{K} \]
Proof 6.5 (of Lemma 3). The sequence is decreasing and the only accumulation point is 0. Let \( a = b^{-1/w} \) with \( b > 1 \) then \( b_{n+1} \geq b_n(1 - 1/b_n)^w \geq b_n(1 + w/b_n) = b_n + w \) so that \( b_n \geq b_0 + nw \) and \( a_n \leq (a_0^{-w} + nw)^{-1/w} \). Use this next for \( a_n = K^{1/w} c_n \).

Proof 6.6 (of Lemma 5).

\[
E[\varphi_n(T_n, \theta)] = E[\langle \theta - \gamma S_n \theta | S^{\beta} \theta - \gamma S_n \theta \rangle] = E[\langle \theta | S^{\beta} \theta \rangle] - \gamma E[\langle \theta | S^{\beta} S_n \theta \rangle] - \gamma E[\langle S_n \theta | S^{\beta} \theta \rangle] + \gamma^2 E[\langle S_n \theta | S^{\beta} S_n \theta \rangle]
\]

\[
= \varphi_n(\theta) - \gamma \theta S^{\beta} \theta - \gamma \langle S_n \theta | S^{\beta} \theta \rangle + \gamma^2 E[\langle \theta | x | S^{\beta} \theta \rangle]
\]

\[
= \varphi_n(\theta) - \gamma \langle S^{\beta} S_n \theta \rangle - \gamma \langle S_n \theta | S^{\beta} \theta \rangle + \gamma^2 E[\langle \theta | x \rangle^2 \langle x | S^{\beta} \theta \rangle]
\]

where we used the definition of \( S = E[S_n] \), the symmetry (Lemma 1) and commutativity of the powers of \( S \), \( S^1 \circ S^{\beta} = S^{1+\beta} = S^{\beta} \circ S^1 \), whenever well defined.

Proof 6.7 (of Corollary 1). By (A) there exists a constant \( C_\beta \) such that for all \( \theta \in \mathbb{H} \) we have, \( E[\langle \theta | x \rangle^2 \langle \theta | S^{\beta} \theta \rangle] \leq C_\beta E[\langle \theta | S^{1+\beta} \theta \rangle] \). Hence by Lemma 5,

\[
E[\langle \theta(n+1) | S^{\beta} \theta(n+1) \rangle] \leq E[\langle \theta(n) | S^{\beta} \theta(n) \rangle] - 2\gamma E[\langle \theta(n) | S^{1+\beta} \theta(n) \rangle] + \gamma^2 C_\beta E[\langle \theta | S^{1+\beta} \theta \rangle]
\]

\[
= E[\langle \theta(n) | S^{\beta} \theta(n) \rangle] - \gamma(2 - \gamma C_\beta) E[\langle \theta(n) | S^{1+\beta} \theta(n) \rangle]
\]

and as \( E[\langle \theta | S^{1+\beta} \theta \rangle] > 0 \) for \( \gamma < 2/C_\beta \), the (positive) quantity \( E[\langle \theta(n) | S^{\beta} \theta(n) \rangle] \) is decreasing in \( n \), and therefore bounded from above for all \( n \) by \( \langle \theta | S^{\beta} \theta \rangle \).

References

[1] R. Berthier, F. R. Bach, P. Gaillard, Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model, Advances in Neural Information Processing Systems (NeurIPS).

[2] M. Reed, B. Simon, I. Functional Analysis, Methods of Modern Mathematical Physics, Elsevier Science, 1981. URL https://books.google.com/books?id=rpFTTjxUFpsC

[3] S. Sheffield, Gaussian free fields for mathematicians, Probability Theory and Related Fields 139 (3) (2007) 521–541. doi:10.1007/s00440-006-0050-1. URL https://doi.org/10.1007/s00440-006-0050-1

[4] M. Ledoux, M. Talagrand, Probability in Banach Spaces: Isoperimetry and Processes, A Series of Modern Surveys in Mathematics Series, Springer, 1991. URL https://books.google.com/books?id=cyFYDfvrXjgC

[5] P. Hall, C. Heyde, Z. Birnbaum, E. Lukacs, Martingale Limit Theory and Its Application, Communication and Behavior, Elsevier Science, 2014. URL https://books.google.com/books?id=gqr16QAAQBAJ