Journal of Machine Learning Research 16 (2015) 3341-3365 Submitted 6/14; Revised 4/15; Published 12/15

Learning Theory of Randomized Kaczmarz Algorithm

Junhong Lin JHLIN5@HOTMAIL.COM
Ding-Xuan Zhou MAZHOU@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
83 Tat Chee Avenue
Kowloon, Hong Kong, China

Editor: Gabor Lugosi

Abstract

A relaxed randomized Kaczmarz algorithm is investigated in a least squares regression setting by a learning theory approach. When the sampling values are accurate and the regression function (conditional means) is linear, such an algorithm has been well studied in the community of non-uniform sampling. In this paper, we are mainly interested in the different case of either noisy random measurements or a nonlinear regression function. In this case, we show that relaxation is needed. A necessary and sufficient condition on the sequence of relaxation parameters or step sizes for the convergence of the algorithm in expectation is presented. Moreover, polynomial rates of convergence, both in expectation and in probability, are provided explicitly. As a result, the almost sure convergence of the algorithm is proved by applying the Borel-Cantelli Lemma.

Keywords: learning theory, relaxed randomized Kaczmarz algorithm, online learning, space of homogeneous linear functions, almost sure convergence

1. Introduction

The Kaczmarz method is an iterative projection algorithm. It was originally proposed for solving (overdetermined) systems of linear equations, and has been adapted to image reconstruction, signal processing and numerous other applications.

Given a matrix $A \in \mathbb{R}^{m \times d}$ and a vector $b \in \mathbb{R}^m$, the classical Kaczmarz algorithm (Kaczmarz, 1937) approximates a solution of the linear systems $Ax = b$ by an iterative scheme as

$$x_{k+1} = x_k + \frac{b_i - \langle a_i, x_k \rangle}{\|a_i\|^2} a_i, \tag{1}$$

where $i = k \mod m$, a_i^T is the i-th row of the matrix A, and $x_1 \in \mathbb{R}^d$ is an initial vector. Here $\langle \ , \ \rangle$ is the inner product in $\mathbb{R}^d$ and $\| \cdot \|$ the induced norm.

The convergence of the Kaczmarz algorithm (2) is well understood (Kaczmarz, 1937), and its convergence rate depends on the order of rows of A. To avoid this dependence, a randomized Kaczmarz algorithm was considered in (Strohmer and Vershynin, 2009) by setting the probability of a row to be proportional to its norm. It takes the form

$$x_{k+1} = x_k + \frac{b_{p(i)} - \langle a_{p(i)}, x_k \rangle}{\|a_{p(i)}\|^2} a_{p(i)}, \tag{2}$$

where $p(i)$ takes values in $\{1, \ldots, m\}$ with probability $\frac{\|a_{p(i)}\|^2}{\|A\|_F^2}$ with $\|A\|_F^2 = \sum_{i=1}^m \sum_{j=1}^d a_{ij}^2$ being the Frobenius norm square of A. Exponential convergence rate was proved for the expected error $\mathbb{E}\|x_{k+1} - x\|^2$ of the randomized Kaczmarz algorithm (2) in (Strohmer and Vershynin, 2009). When noise exists in the sample value $b = Ax + \xi$ with ξ being a noise vector, a bound for the expected error was obtained in (Needell, 2010) and divergence was proved when the variance of ξ is positive. The error bound consists of an exponentially convergent part and a noise-driven term proportional to the noise level $\max_i \frac{|\xi_i|}{\|a_i\|^2}$.

The randomized Kaczmarz algorithm (2) was generalized in Chen and Powell (2012) to a setting with a sequence of independent random measurement vectors $\{\varphi_t \in \mathbb{R}^d\}_t$ as

$$x_{k+1} = x_k + \frac{y_k - \langle \varphi_k, x_k \rangle}{\|\varphi_k\|^2} \varphi_k. \tag{3}$$

When the measurements have no noise $y_k = \langle \varphi_k, x \rangle$, almost sure convergence was proved and quantitative error bounds were provided in (Chen and Powell, 2012).

When the linear system $Ax = b$ is overdetermined ($m > d$) and has no solution, the Kaczmarz algorithm (2) can be modified by introducing a relaxation parameter $\eta_k > 0$ in front of $\frac{b_i - \langle a_i, x_k \rangle}{\|a_i\|^2} a_i$ and the output sequence $\{x_k\}$ converges to the least squares solution $\arg\min_{x \in \mathbb{R}^d} \|Ax - b\|^2$ when $\lim_{k \to \infty} \eta_k = 0$. See, e.g., (Zouzias and Freris, 2013) and references therein.

Setting $\psi_k = \frac{1}{\|\varphi_k\|} \varphi_k \in \mathbb{S}^{d-1}$ and $\widetilde{y}_k = \frac{1}{\|\varphi_k\|} y_k$ yields an equivalent form of the scheme (3) as

$$x_{k+1} = x_k + \{\widetilde{y}_k - \langle \psi_k, x_k \rangle\} \psi_k.$$

This form is similar to those in the literature of online learning for least squares regression and together with the relaxed Kaczmarz method (Zouzias and Freris, 2013) motivates us to consider the following relaxed randomized Kaczmarz algorithm.

Definition 1 *With normalized measurement vectors $\{\psi_t \in \mathbb{S}^{d-1}\}_t$ and sample values $\{\widetilde{y}_t \in \mathbb{R}\}_t$, the relaxed randomized Kaczmarz algorithm is defined by*

$$x_{t+1} = x_t + \eta_t \{\widetilde{y}_t - \langle \psi_t, x_t \rangle\} \psi_t, \qquad t = 1, \ldots, \tag{4}$$

where $x_1 \in \mathbb{R}^d$ is an initial vector and $\{\eta_t\}$ is a sequence of relaxation parameters or step sizes.

The purpose of this paper is to provide learning theory analysis for the relaxed randomized Kaczmarz algorithm. We shall assume throughout the paper that $0 < \eta_t \leq 2$ for each $t \in \mathbb{N}$ and that the sequence $\{z_t := (\psi_t, \widetilde{y}_t)\}_{t \in \mathbb{N}}$ is independently drawn according to a Borel probability measure ρ on $Z := \mathbb{S}^{d-1} \times \mathbb{R}$ which satisfies $\mathbb{E}[|\widetilde{y}|^2] < \infty$.

Our first goal is to deal with the noisy setting for the randomized Kaczmarz algorithm. When the sampling process is noisy or nonlinear (to be defined below), we show that $\{x_t\}_t$ converges to some $x^* \in \mathbb{R}^d$ in expectation if and only if $\lim_{t \to \infty} \eta_t = 0$ and $\sum_{t=1}^\infty \eta_t = \infty$. Moreover, the rate of convergence in expectation cannot be too fast. It tells us that the relaxation parameter is necessary for the convergence in the noisy setting. When $\{\eta_t\}_t$ takes the form $\eta_t = \eta_1 t^{-\theta}$, we provide convergence rates in expectation and in confidence and prove the almost sure convergence. Such results were presented in the case of no noise in (Strohmer and Vershynin, 2009; Chen and Powell, 2012) and are new in the noisy setting.

Our second goal is to give the first almost sure convergence result in online learning for least squares regression when regularization is not needed. Such a result can be found in (Tarrés and Yao, 2014) when regularization is imposed, while the convergence in expectation without regularization was proved in (Ying and Pontil, 2008). We also present the first consistency result for online learning when the approximation error (to be defined below) does not tend to zero.

2. Main Results

To introduce our learning theory approach to the relaxed randomized Kaczmarz algorithm (4), we decompose the probability measure ρ on $Z = \mathbb{S}^{d-1} \times \mathbb{R}$ into its marginal distribution ρ_X on $X := \mathbb{S}^{d-1}$ and conditional distributions $\rho(\cdot|\psi)$ at $\psi \in X$. The conditional means define the *regression function* $f_\rho : X \to \mathbb{R}$ as

$$f_\rho(\psi) = \int_{\mathbb{R}} \widetilde{y} d\rho(\widetilde{y}|\psi), \qquad \psi \in X. \tag{5}$$

The hypothesis space for the Kaczmarz algorithm (4) consists of homogeneous linear functions

$$\mathcal{H} = \left\{ f_x \in L^2_{\rho_X} : \ x \in \mathbb{R}^d \right\}, \qquad \text{where } f_x(\psi) := \langle x, \psi \rangle, \quad \psi \in X. \tag{6}$$

Definition 2 *The sampling process associated with ρ is said to be noise-free if $\widetilde{y} = f_\rho(\psi)$ almost surely. Otherwise, it is called noisy. It is said to be linear if $f_\rho \in \mathcal{H}$ as a function in $L^2_{\rho_X}$. Otherwise, it is called nonlinear.*

The main difference between our analysis in this paper and that in the literature (Strohmer and Vershynin, 2009; Needell, 2010; Chen and Powell, 2012) lies in the setting when the sampling process is either noisy or nonlinear. These two situations can be handled simultaneously by means of the least squares generalization error $\mathcal{E}(f) = \int_Z (\widetilde{y} - f(\psi))^2 d\rho$, a well developed concept in learning theory. The assumption $\mathbb{E}[|\widetilde{y}|^2] < \infty$ on ρ ensures $f_\rho \in L^2_{\rho_X}$ and $\mathcal{E}(f_\rho) < \infty$. The noise-free condition can be stated as $\mathcal{E}(f_\rho) = 0$.

It is well known that the regression function minimizes $\mathcal{E}(f)$ among all the square integral (with respect to ρ_X) functions $f \in L^2_{\rho_X}$, and satisfies

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_X}} = \int_X (f(\psi) - f_\rho(\psi))^2 d\rho_X. \tag{7}$$

Since the hypothesis space $\mathcal{H}$ is a finite dimensional subspace of $L^2_{\rho_X}$, the continuous functional $\mathcal{E}(f)$ achieves a minimizer

$$f_{\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}(f). \tag{8}$$

From (7) we see that $f_{\mathcal{H}}$ is the best approximation of f_ρ in the subspace $\mathcal{H}$. It is unique as the orthogonal projection of f_ρ onto $\mathcal{H}$. It can be written as $f_{\mathcal{H}} = f_{x^*}$ for some $x^* \in \mathbb{R}^d$. But such a vector x^* is not necessarily unique.

The linear condition can be stated as $f_\rho = f_{\mathcal{H}}$ or $f_\rho \in \mathcal{H}$ as functions in $L^2_{\rho_X}$. So we see that the sampling process is noisy or nonlinear if and only if $\mathcal{E}(f_{\mathcal{H}}) > 0$. Now we can state our first main result, to be proved in Section 4, which gives a characterization of the convergence of $\{x_t\}_t$ to some $x^* \in \mathbb{R}^d$ in expectation.

Theorem 3 *Define the sequence $\{x_t\}_t$ by (4). Assume $\mathcal{E}(f_{\mathcal{H}}) > 0$. Then we have the limit $\lim_{T\to\infty} \mathbb{E}_{z_1,\ldots,z_T}\|x_{T+1} - x^*\|^2 = 0$ for some $x^* \in \mathbb{R}^d$ if and only if*

$$\lim_{t\to\infty} \eta_t = 0 \quad \textit{and} \quad \sum_{t=1}^{\infty} \eta_t = \infty. \tag{9}$$

In this case, we have

$$\sum_{T=1}^{\infty} \sqrt{\mathbb{E}_{z_1,\ldots,z_T}\|x_{T+1} - x^*\|^2} = \infty. \tag{10}$$

Compared with the result on exponential convergence in expectation in the linear case without noise (Strohmer and Vershynin, 2009), the somewhat negative result (10) tells us that in the noisy setting the convergence in expectation cannot be as fast as $\mathbb{E}_{z_1,\ldots,z_T}\|x_{T+1} - x^*\|^2 \neq O(T^{-\theta})$ for any $\theta > 2$. But for $\theta < 1$, such learning rates can be achieved by taking $\eta_t = \eta_1 t^{-\theta}$, as shown in the following second main result, to be proved in Section 4.

Theorem 4 *Let $\eta_t = \eta_1 t^{-\theta}$ for some $\theta \in (0, 1]$ and $\eta_1 \in (0, 1)$. Define the sequence $\{x_t\}_t$ by (4). Then for some $x^* \in \mathbb{R}^d$ we have*

$$\mathbb{E}_{z_1,\ldots,z_T}\|x_{T+1} - x^*\|^2 \leq \begin{cases} \widetilde{C}_0 T^{-\theta}, & \textit{if } \theta < 1, \\ \widetilde{C}_0 T^{-\lambda_r \eta_1}, & \textit{if } \theta = 1, \end{cases} \tag{11}$$

where $\widetilde{C}_0$ is a constant independent of $T \in \mathbb{N}$ (given explicitly in the proof) and λ_r is the smallest positive eigenvalue of the covariance matrix C_{ρ_X} of the probability measure ρ_X defined by

$$C_{\rho_X} = \mathbb{E}_{\rho_X}[\psi\psi^T] = \int_X \psi\psi^T d\rho_X. \tag{12}$$

Our third main result is the following confidence-based estimate for the error which will be proved in Section 5.

Theorem 5 *Assume that for some constant $M > 0$, $|\widetilde{y}| \leq M$ almost surely. Let $\theta \in [1/2, 1]$, $\eta_t = \eta_1 t^{-\theta}$ with $0 < \eta_1 < \min\{1, \frac{1}{2\lambda_r}\}$, and $2 \leq T \in \mathbb{N}$. Then for some $x^* \in \mathbb{R}^d$ and for any $0 < \delta < 1$, with confidence at least $1 - \delta$ we have*

$$\|x_{T+1} - x^*\| \leq \begin{cases} \widetilde{C}_1 T^{-\theta/2} \left(\log \frac{4}{\delta}\right)^2 \log T, & \textit{when } \theta \in [1/2, 1), \\ \widetilde{C}_1 T^{-\lambda_r \eta_1} \log \frac{2}{\delta} \sqrt{\log T}, & \textit{when } \theta = 1, \end{cases} \tag{13}$$

where $\widetilde{C}_1$ is a positive constant independent of T or δ (given explicitly in the proof).

Our last main result is about the almost sure convergence of the algorithm, which will be proved in Section 6.

Theorem 6 *Under the assumptions of Theorem 5, we have for any $\epsilon \in (0,1]$, the following holds for some $x^* \in \mathbb{R}^d$:*

(A) When $1/2 \leq \theta < 1$, $\lim_{t\to\infty} t^{\theta(1-\epsilon)/2}\|x_{t+1} - x^*\| = 0$ almost surely.

(B) When $\theta = 1$, $\lim_{t\to\infty} t^{\lambda_r \eta_1(1-\epsilon)}\|x_{t+1} - x^*\| = 0$ almost surely.

Let us demonstrate our setting by two examples without noise considered in the literature. The first example appeared in (Chen and Powell, 2012).

Example 1 *If random measurement vectors $\{\varphi_t\}_{t=1}^{\infty}$ are independent and nonzero almost surely, then $\{\psi_k = \frac{1}{\|\varphi_k\|}\varphi_k \in \mathbb{S}^{d-1}\}$ are independent.*

The second example is from (Strohmer and Vershynin, 2009).

Example 2 *Define the random vector φ which is a normalized row of a full rank matrix $A \in \mathbb{R}^{m\times d}$, with probabilities as*

$$\varphi = \frac{a_j}{\|a_j\|} \quad \textit{with probability} \quad \frac{\|a_j\|^2}{\|A\|_F^2} \quad j = 1, \cdots, m.$$

It was shown in Strohmer and Vershynin (2009) that the smallest eigenvalue of the covariance matrix is positive:

$$\lambda_{\min}(\mathbb{E}[\varphi\varphi^T]) \geq \frac{1}{\|A\|_F^2\|A^{-1}\|^2}.$$

It means $r = d$ and $\lambda_r \geq \frac{1}{\|A\|_F^2\|A^{-1}\|^2}$.

The third example is on homoskedastic models (Johnston, 1963).

Example 3 *In the literature of homoskedastic models, it is assumed that the sample value $\{y_t\}_t$ satisfies $y_t = \langle x^*, \psi_t\rangle + \xi_t$ with $\{\xi_t\}_t$ being independently drawn according to a zero mean probability measure ξ. This corresponds to the special case when the conditional distributions $\rho(\cdot|\psi)$ are given by $\rho(\cdot|\psi) = f_\rho(\psi) + \xi$. Our setting induced by ρ is more general and allows heteroskedastic models.*

3. Connections to Learning Theory

The relaxed randomized Kaczmarz algorithm defined by (4) may be rewritten as an online learning algorithm with output functions from the hypothesis space (6), and our main results stated in the last section are new even in the online learning literature. To demonstrate this, we denote the tth output function F_t on X induced by the vector x_t to be given by $F_t(\psi) = \langle x_t, \psi\rangle$ for $\psi \in X$. Then the iteration relation (4) gives

$$F_{t+1} = F_t + \eta_t \left\{\widetilde{y}_t - F_t(\psi_t)\right\} \langle \cdot, \psi_t\rangle. \tag{14}$$

This is a special kernel-based least squares online learning algorithm. Here a (Mercer) kernel on a metric space $\mathcal{X}$ means a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is continuous, symmetric and the matrix $(K(x_i, x_j))_{i,j=1}^{\ell}$ is positive semidefinite for any finite subset $\{x_i\}_{i=1}^{\ell} \subseteq \mathcal{X}$.

It generates a reproducing kernel Hilbert space $(\mathcal{H}_K, \|\cdot\|_K)$ by the set of fundamental functions $\{K(\cdot,x) : x \in \mathcal{X}\}$ with the inner product $\langle K(\cdot,x), K(\cdot,y)\rangle_K = K(x,y)$. A least squares regularized online learning algorithm in $\mathcal{H}_K$ is defined with $\{(\psi_t, \widetilde{y}_t) \in \mathcal{X} \times \mathbb{R}\}_t$ drawn independently according to a probability measure on $Z = \mathcal{X} \times \mathbb{R}$ as

$$F_{t+1} = F_t - \eta_t \left\{ \left(F_t(\psi_t) - \widetilde{y}_t\right) K(\cdot, \psi_t) + \lambda F_t \right\}, \qquad t = 1, \ldots, \tag{15}$$

where $\lambda \geq 0$ is a regularization parameter. The consistency of the online learning algorithm (15) is well understood when the approximation error $\mathcal{D}(\lambda)$ tends to zero as $\lambda \to 0$.

Definition 7 *The approximation error (or regularization error) of the pair (ρ, K) is defined for $\lambda > 0$ as*

$$\mathcal{D}(\lambda) = \inf_{f \in \mathcal{H}_K} \left\{ \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda \|f\|_K^2 \right\} = \inf_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|_{L^2_{\rho_\mathcal{X}}}^2 + \lambda \|f\|_K^2 \right\}. \tag{16}$$

When $\lambda > 0$ (with regularization) and $\lim_{\lambda \to 0} \mathcal{D}(\lambda) = 0$, the error $\|F_{T+1} - f_\rho\|_{L^2_{\rho_\mathcal{X}}}^2$ in expectation and in confidence was bounded in (Smale and Yao, 2005; Ying and Zhou, 2006; Smale and Zhou, 2009; Tarrés and Yao, 2014) by means of the decay of $\mathcal{D}(\lambda)$ and T. The error analysis was done in Ying and Pontil (2008) without regularization ($\lambda = 0$) but under the approximation error condition $\lim_{\lambda \to 0} \mathcal{D}(\lambda) = 0$. The error $\|F_{T+1} - f_\rho\|_K^2$ with the $\mathcal{H}_K$-metric was also analyzed when $f_\rho \in \mathcal{H}_K$.

If we take the kernel to be the linear one: $K(x,y) = \langle x, y\rangle$ with $\mathcal{X} = \mathbb{R}^d$, and assume that the marginal distribution $\rho_\mathcal{X}$ is supported on $X = \mathbb{S}^{d-1}$, then $\psi_t \in \mathbb{S}^{d-1}$ almost surely. Set $\lambda = 0$, we see that the relaxed randomized Kaczmarz algorithm expressed in the form (14) is the least squares online learning algorithm (15) without regularization. So the error analysis from Ying and Pontil (2008) applies, but the condition $\lim_{\lambda \to 0} \mathcal{D}(\lambda) = 0$ is required for the consistency in $L^2_{\rho_\mathcal{X}}$ and even stronger conditions (stronger than $f_\rho \in \mathcal{H}_K$) are needed for the consistency in the $\mathcal{H}_K$-metric.

Notice that for the linear kernel, $\|x\| = \|\langle \cdot, x\rangle\|_K$. So the error analysis carried out in this paper provides bounds for the error $\|F_{T+1} - f_\mathcal{H}\|_K$ without the condition $\lim_{\lambda \to 0} \mathcal{D}(\lambda) = 0$. Such results cannot be found in the literature of online learning. It leads to the problem of carrying our similar error analysis for more general online learning algorithms associated with more general kernels. Moreover, the best convergence rate in expectation of the general kernel-based least squares online learning algorithm is $O(T^{-1/2})$ in the literature (Smale and Yao, 2005; Ying and Zhou, 2006; Smale and Zhou, 2009; Tarrés and Yao, 2014; Hu et al., 2015). Theorem 4 demonstrates that the special online learning algorithm (4) has convergence rates of type $O(T^{-(1-\epsilon)})$ for any $\epsilon > 0$ and even of type $O(T^{-1} \log T)$ shown in Theorem 8 below, which is a great improvement.

Note that there is a gap between the negative result (10) and the positive one (11), which leads to the natural question whether learning rates of type $\mathbb{E}_{z_1,\ldots,z_T} \|x_{T+1} - x^*\|^2 = O(T^{-\theta})$ are possible for $1 < \theta \leq 2$. We conjecture that this is impossible for a general probability measure ρ, but a noise condition might help. The case $\theta = 1$ with a slight logarithmic modification $O(T^{-1} \log T)$ can be achieved by imposing a minor restriction on the step size in the following theorem which will be proved in the next section. The authors thank Dr. Yiming Ying for pointing out this result.

Theorem 8 *Let λ_r be as in Theorem 4 and $\eta_t = \frac{1}{\lambda_r(t+t_0)}$ for some $t_0 \in \mathbb{N}$ such that $t_0\lambda_r \geq 1$. Define the sequence $\{x_t\}_t$ by (4). Then for some $x^* \in \mathbb{R}^d$, we have*

$$\mathbb{E}_{z_1,\cdots,z_T}\|x_{T+1} - x^*\|^2 \leq \widetilde{C}_3(T+t_0)^{-1}\log T,$$

where $\widetilde{C}_3$ is a constant independent of $T \in \mathbb{N}$ (given explicitly in the proof).

4. Convergence in Expectation

In this section we prove our main results on convergence in expectation. To this end, we need some preliminary analysis.

Recall the function $f_{\mathcal{H}}$ defined by (8). It equals f_{x^*} for some $x^* \in \mathbb{R}^d$. As the orthogonal projection of f_ρ onto the finite dimensional subspace $\mathcal{H}$ in the Hilbert space $L^2_{\rho_X}$, it satisfies

$$\langle f_\rho - f_{x^*}, f_x\rangle_{L^2_{\rho_X}} = \int_X \left(f_\rho(\psi) - \langle x^*, \psi\rangle\right)\langle x, \psi\rangle d\rho_X(\psi) = 0, \quad \forall x \in \mathbb{R}^d. \tag{17}$$

The vector x^* is not necessarily unique. To see this, we use the covariance matrix C_{ρ_X} of the measure ρ_X defined by (12) and denote its eigenvalues to be $\lambda_1 \geq \ldots \geq \lambda_r > \lambda_{r+1} = \ldots = \lambda_d = 0$ where $r \in \{1, \ldots, d\}$ is the rank of C_{ρ_X}. Denote the eigenspace of C_{ρ_X} associated with the eigenvalue 0 as V_0 and the orthogonal projection onto V_0 as P_0. Then any vector $x^* + v$ from the set $x^* + V_0$ is also a minimizer of $\mathcal{E}(f_x)$ in $\mathbb{R}^d$, but $f_{x^*+v} = f_{x^*} = f_{\mathcal{H}}$ as functions in the space $L^2_{\rho_X}$.

The following lemma about the residual vectors $\{r_t = x_t - x^*\}_t$ is a crucial step in our analysis in this section.

Lemma 9 *Define the sequence $\{x_t\}_t$ by (4). Let $x^* \in \mathbb{R}^d$ be such that $f_{x^*} = f_{\mathcal{H}}$. Denote $r_t = x_t - x^*$. Then there holds*

$$\mathbb{E}_{z_t}[\|r_{t+1}\|^2] = \|r_t\|^2 + (-2\eta_t + \eta_t^2)\|f_{r_t}\|^2_{L^2_{\rho_X}} + \eta_t^2\mathcal{E}(f_{\mathcal{H}}), \qquad \forall\ t \in \mathbb{N}. \tag{18}$$

Proof Subtract x^* from both sides of (4) and take inner products. We see from $\|\psi_t\| = 1$ that

$$\|r_{t+1}\|^2 = \|r_t\|^2 + 2\eta_t\left\{\widetilde{y}_t - \langle\psi_t, x_t\rangle\right\}\langle\psi_t, r_t\rangle + \eta_t^2\left\{\widetilde{y}_t - \langle\psi_t, x_t\rangle\right\}^2. \tag{19}$$

Since x_t does not depend on z_t, taking expectation with respect to z_t, we see from $\mathbb{E}[\widetilde{y}_t|\psi_t] = f_\rho(\psi_t)$ and $\mathbb{E}_{z_t}\left\{\widetilde{y}_t - \langle\psi_t, x_t\rangle\right\}^2 = \mathcal{E}(f_{x_t})$ that

$$\mathbb{E}_{z_t}[\|r_{t+1}\|^2] = \|r_t\|^2 + 2\eta_t\mathbb{E}_{\psi_t}\left[\left\{f_\rho(\psi_t) - \langle\psi_t, x_t\rangle\right\}\langle\psi_t, r_t\rangle\right] + \eta_t^2\mathcal{E}(f_{x_t}).$$

By (17), we know that the middle term above equals

$$2\eta_t\mathbb{E}_{\psi_t}\left[\left\{\langle\psi_t, x^*\rangle - \langle\psi_t, x_t\rangle\right\}\langle\psi_t, r_t\rangle\right] = 2\eta_t\mathbb{E}_{\psi_t}\left[\left\{\langle\psi_t, -r_t\rangle\right\}\langle\psi_t, r_t\rangle\right] = -2\eta_t\|f_{r_t}\|^2_{L^2_{\rho_X}}.$$

Since f_{x^*} is the orthogonal projection of f_ρ onto $\mathcal{H}$, there holds $\mathcal{E}(f_{x_t}) = \mathcal{E}(f_\rho) + \|f_\rho - f_{x^*}\|^2_{L^2_{\rho_X}} + \|f_{x^*} - f_{x_t}\|^2_{L^2_{\rho_X}} = \mathcal{E}(f_{\mathcal{H}}) + \|f_{r_t}\|^2_{L^2_{\rho_X}}$. Then the desired identity (18) follows. ■

We are in a position to prove our first main result.

Proof of Theorem 3 *Necessity.* We first analyze the first two terms of the right hand side of the identity (18) in Lemma 9. Since $0 < \eta_t \leq 2$, we have $-2\eta_t + \eta_t^2 < 0$. Observe from the Schwarz inequality that $|f_{r_t}(\psi)|^2 = |\langle r_t, \psi \rangle|^2 \leq \|r_t\|^2 \|\psi\|^2 = \|r_t\|^2$ and thereby $\|f_{r_t}\|^2_{L^2_{\rho_X}} \leq \|r_t\|^2$. It follows that

$$\|r_t\|^2 + (-2\eta_t + \eta_t^2)\|f_{r_t}\|^2_{L^2_{\rho_X}} \geq \|r_t\|^2 + (-2\eta_t + \eta_t^2)\|r_t\|^2 = (1 - \eta_t)^2 \|r_t\|^2.$$

This together with (18) implies

$$\mathbb{E}_{z_t}[\|r_{t+1}\|^2] \geq (1 - \eta_t)^2 \|r_t\|^2 + \eta_t^2 \mathcal{E}(f_{\mathcal{H}}). \tag{20}$$

Then we can proceed with proving the necessity. If $\lim_{T\to\infty} \mathbb{E}_{z_1,\ldots,z_T} \|x_{T+1} - x^*\|^2 = 0$ for some $x^* \in \mathbb{R}^d$ and $\mathcal{E}(f_{\mathcal{H}}) > 0$, we know from (20) that $\lim_{T\to\infty} \eta_T = 0$. It ensures the existence of some integer $t_0 \geq 2$ such that $\eta_t \leq \frac{1}{3}$ for any $t \geq t_0$. Since $1 - \eta \geq \exp\{-2\eta\}$ for $0 < \eta \leq \frac{1}{3}$, we know that for any $t \geq t_0$, $(1 - \eta_t)^2 \geq \exp\{-4\eta_t\}$. Combining this with (20) yields

$$\mathbb{E}_{z_1,\ldots,z_T} \|x_{T+1} - x^*\|^2 \geq \Pi_{t=t_0}^T \exp\{-4\eta_t\} \, \mathbb{E}_{z_1,\ldots,z_{t_0-1}} \|r_{t_0}\|^2.$$

But (20) also tells us that $\mathbb{E}_{z_1,\ldots,z_{t_0-1}} \|r_{t_0}\|^2 \geq \eta_{t_0-1}^2 \mathcal{E}(f_{\mathcal{H}}) > 0$. So

$$\mathbb{E}_{z_1,\ldots,z_T} \|x_{T+1} - x^*\|^2 \geq \exp\left\{-4 \sum_{t=t_0}^T \eta_t \right\} \eta_{t_0-1}^2 \mathcal{E}(f_{\mathcal{H}}).$$

Since $\lim_{T\to\infty} \mathbb{E}_{z_1,\ldots,z_T} \|x_{T+1} - x^*\|^2 = 0$, we must have $\sum_{t=1}^\infty \eta_t = \infty$. This proves the necessity.

Sufficiency. Recall that V_0 is the eigenspace of the covariance matrix C_{ρ_X} associated with the eigenvalue 0 and P_0 is the orthogonal projection onto V_0. Then ψ_t is orthogonal to V_0 almost surely for each t. It follows that $P_0(x_{t+1}) = P_0(x_t)$ and thereby $P_0(x_t) = P_0(x_1)$ for each t. Take the vector x^* to be the minimizer of $\mathcal{E}(f_x)$ in $\mathbb{R}^d$ such that $P_0(x^*) = P_0(x_1)$. With this choice, r_t is orthogonal to V_0 for each t, and belongs to the orthogonal complement $V_0^\perp$. Note that the eigenvalues of C_{ρ_X} restricted to the subspace $V_0^\perp$ is at least $\lambda_r > 0$. So we have

$$\|f_{r_t}\|^2_{L^2_{\rho_X}} = \int_X |\langle \psi, r_t \rangle|^2 \, d\rho_X = \int_X r_t^T \psi \psi^T r_t d\rho_X = r_t^T C_{\rho_X} r_t \tag{21}$$

and $\|f_{r_t}\|^2_{L^2_{\rho_X}} \geq \lambda_r \|r_t\|^2$. The condition $\lim_{t\to\infty} \eta_t = 0$ ensures the existence of some $t_1 \in \mathbb{N}$ such that $\eta_t \leq 1$ for any $t \geq t_1$. Thus, we see from (18) in Lemma 9 that for $t \geq t_1$,

$$\mathbb{E}_{z_t}[\|r_{t+1}\|^2] \leq \|r_t\|^2 - \eta_t \|f_{r_t}\|^2_{L^2_{\rho_X}} + \eta_t^2 \mathcal{E}(f_{\mathcal{H}}) \leq (1 - \eta_t \lambda_r)\|r_t\|^2 + \eta_t^2 \mathcal{E}(f_{\mathcal{H}}).$$

Applying this inequality iteratively for $t = T, \cdots t_1$ yields

$$\mathbb{E}_{z_1,\ldots,z_T}[\|r_{T+1}\|^2] \leq \mathbb{E}_{z_1,\ldots,z_{t_1-1}}[\|r_{t_1}\|^2] \prod_{t=t_1}^T (1 - \eta_t \lambda_r) + \mathcal{E}(f_{\mathcal{H}}) \sum_{t=t_1}^T \eta_t^2 \prod_{k=t+1}^T (1 - \eta_k \lambda_r), \tag{22}$$

where we denote $\prod_{k=t+1}^{T}(1-\eta_k\lambda_r)=1$ for $t=T$. By the condition $\sum_{t=1}^{\infty}\eta_t=\infty$, one has

$$\prod_{t=t_1}^{T}(1-\eta_t\lambda_r)\leq \exp\left\{-\lambda_r\sum_{t=t_1}^{T}\eta_t\right\}\to 0 \quad \text{as } T\to\infty.$$

Thus for any $\varepsilon>0$, there exists $t_2=t_2(\varepsilon)\in\mathbb{N}$ such that for any $T\geq t_2$,

$$\mathbb{E}_{z_1,\dots,z_{t_1-1}}[\|r_{t_1}\|^2]\prod_{t=t_1}^{T}(1-\eta_t\lambda_r)\leq\varepsilon.$$

To deal with the other term of the bound (22) for $\|r_{T+1}\|^2$, we use the assumption $\lim_{t\to\infty}\eta_t=0$, and find some integer $t(\varepsilon)\geq t_1$ such that $\eta_t\leq\lambda_r\varepsilon$ for any $t\geq t(\varepsilon)$. Write

$$\sum_{t=t_1}^{T}\eta_t^2\prod_{k=t+1}^{T}(1-\eta_k\lambda_r)=\sum_{t=t_1}^{t(\varepsilon)}\eta_t^2\prod_{k=t+1}^{T}(1-\eta_k\lambda_r)+\sum_{t=t(\varepsilon)+1}^{T}\eta_t^2\prod_{k=t+1}^{T}(1-\eta_k\lambda_r). \tag{23}$$

The second term of (23) can be bounded as

$$\begin{aligned}\sum_{t=t(\varepsilon)+1}^{T}\eta_t^2\prod_{k=t+1}^{T}(1-\eta_k\lambda_r) &= \varepsilon\sum_{t=t(\varepsilon)+1}^{T}\eta_t\lambda_r\prod_{k=t+1}^{T}(1-\eta_k\lambda_r)\\ &= \varepsilon\sum_{t=t(\varepsilon)+1}^{T}(1-(1-\eta_t\lambda_r))\prod_{k=t+1}^{T}(1-\eta_k\lambda_r)\\ &= \varepsilon\left(1-\prod_{k=t(\varepsilon)+1}^{T}(1-\eta_k\lambda_r)\right)\leq\varepsilon.\end{aligned}$$

To bound the first term of (23), we apply the condition $\sum_{t=1}^{\infty}\eta_t=\infty$ again and find some integer $t_3=t_3(\varepsilon)>t(\varepsilon)$ such that $\sum_{k=t(\varepsilon)+1}^{t_3}\eta_k\geq\frac{1}{\lambda_r}\log\frac{t(\varepsilon)}{\varepsilon}$. Hence

$$\sum_{k=t(\varepsilon)+1}^{T}\eta_k\geq\sum_{k=t(\varepsilon)+1}^{t_3}\eta_k\geq\frac{1}{\lambda_r}\log\frac{t(\varepsilon)}{\varepsilon},\qquad\forall\, T\geq t_3.$$

It thus follows that for each $t\in\{t_1,\dots,t(\varepsilon)\}$,

$$\prod_{k=t+1}^{T}(1-\eta_k\lambda_r)\leq\exp\left\{-\lambda_r\sum_{k=t+1}^{T}\eta_k\right\}\leq\exp\left\{-\lambda_r\sum_{k=t(\varepsilon)+1}^{T}\eta_k\right\}\leq\frac{\varepsilon}{t(\varepsilon)}.$$

Combining with the fact $\eta_t\leq 1$ for each $t\geq t_1$, we see that the first term of (23) can be bounded as

$$\sum_{t=t_1}^{t(\varepsilon)}\eta_t^2\prod_{k=t+1}^{T}(1-\eta_k\lambda_r)\leq\frac{\varepsilon}{t(\varepsilon)}\sum_{t=t_1}^{t(\varepsilon)}\eta_t^2\leq\varepsilon.$$

From the above analysis, we know that when $T \geq \max\{t_1, t(\varepsilon), t_2, t_3\}$,

$$\mathbb{E}_{z_1,\ldots,z_T}[\|r_{T+1}\|^2] \leq \varepsilon + 2\mathcal{E}(f_{\mathcal{H}})\varepsilon.$$

This proves the convergence $\lim_{T\to\infty} \mathbb{E}_{z_1,\ldots,z_T}\|x_{T+1} - x^*\|^2 = 0$ for some $x^* \in \mathbb{R}^d$ and the sufficiency is verified.

From the bound (20), we also see that

$$\mathbb{E}_{z_1,\ldots,z_T}\|x_{T+1} - x^*\|^2 \geq \eta_T^2 \mathcal{E}(f_{\mathcal{H}}), \qquad \forall\, T \in \mathbb{N}.$$

This implies that

$$\sum_{T=1}^{\infty} \sqrt{\mathbb{E}_{z_1,\ldots,z_T}\|x_{T+1} - x^*\|^2} \geq \sqrt{\mathcal{E}(f_{\mathcal{H}})} \sum_{T=1}^{\infty} \eta_T = \infty.$$

The proof of Theorem 3 is complete. ■

In the proof of our second main result, we need some elementary inequalities.

Lemma 10 *(a) For $\nu, a > 0$, there holds*

$$\exp\{-\nu x\} \leq \left(\frac{a}{\nu \mathrm{e}}\right)^a x^{-a}, \qquad \forall x > 0. \tag{24}$$

(b) Let $\nu > 0$ and $q_2 \geq 0$. If $0 < q_1 < 1$, then for any $t \in \mathbb{N}$, we have

$$\sum_{i=1}^{t-1} i^{-q_2} \exp\left\{-\nu \sum_{j=i+1}^{t} j^{-q_1}\right\} \leq \left(\frac{2^{q_1+q_2}}{\nu} + \left(\frac{1+q_2}{\nu(1-2^{q_1-1})\mathrm{e}}\right)^{\frac{1+q_2}{1-q_1}}\right) t^{q_1-q_2}. \tag{25}$$

For $q_1 = 1$, we have

$$\sum_{i=1}^{t-1} i^{-q_2} \exp\left\{-\nu \sum_{j=i+1}^{t} j^{-1}\right\} \leq \begin{cases} \frac{2^{q_2}}{|\nu - q_2 + 1|} t^{-\min\{\nu, q_2-1\}}, & \text{if } \nu \neq q_2 - 1, \\ 2^{q_2} t^{-\nu} \log t, & \text{if } \nu = q_2 - 1. \end{cases} \tag{26}$$

(c) For any $t < T \in \mathbb{N}$ and $\theta \in (0, 1]$, there holds

$$\sum_{k=t+1}^{T} k^{-\theta} \geq \begin{cases} \frac{1}{1-\theta}[(T+1)^{1-\theta} - (t+1)^{1-\theta}], & \text{if } \theta < 1, \\ \log(T+1) - \log(t+1), & \text{if } \theta = 1. \end{cases} \tag{27}$$

(d) For $\theta \in (0, 1]$, $\mu > 0$, and $T \in \mathbb{N}$, there holds

$$\exp\left\{-\mu \sum_{t=1}^{T} t^{-\theta}\right\} \leq \begin{cases} \exp\left\{\frac{\mu}{1-\theta}\right\} \left(\frac{\theta}{\mu \mathrm{e}}\right)^{\frac{\theta}{1-\theta}} T^{-\theta}, & \text{if } \theta < 1, \\ T^{-\mu}, & \text{if } \theta = 1. \end{cases} \tag{28}$$

Proof The inequalities in parts (a) and (b) can be found in (Smale and Zhou, 2009, Lemma 2).

Part (c) can be proved by noting that

$$\sum_{k=t+1}^{T} k^{-\theta} \geq \sum_{k=t+1}^{T} \int_{k}^{k+1} x^{-\theta} dx = \int_{t+1}^{T+1} x^{-\theta} dx.$$

For part (d), we use the inequality (27) in part (c) to derive

$$\exp\left\{-\mu \sum_{t=1}^{T} t^{-\theta}\right\} \leq \begin{cases} \exp\left\{\frac{\mu}{1-\theta}\right\} \exp\left\{-\frac{\mu}{1-\theta} T^{1-\theta}\right\}, & \text{if } \theta < 1, \\ T^{-\mu}, & \text{if } \theta = 1. \end{cases}$$

For $\theta \in (0,1)$, by applying (24) with $\nu = \frac{\mu}{1-\theta}$, $x = T^{1-\theta}$ and $a = \frac{\theta}{1-\theta}$, we get

$$\exp\left\{-\frac{\mu}{1-\theta} T^{1-\theta}\right\} \leq \left(\frac{\theta}{\mu \mathrm{e}}\right)^{\frac{\theta}{1-\theta}} T^{-\theta}.$$

This proves the result. ■

We can now prove our second main result. This is done by following the estimate (22) in the proof of Theorem 3.

Proof of Theorem 4 Since $\eta_1 < 1$, we have $\eta_t < 1$ for all $t \in \mathbb{N}$. Therefore, we can take $t_1 = 1$ in (22) and obtain

$$\begin{aligned} \mathbb{E}_{z_1,\ldots,z_T}[\|r_{T+1}\|^2] &\leq \|r_1\|^2 \prod_{t=1}^{T}(1-\eta_t\lambda_r) + \mathcal{E}(f_{\mathcal{H}}) \sum_{t=1}^{T} \eta_t^2 \prod_{k=t+1}^{T}(1-\eta_k\lambda_r) \\ &\leq \|r_1\|^2 \exp\left\{-\lambda_r\eta_1 \sum_{t=1}^{T} t^{-\theta}\right\} \\ &\quad + \mathcal{E}(f_{\mathcal{H}})\eta_1^2 \sum_{t=1}^{T} t^{-2\theta} \exp\left\{-\lambda_r\eta_1 \sum_{k=t+1}^{T} k^{-\theta}\right\}. \end{aligned} \tag{29}$$

Applying part (d) with $\mu = \lambda_r\eta_1$ of Lemma 10, we know that the first term of (29) can be bounded as

$$\|r_1\|^2 \exp\left\{-\lambda_r\eta_1 \sum_{t=1}^{T} t^{-\theta}\right\} \leq \begin{cases} \|r_1\|^2 \exp\left\{\frac{\lambda_r\eta_1}{1-\theta}\right\} \left(\frac{\theta}{\lambda_r\eta_1 \mathrm{e}}\right)^{\frac{\theta}{1-\theta}} T^{-\theta}, & \text{if } \theta < 1, \\ \|r_1\|^2 T^{-\lambda_r\eta_1}, & \text{if } \theta = 1. \end{cases}$$

Applying part (b) of Lemma 10 with $q_1 = \theta, q_2 = 2\theta$, $\nu = \lambda_r\eta_1$, and noting that $\lambda_r\eta_1 < 1$ by $\eta_1 \in (0,1)$ and $\lambda_r \in (0,1]$, we know that the second term of (29) can be bounded by

$$\begin{cases} \mathcal{E}(f_{\mathcal{H}})\eta_1^2 \left(1 + \frac{2^{3\theta}}{\lambda_r\eta_1} + \left(\frac{1+2\theta}{\lambda_r\eta_1(1-2^{\theta-1})\mathrm{e}}\right)^{\frac{1+2\theta}{1-\theta}}\right) T^{-\theta}, & \text{if } \theta < 1, \\ \mathcal{E}(f_{\mathcal{H}})\eta_1^2 \left(1 + \frac{4}{1-\lambda_r\eta_1}\right) T^{-\lambda_r\eta_1}, & \text{if } \theta = 1. \end{cases}$$

Thus, we get our desired result with $\widetilde{C}_0$ given by

$$\widetilde{C}_0 = \begin{cases} \|r_1\|^2 \exp\left\{\frac{\lambda_r\eta_1}{1-\theta}\right\}\left(\frac{\theta}{\lambda_r\eta_1\mathrm{e}}\right)^{\frac{\theta}{1-\theta}} & \\ \qquad +\mathcal{E}(f_{\mathcal{H}})\eta_1^2\left(\frac{\lambda_r\eta_1+2^{3\theta}}{\lambda_r\eta_1} + \left(\frac{1+2\theta}{\lambda_r\eta_1(1-2^{\theta-1})\mathrm{e}}\right)^{\frac{1+2\theta}{1-\theta}}\right), & \text{if } \theta < 1, \\ \|r_1\|^2 + \mathcal{E}(f_{\mathcal{H}})\eta_1^2\left(1 + \frac{4}{1-\lambda_r\eta_1}\right), & \text{if } \theta = 1. \end{cases}$$

This completes the proof of Theorem 4. ■

Remark 11 *From the proof of Theorem 4, we see that if $\mathcal{E}(f_{\mathcal{H}}) = 0$, then for some $x^* \in \mathbb{R}^d$, we have*

$$\mathbb{E}_{z_1,\ldots,z_T}[\|r_{T+1}\|^2] \le \|r_1\|^2 \prod_{t=1}^{T}(1-\eta_t\lambda_r).$$

The above argument actually can be used to prove Theorem 8.

Proof of Theorem 8 Since $\eta_1 \le 1$, we have $\eta_t \le 1$ for all $t \in \mathbb{N}$. Thus, we can take $t_1 = 1$ in (22) and obtain

$$\begin{aligned} \mathbb{E}_{z_1,\ldots,z_T}[\|r_{T+1}\|^2] &\le \|r_1\|^2 \prod_{t=1}^{T}(1-\eta_t\lambda_r) + \mathcal{E}(f_{\mathcal{H}})\sum_{t=1}^{T}\eta_t^2 \prod_{k=t+1}^{T}(1-\eta_k\lambda_r) \\ &= \|r_1\|^2 \prod_{t=1}^{T}\left(1-\frac{1}{t+t_0}\right) \\ &\quad + \frac{\mathcal{E}(f_{\mathcal{H}})}{\lambda_r^2}\sum_{t=1}^{T}\frac{1}{(t+t_0)^2}\prod_{k=t+1}^{T}\left(1-\frac{1}{k+t_0}\right). \end{aligned}$$

We note that

$$\prod_{k=t+1}^{T}\left(1-\frac{1}{k+t_0}\right) = \prod_{k=t+1}^{T}\frac{k+t_0-1}{k+t_0} = \frac{t+t_0}{T+t_0}.$$

It thus follows that

$$\mathbb{E}_{z_1,\ldots,z_T}[\|r_{T+1}\|^2] \le \|r_1\|^2\frac{t_0}{T+t_0} + \frac{\mathcal{E}(f_{\mathcal{H}})}{\lambda_r^2}\frac{1}{T+t_0}\sum_{t=1}^{T}\frac{1}{t+t_0}.$$

With $\sum_{t=1}^{T}\frac{1}{t+t_0} \le \log\frac{T+t_0}{t_0+1} \le \log T$, we get the desired result with $\widetilde{C}_3$ given by

$$\widetilde{C}_3 = t_0\|r_1\|^2 + \frac{\mathcal{E}(f_{\mathcal{H}})}{\lambda_r^2}.$$

This proves Theorem 8. ■

5. Confidence-Based Estimates for Convergence

In this section, we prove our third main result, Theorem 5. Recall that V_0 is the eigenspace of the covariance matrix C_{ρ_X} associated with the eigenvalue 0. We choose x^* as in the proof of the sufficiency part of Theorem 3. With this choice, r_t belongs to the orthogonal complement $V_0^\perp$ almost surely. Our error analysis is based on the following error decomposition.

5.1 Error Decomposition

For $t \in \mathbb{N}$, set the operator $\Pi_k^t = \prod_{j=k}^t (I - \eta_j C_{\rho_X})$ on $\mathbb{R}^d$ for $k \leq t$ and $\Pi_{t+1}^t = I$. Subtracting x^* from both sides of (4), we have

$$r_{k+1} = (I - \eta_k C_{\rho_X}) r_k + \eta_k \chi_k, \tag{30}$$

where

$$\chi_k = (\tilde{y}_k - \langle \psi_k, x^* \rangle)\psi_k + (C_{\rho_X} - \psi_k \psi_k^T) r_k.$$

Applying this relationship iteratively for $k = t, \cdots, 1$, we get

$$r_{t+1} = \Pi_1^t r_1 + \sum_{k=1}^t \eta_k \Pi_{k+1}^t \chi_k.$$

Thus

$$\|r_{t+1}\| \leq \left\| \Pi_1^t r_1 \right\| + \left\| \sum_{k=1}^t \eta_k \Pi_{k+1}^t \chi_k \right\|. \tag{31}$$

The first term of the bound (31) is caused by the initial error, which is deterministic and will be estimated in subsection 5.2. The second term is the sample error depending on the sample. Since r_k is independent of z_k, by $\mathbb{E}[\tilde{y}_k | \psi_k] = f_\rho(\psi_k)$ and (17),

$$\mathbb{E}[\chi_k | z_1, \ldots, z_{k-1}] = \int_X (f_\rho(\psi) - \langle x^*, \psi \rangle)\psi + (C_{\rho_X} - \psi\psi^T) r_k d\rho_X(\psi) = 0.$$

It tells us that $\{\omega_k := \eta_k \Pi_{k+1}^t \chi_k\}_k$ is a martingale difference sequence. The idea of analyzing the sample error by properties of martingale difference sequences can be found in the recent work in (Tarrés and Yao, 2014) to which details about martingale difference sequences are referred. In particular, we can apply the following Pinelis-Bernstein inequality from (Tarrés and Yao, 2014) (derived from (Pinelis, 1994, Theorem 3.4)) to estimate the sample error.

Lemma 12 *Let $\{\omega_k\}_k$ be a martingale difference sequence in a Hilbert space. Suppose that almost surely $\|\omega_k\| \leq B$ and $\sum_{k=1}^t \mathbb{E}[\|\omega_k\|^2 | \omega_1, \ldots, \omega_{k-1}] \leq L_t^2$. Then for any $0 < \delta < 1$, the following holds with probability at least $1 - \delta$,*

$$\sup_{1 \leq j \leq t} \left\| \sum_{k=1}^j \omega_k \right\| \leq 2 \left(\frac{B}{3} + L_t \right) \log \frac{2}{\delta}.$$

The required bounds B and L_t will be presented in subsections 5.3 and 5.4, respectively.

5.2 Initial Error

Lemma 13 *Let $\eta_k = \eta_1 k^{-\theta}$ with $\theta \in (0,1]$ and $\eta_1 \in (0,1)$. Then*

$$\|\Pi_1^t r_1\| \leq \begin{cases} C_0 t^{-\theta}, & \text{when } \theta < 1, \\ C_0 t^{-\lambda_r \eta_1}, & \text{when } \theta = 1, \end{cases}$$

where

$$C_0 = \begin{cases} \|r_1\| \exp\left\{\frac{\lambda_r \eta_1}{1-\theta}\right\} \left(\frac{\theta}{\lambda_r \eta_1 \mathrm{e}}\right)^{\frac{\theta}{1-\theta}}, & \text{when } \theta < 1, \\ \|r_1\|, & \text{when } \theta = 1. \end{cases}$$

Proof By our choice of x^*, we know that r_1 belongs to the subspace $V_0^\perp$. Thus, we have

$$\|\Pi_1^t r_1\| \leq \|\Pi_1^t|_{V_0^\perp}\| \|r_1\|.$$

Here $\Pi_1^t|_{V_0^\perp}$ denotes the restriction of the self adjoint operator Π_1^t onto $V_0^\perp$. Since $\{\lambda_l : l = 1, 2, \cdots, r\}$ are the eigenvalues of C_{ρ_X} restricted to $V_0^\perp$, $\lambda_1 \leq 1$ and $\eta_1 < 1$, we have

$$\|\Pi_1^t|_{V_0^\perp}\| = \sup_{1 \leq l \leq r} \prod_{k=1}^t (1 - \eta_k \lambda_l) \leq \prod_{k=1}^t (1 - \eta_1 \lambda_r k^{-\theta}) \leq \exp\left\{ -\lambda_r \eta_1 \sum_{k=1}^t k^{-\theta} \right\}.$$

Applying part (d) of Lemma 10, we get our desired result. ■

5.3 Bounding the Residual Sequence

To bound $\omega_k = \eta_k \Pi_{k+1}^t \chi_k$, we start with a rough bound for $\|r_t\|$.

Lemma 14 *Assume that for some constant $M > 0$, $|\widetilde{y}| \leq M$ almost surely. Let $\theta \in [0,1]$ and $\eta_t = \eta_1 t^{-\theta}$ with $\eta_1 \in (0,1)$. Then for any $t \in \mathbb{N}$, we have almost surely*

$$\|r_t\| \leq \begin{cases} C_1 t^{\frac{1-\theta}{2}}, & \text{when } \theta \in [0,1), \\ C_1 \sqrt{\log(et)}, & \text{when } \theta = 1, \end{cases} \tag{32}$$

where C_1 is a constant independent of t given by

$$C_1 = \begin{cases} \sqrt{\frac{\|r_1\|^2 + \eta_1 (M + \|x^*\|)^2}{1-\theta}}, & \text{when } \theta \in [0,1), \\ \sqrt{\|r_1\|^2 + \eta_1 (M + \|x^*\|)^2}, & \text{when } \theta = 1. \end{cases}$$

Proof Rewrite (19) with $x_t = x^* + r_t$ as

$$\begin{aligned} \|r_{t+1}\|^2 &= \|r_t\|^2 + 2\eta_t (\tilde{y}_t - \langle \psi_t, x^* \rangle - \langle \psi_t, r_t \rangle) \langle \psi_t, r_t \rangle \\ &\quad + \eta_t^2 (\tilde{y}_t - \langle \psi_t, x^* \rangle - \langle \psi_t, r_t \rangle)^2 \\ &= \|r_t\|^2 + \mathcal{F}(\langle \psi_t, r_t \rangle), \end{aligned}$$

where $\mathcal{F}:\mathbb{R}\to\mathbb{R}$ is a quadratic function given by

$$\mathcal{F}(\mu)=\mathcal{F}_{\eta_t,\tilde{y}_t,\psi_t,x^*}(\mu)=\eta_t(\eta_t-2)\mu^2+2\eta_t(1-\eta_t)(\tilde{y}_t-\langle\psi_t,x^*\rangle)\mu+\eta_t^2(\tilde{y}_t-\langle\psi_t,x^*\rangle)^2.$$

Note that $\eta_t(\eta_t-2)\le 0$ by $0<\eta_t\le\eta_1\le 1$. A simple calculation shows that

$$\max_{x\in\mathbb{R}}\mathcal{F}(x)=-\frac{\eta_t^2(1-\eta_t)^2(\tilde{y}_t-\langle\psi_t,x^*\rangle)^2}{\eta_t(\eta_t-2)}+\eta_t^2(\tilde{y}_t-\langle\psi_t,x^*\rangle)^2=\frac{\eta_t(\tilde{y}_t-\langle\psi_t,x^*\rangle)^2}{2-\eta_t}.$$

Since $|\tilde{y}_t|\le M$ almost surely and $\|\psi_t\|=1$,

$$|\tilde{y}_t-\langle\psi_t,x^*\rangle|\le|\tilde{y}_t|+\|\psi_t\|\|x^*\|\le M+\|x^*\|.$$

Thus,

$$\|r_{t+1}\|^2\le\|r_t\|^2+\frac{\eta_t(\tilde{y}_t-\langle\psi_t,x^*\rangle)^2}{2-\eta_t}\le\|r_t\|^2+\eta_t(M+\|x^*\|)^2.$$

Using this relationship iteratively yields

$$\|r_{t+1}\|^2\le\|r_1\|^2+\sum_{k=1}^t\eta_k(M+\|x^*\|)^2=\|r_1\|^2+\eta_1(M+\|x^*\|)^2\sum_{k=1}^t k^{-\theta}.$$

Since that

$$\sum_{k=1}^t k^{-\theta}\le 1+\sum_{k=2}^t\int_{k-1}^k x^{-\theta}dx=\begin{cases}\frac{t^{1-\theta}-\theta}{1-\theta}, & \text{when } \theta\in[0,1),\\ \log(et), & \text{when } \theta=1,\end{cases}$$

we get

$$\|r_t\|^2\le\begin{cases}\frac{\|r_1\|^2+\eta_1(M+\|x^*\|)^2}{1-\theta}t^{1-\theta}, & \text{when } \theta\in[0,1),\\ (\|r_1\|^2+\eta_1(M+\|x^*\|)^2)\log(et), & \text{when } \theta=1,\end{cases}$$

which leads to the desired result. ■

5.4 Estimating Conditional Variance and Upper Bound

In this subsection, we give bounds for the two terms $\sum_{k=1}^t\eta_k^2\mathbb{E}[\|\Pi_{k+1}^t\chi_k\|^2|z_1,\ldots,z_{k-1}]$ and $\sup_{1\le k\le t}\|\eta_k\Pi_{k+1}^t\chi_k\|$ required in applying the Pinelis-Bernstein inequality.

Lemma 15 *Let $\eta_k=\eta_1k^{-\theta}$ with $\theta\in(0,1]$ and $\eta_1\in(0,1)$. Then almost surely we have*

$$\begin{aligned}&\sum_{k=1}^t\eta_k^2\mathbb{E}[\|\Pi_{k+1}^t\chi_k\|^2|z_1,\ldots,z_{k-1}]\\&\le\sum_{k=1}^t\eta_1^2k^{-2\theta}\exp\left\{-2\eta_1\lambda_r\sum_{j=k+1}^t j^{-\theta}\right\}\left(\mathcal{E}(f_{\mathcal{H}})+\|r_k\|_2^2\right).\end{aligned}\tag{33}$$

Proof Recall that both ψ_k and r_k belong to $V_0^\perp$ almost surely for each $k \in \mathbb{N}$. As a result, χ_k also belongs to $V_0^\perp$ almost surely for each k. Hence

$$\sum_{k=1}^t \eta_k^2 \mathbb{E}[\|\Pi_{k+1}^t \chi_k\|^2 | z_1, \ldots, z_{k-1}] \leq \sum_{k=1}^t \eta_k^2 \|\Pi_{k+1}^t|_{V_0^\perp}\|^2 \mathbb{E}[\|\chi_k\|^2 | z_1, \ldots, z_{k-1}].$$

Since $\eta_1 < 1$ and $\lambda_1 \leq 1$, we have

$$\begin{aligned} \left\|\Pi_{k+1}^t|_{V_0^\perp}\right\| &= \sup_{1 \leq l \leq r} \prod_{j=k+1}^t (1 - \eta_j \lambda_l) \leq \prod_{j=k+1}^t (1 - \eta_j \lambda_r) \\ &\leq \exp\left\{-\lambda_r \sum_{j=k+1}^t \eta_j\right\} = \exp\left\{-\lambda_r \eta_1 \sum_{j=k+1}^t j^{-\theta}\right\}. \end{aligned} \tag{34}$$

Thus,

$$\begin{aligned} &\sum_{k=1}^t \eta_k^2 \mathbb{E}[\|\Pi_{k+1}^t \chi_k\|^2 | z_1, \ldots, z_{k-1}] \\ &\leq \sum_{k=1}^t \eta_k^2 \exp\left\{-2\lambda_r \eta_1 \sum_{j=k+1}^t j^{-\theta}\right\} \mathbb{E}[\|\chi_k\|^2 | z_1, \ldots, z_{k-1}]. \end{aligned} \tag{35}$$

Since r_k does not depend on z_k, we see from $\|\psi_k\| = 1$, $\mathbb{E}[\tilde{y}_k | \psi_k] = f_\rho(\psi_k)$ and (17) that

$$\begin{aligned} &\mathbb{E}_{z_k}[\langle(\tilde{y}_k - \langle\psi_k, x^*\rangle)\psi_k, (C_{\rho_X} - \psi_k\psi_k^T)r_k\rangle] \\ &= \mathbb{E}_{z_k}[(\tilde{y}_k - \langle\psi_k, x^*\rangle)\langle\psi_k, C_{\rho_X} r_k\rangle] - \mathbb{E}_{z_k}[(\tilde{y}_k - \langle\psi_k, x^*\rangle)\langle\psi_k, r_k\rangle\|\psi_k\|^2] \\ &= \mathbb{E}_{\psi_k}[(f_\rho(\psi_k) - \langle\psi_k, x^*\rangle)\langle\psi_k, C_{\rho_X} r_k\rangle] - \mathbb{E}_{\psi_k}[(f_\rho(\psi_k) - \langle\psi_k, x^*\rangle)\langle\psi_k, r_k\rangle] = 0. \end{aligned}$$

It thus follows that

$$\begin{aligned} &\mathbb{E}[\|\chi_k\|^2 | z_1, \ldots, z_{k-1}] = \mathbb{E}_{z_k}[\|\chi_k\|^2] \\ &= \mathbb{E}_{z_k}[(\tilde{y}_k - \langle\psi_k, x^*\rangle)^2] + \mathbb{E}_{z_k}[\|(C_{\rho_X} - \psi_k\psi_k^T)r_k\|^2] \\ &= \mathcal{E}(f_{\mathcal{H}}) + \mathbb{E}_{z_k}\langle(C_{\rho_X} - C_{\rho_X}^2)r_k, r_k\rangle \\ &\leq \mathcal{E}(f_{\mathcal{H}}) + \|r_k\|_2^2. \end{aligned}$$

Putting the above bound into (35), we get the desire result. ■

Lemma 16 *Assume that for some constant $M > 0$, $|\widetilde{y}| \leq M$ almost surely. Let $\theta \in [0, 1]$ and $\eta_t = \eta_1 t^{-\theta}$ with $\eta_1 \in (0, 1)$. Then for any $t \in \mathbb{N}$, we have almost surely*

$$\sup_{1 \leq k \leq t} \|\eta_k \Pi_{k+1}^t \chi_k\| \leq \begin{cases} C_2 t^{-\theta} \max\left\{\sup_{1 \leq k \leq t} \|r_k\|,\ 1\right\}, & \textit{when } \theta < 1, \\ C_2 t^{-\lambda_r \eta_1} \max\left\{\sup_{1 \leq k \leq t} \|r_k\|,\ 1\right\}, & \textit{when } \theta = 1, \end{cases} \tag{36}$$

where C_2 is a constant given by

$$C_2 = \begin{cases} \eta_1(M + \|x^*\| + 2)\left(2^\theta + \left(\frac{\theta}{\mathrm{e}\lambda_r\eta_1(1 - 2^{\theta-1})}\right)^{\frac{\theta}{1-\theta}}\right), & \textit{when } \theta < 1, \\ \eta_1(M + \|x^*\| + 2)2^{\lambda_r \eta_1}, & \textit{when } \theta = 1. \end{cases}$$

Proof Let $k \in \{1, \ldots, t\}$. From the definition of χ_k, we have

$$\|\chi_k\| \leq (|\tilde{y}_k| + \|\psi_k\|\|x^*\|)\|\psi_k\| + \|C_{\rho_X} - \psi_k\psi_k^T\|\|r_k\|.$$

But $|\tilde{y}_k| \leq M$, $\|\psi_k\| = 1$ and $\|C_{\rho_X}\| \leq 1$. So we have

$$\|\chi_k\| \leq M + \|x^*\| + 2\|r_k\| \leq (M + \|x^*\| + 2)\max\{\|r_k\|, 1\}.$$

This together with (34) and the fact that χ_k belongs to $V_0^\perp$ implies

$$\begin{aligned}
\|\eta_k \Pi_{k+1}^t \chi_k\| &\leq \eta_1 k^{-\theta} \|\Pi_{k+1}^t|_{V_0^\perp}\| \|\chi_k\| \\
&\leq \eta_1 (M + \|x^*\| + 2) k^{-\theta} \|\Pi_{k+1}^t|_{V_0^\perp}\| \max\{\|r_k\|, 1\} \\
&\leq \eta_1 (M + \|x^*\| + 2) k^{-\theta} \exp\left\{-\lambda_r \eta_1 \sum_{j=k+1}^{t} j^{-\theta}\right\} \max\{\|r_k\|, 1\}.
\end{aligned}$$

What is left is to estimate

$$I_k := k^{-\theta} \exp\left\{-\lambda_r \eta_1 \sum_{j=k+1}^{t} j^{-\theta}\right\}.$$

For $\theta \in [1/2, 1)$, applying part (c) of Lemma 10 gives

$$I_k \leq k^{-\theta} \exp\left\{-\frac{\lambda_r \eta_1}{1-\theta}[(t+1)^{1-\theta} - (k+1)^{1-\theta}]\right\}.$$

If $k \geq t/2$, then $k^{-\theta} \leq 2^\theta t^{-\theta}$ and thus

$$I_k \leq 2^\theta t^{-\theta}.$$

If $1 \leq k < t/2$, then we have $k+1 \leq (t+1)/2$ and $(t+1)^{1-\theta} - (k+1)^{1-\theta} \geq (1-2^{\theta-1})(t+1)^{1-\theta}$. It follows that

$$I_k \leq \exp\left\{-\frac{\lambda_r \eta_1 (1-2^{\theta-1})}{1-\theta} t^{1-\theta}\right\}.$$

Applying part (a) of Lemma 10 with $x = t^{1-\theta}$, $\nu = \frac{\lambda_r \eta_1 (1-2^{\theta-1})}{1-\theta}$ and $a = \frac{\theta}{1-\theta}$, we get

$$I_k \leq \left(\frac{\theta}{\mathrm{e}\lambda_r \eta_1 (1-2^{\theta-1})}\right)^{\frac{\theta}{1-\theta}} t^{-\theta}.$$

For $\theta = 1$, by part (c) of Lemma 10, with $\lambda_r \eta_1 < 1$, we have

$$I_k \leq k^{-1}\left(\frac{t+1}{k+1}\right)^{-\lambda_r \eta_1} = \left(\frac{t}{t+1} \cdot \frac{k+1}{k}\right)^{\lambda_r \eta_1} t^{-\lambda_r \eta_1} k^{\lambda_r \eta_1 - 1} \leq 2^{\lambda_r \eta_1} t^{-\lambda_r \eta_1}.$$

From the above analysis, we conclude the desired result. ■

5.5 Preliminary Error Analysis

Based on the above estimates, we can apply Lemma 12 to obtain an error bound.

Proposition 17 *Under the assumptions of Theorem 5, for some $x^* \in \mathbb{R}^d$ and for any $0 < \delta < 1$ and fixed $t \in \mathbb{N}$, with confidence at least $1-\delta$, we have*

$$\|x_{t+1} - x^*\| \leq \begin{cases} \widetilde{C}_2 t^{\frac{1}{2}-\theta} \log\frac{2}{\delta}, & \text{when } \theta \in [\frac{1}{3}, 1), \\ \widetilde{C}_2 t^{-\lambda_r \eta_1} \sqrt{\log(\mathrm{e}t)} \log\frac{2}{\delta}, & \text{when } \theta = 1, \end{cases} \tag{37}$$

where $\widetilde{C}_2$ is a positive constant independent of t or δ (given explicitly in the proof).

Proof To apply the Pinelis-Bernstein inequality to estimate $\|\sum_{k=1}^t \eta_k \Pi_{k+1}^t \chi_k\|$, we need bounds B and L_t.

By Lemmas 14 and 16, we have

$$\sup_{1\leq k\leq t} \|\eta_k \Pi_{k+1}^t \chi_k\| \leq \begin{cases} C_2(C_1+1)t^{\frac{1-3\theta}{2}}, & \text{when } \theta < 1, \\ C_2(C_1+1)t^{-\lambda_r\eta_1}\sqrt{\log(et)}, & \text{when } \theta = 1. \end{cases} \tag{38}$$

By Lemmas 15 and 14, we get

$$\begin{aligned} &\sum_{k=1}^t \eta_k^2 \mathbb{E}[\|\Pi_{k+1}^t \chi_k\|^2 | z_1, \ldots, z_{k-1}] \\ &\leq \begin{cases} C_3 \sum_{k=1}^t k^{-(3\theta-1)} \exp\left\{-2\lambda_r\eta_1 \sum_{j=k+1}^t j^{-\theta}\right\}, & \text{when } \theta \in [\frac{1}{3}, 1), \\ C_3 \log(et) \sum_{k=1}^t k^{-2} \exp\left\{-2\lambda_r\eta_1 \sum_{j=k+1}^t j^{-1}\right\}, & \text{when } \theta = 1, \end{cases} \end{aligned}$$

where

$$C_3 = (\mathcal{E}(f_{\mathcal{H}}) + C_1^2)\eta_1^2.$$

Applying part (b) of Lemma 10 with $\nu = 2\lambda_r\eta_1 < 1$, $q_1 = \theta$ and $q_2 = 3\theta - 1$, we have for $\theta < 1$,

$$\begin{aligned} &\sum_{k=1}^t k^{-(3\theta-1)} \exp\left\{-2\lambda_r\eta_1 \sum_{j=k+1}^t j^{-\theta}\right\} \\ &\leq \left(\frac{2^{4\theta-1}}{2\lambda_r\eta_1} + \left(\frac{3\theta}{2\lambda_r\eta_1 \mathrm{e}(1-2^{\theta-1})}\right)^{\frac{3\theta}{1-\theta}}\right) t^{1-2\theta} + t^{1-3\theta}, \end{aligned}$$

and for $\theta = 1$,

$$\sum_{k=1}^t k^{-2} \exp\left\{-2\lambda_r\eta_1 \sum_{j=k+1}^t j^{-1}\right\} \leq \frac{4}{1-2\lambda_r\eta_1} t^{-2\lambda_r\eta_1} + t^{-2}.$$

Therefore, we get

$$\sum_{k=1}^t \eta_k^2 \mathbb{E}[\|\Pi_{k+1}^t \chi_k\|^2 | z_1, \ldots, z_{k-1}] \leq \begin{cases} C_4 t^{1-2\theta}, & \text{when } \theta \in [\frac{1}{3}, 1), \\ C_4 t^{-2\lambda_r\eta_1} \log(et), & \text{when } \theta = 1, \end{cases} \tag{39}$$

with

$$C_4 = \begin{cases} C_3 \left(\frac{2^{4\theta-1}}{2\lambda_r\eta_1} + \left(\frac{3\theta}{2\lambda_r\eta_1 \mathrm{e}(1-2^{\theta-1})} \right)^{\frac{3\theta}{1-\theta}} + 1 \right), & \text{when } \theta \in [\frac{1}{3}, 1), \\ C_3 \frac{5-2\lambda_r\eta_1}{1-2\lambda_r\eta_1}, & \text{when } \theta = 1. \end{cases}$$

Applying Lemma 12 to the martingale difference sequence $\{\omega_k := \eta_k \Pi_{k+1}^t \chi_k\}_k$ with B and L_t given by (38) and (39) respectively, we know that with probability at least $1-\delta$,

$$\sup_{1\le j\le t} \left\| \sum_{k=1}^{j} \eta_k \Pi_{k+1}^t \chi_k \right\| \le \begin{cases} C_5 t^{\frac{1-2\theta}{2}} \log \frac{2}{\delta}, & \text{when } \theta \in [\frac{1}{3}, 1), \\ C_5 t^{-\lambda_r\eta_1} \sqrt{\log(\mathrm{e}t)} \log \frac{2}{\delta}, & \text{when } \theta = 1, \end{cases}$$

where

$$C_5 = 2\left(C_2(C_1+1)/3 + \sqrt{C_4} \right).$$

Putting this bound into (31) with t replaced by j, and then applying Lemma 13 to bound the initial error, we get the desired result with $\widetilde{C}_2 = C_0 + C_5$ from Lemma 12. ■

In the above procedure, we have used a rough bound (32) for $\|r_t\|$. This rough bound tends to ∞ as t becomes large. In contrast, the bound provided in Proposition 17 tends to 0 (when $\theta \in (1/2, 1]$) and is much better. But this bound holds with confidence. We shall use this refined bound to improve our estimates in the following subsection.

5.6 Improved Error Analysis

In this subsection, we prove our third main result by improving the preliminary confidence-based error bound in Proposition 17.

Proof of Theorem 5 When $\theta = 1$, our desired bound follows from (37) with $\widetilde{C}_1 = 2\widetilde{C}_2$.

It remains to prove the case $\theta \in [1/2, 1)$. Let $T \in N$. Applying Proposition 17 with $t = 1, \cdots, T$, and taking the union event followed by rescaling, we know that there exists a subset Z_δ^T of Z^T with measure at least $1-\delta$ such that

$$\|r_t\| \le C_6 \log \frac{2}{\delta} \log T, \qquad \forall\ t = 1, \ldots, T+1,\ (z_1, \ldots, z_T) \in Z_\delta^T, \tag{40}$$

where $C_6 = 2\widetilde{C}_2 + \|r_1\|$.

Now we turn to the essential part of the proof. Define another martingale difference sequence $\{\widetilde{\omega}_k\}_k$ by multiplying the one in the proof of Proposition 17 by a characteristic function $\mathbf{1}_{\{\|r_k\| \le C_6 \log \frac{2}{\delta} \log T\}}$ as

$$\widetilde{\omega}_k = \eta_k \Pi_{k+1}^T \chi_k \mathbf{1}_{\{\|r_k\| \le C_6 \log \frac{2}{\delta} \log T\}}.$$

From (36) and the multiplication with the characteristic function $\mathbf{1}_{\{\|r_k\| \le C_6 \log \frac{2}{\delta} \log T\}}$, we have

$$\sup_{1\le k\le T} \|\widetilde{\omega}_k\| \le C_2 C_6 \log\left(\frac{2}{\delta}\right) (\log T) T^{-\theta}. \tag{41}$$

Notice that the characteristic function $\mathbf{1}_{\{\|r_k\|\leq C_6 \log\frac{2}{\delta}\log T\}}$ is independent of z_k. Also, from the proof of Lemma 15, we know that for each $k \in \{1, \ldots, T\}$,

$$\mathbb{E}[\|\Pi_{k+1}^t \chi_k\|^2 | z_1, \ldots, z_{k-1}] \leq \exp\left\{-\lambda_r \eta_1 \sum_{j=k+1}^{t} j^{-\theta}\right\} \left(\mathcal{E}(f_{\mathcal{H}}) + \|r_k\|_2^2\right).$$

It follows by setting $C_7 = (\mathcal{E}(f_{\mathcal{H}}) + C_6^2)\eta_1^2$ that

$$\sum_{k=1}^{T} \mathbb{E}\left[\|\widetilde{\omega}_k\|^2 | z_1, \ldots, z_{k-1}\right] \leq C_7 \left(\log\frac{2}{\delta}\log T\right)^2 \sum_{k=1}^{T} k^{-2\theta} \exp\left\{-2\lambda_r \eta_1 \sum_{j=k+1}^{T} j^{-\theta}\right\}.$$

Applying part (b) of Lemma 10 yields

$$\begin{aligned}&\sum_{k=1}^{T} \mathbb{E}\left[\|\widetilde{\omega}_k\|^2 | z_1, \ldots, z_{k-1}\right] \\ &\leq C_7 \left(\log\frac{2}{\delta}\log T\right)^2 \left(\frac{2^{3\theta}}{2\lambda_r\eta_1} + \left(\frac{1+2\theta}{2\lambda_r\eta_1 \mathrm{e}(1-2^{\theta-1})}\right)^{\frac{1+2\theta}{1-\theta}} + 1\right) T^{-\theta}.\end{aligned}$$

Using this bound as L_T and (41) as the bound B in Lemma 12, we know that there exists another subset $\tilde{Z}_\delta^T$ of Z^T with measure at least $1-\delta$ such that for every $(z_1, \ldots, z_T) \in \tilde{Z}_\delta^T$, there holds

$$\left\|\sum_{k=1}^{T} \widetilde{\omega}_k\right\| \leq C_8 T^{\frac{-\theta}{2}} \left(\log\frac{2}{\delta}\right)^2 \log T,$$

where

$$C_8 = \frac{2C_2C_6}{3} + 2\sqrt{C_7}\left(\frac{2^{3\theta} + 2\lambda_r\eta_1}{2\lambda_r\eta_1} + \left(\frac{1+2\theta}{2\lambda_r\eta_1 \mathrm{e}(1-2^{\theta-1})}\right)^{\frac{1+2\theta}{1-\theta}}\right)^{\frac{1}{2}}.$$

This together with (40) tells us that for every $(z_1, \ldots, z_T) \in Z_\delta^T \cap \tilde{Z}_\delta^T$, there holds

$$\left\|\sum_{k=1}^{T} \eta_k \Pi_{k+1}^T \chi_k\right\| \leq C_8 T^{\frac{-\theta}{2}} \left(\log\frac{2}{\delta}\right)^2 \log T. \tag{42}$$

The subset $Z_\delta^T \cap \tilde{Z}_\delta^T$ has measure at least $1-2\delta$. Therefore, we can put (42) into (31), and apply Lemma 13 to bound the initial error, which proves Theorem 5 for the case $\theta \in [1/2, 1)$ after scaling δ to $\delta/2$ and setting the constant $\widetilde{C}_1 = C_0 + C_8$. ■

6. Almost Sure Convergence

In this section, we prove the almost sure convergence of the randomized Kaczmarz algorithm. Recall that the almost sure convergence of a sequence of random variables $\{X_n\}$ towards X means that

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1,$$

or equivalently,

$$\lim_{n\to\infty} \mathbb{P}\Big(\sup_{k\geq n} |X_k - X| > \varepsilon\Big) = 0 \quad \text{for any } \varepsilon > 0.$$

The Borel-Cantelli Lemma (see e.g. (Klenke, 2010)) asserts for a sequence $(E_n)_n$ of events that if the sum of the probabilities is finite $\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty$, then the probability that infinitely many of them occur is 0, that is, $\mathbb{P}\left(\limsup_{n\to\infty} E_n\right) = \mathbb{P}\left(\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} E_n\right) = 0$. The following lemma is an easy consequence of the Borel-Cantelli Lemma. We give the proof for completeness.

Lemma 18 *Let $\{X_n\}$ be a sequence of events in some probability space and $\{\varepsilon_n\}$ be a sequence of positive numbers satisfying $\lim_{n\to\infty} \varepsilon_n = 0$. If*

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|X_n - X| > \varepsilon_n\right) < \infty,$$

then X_n converges to X almost surely.

Proof Since $\lim_{n\to\infty} \varepsilon_n = 0$, for any $\varepsilon > 0$, there exists some $n \in \mathbb{N}$ such that for all $k \geq n$, $\varepsilon_k < \varepsilon$. Thus,

$$\mathbb{P}\Big(\sup_{k\geq n} |X_k - X| > \varepsilon\Big) \leq \mathbb{P}\Big(\bigcup_{k\geq n} (|X_k - X| > \varepsilon_k)\Big) \leq \sum_{k\geq n} \mathbb{P}\Big(|X_k - X| > \varepsilon_k\Big).$$

Letting $n \to \infty$, one gets $\mathbb{P}\left(\sup_{k\geq n} |X_k - X| > \varepsilon\right) \to 0$. This proves the result. ■

Now we can apply Lemma 18 to prove our last main result.

Proof of Theorem 6 Set

$$\Lambda_t = \begin{cases} t^{-\theta/2} & \text{when } \theta < 1, \\ t^{-\lambda_r \eta_1} & \text{when } \theta = 1. \end{cases}$$

By Theorem 5, we have for any $t \geq 2$ and $0 < \delta_t < 1$,

$$\mathbb{P}\left(\Lambda_t^{\epsilon-1} \|x_{t+1} - x^*\| > \widetilde{C}_1 \Lambda_t^{\epsilon} \left(\log \frac{4}{\delta_t}\right)^2 \log t\right) \leq \delta_t.$$

Choose $\delta_t = t^{-2}$, and $\varepsilon_t = \widetilde{C}_1 \Lambda_t^{\epsilon} \left(\log 4/\delta_t\right)^2 \log t$. Obviously

$$\sum_{t=2}^{\infty} \mathbb{P}\left(\Lambda_t^{\epsilon-1} \|x_{t+1} - x^*\| > \varepsilon_t\right) \leq \sum_{t=2}^{\infty} \delta_t < \infty$$

and

$$\varepsilon_t \leq 4\widetilde{C}_1 \Lambda_t^{\epsilon} \log^3(2t) \to 0, \quad \text{as } t \to \infty.$$

Then our conclusion of Theorem 6 follows from Lemma 18. ■

Remark 19 *The above method of proof can be used to get a more quantitative estimate for the almost sure convergence of the Kaczmarz algorithm with noiseless random measurements (Chen and Powell, 2012). In that setting, $\eta_t \equiv 1$, $y_t = f_\rho(\psi_t)$ and $r = d$. It was shown in (Strohmer and Vershynin, 2009; Chen and Powell, 2012) that with $q = 1 - \lambda_r$,*

$$\mathbb{E}\|x_{t+1} - x^*\|^2 \leq q^t \|r_1\|^2.$$

It follows from the Chebyshev inequality that for any $\epsilon \in (0,1)$,

$$\mathbb{P}\left(q^{t(\epsilon-1)}\|x_{t+1} - x^*\|^2 > q^{t\epsilon}t^2\right) = \mathbb{P}\left(\|x_{t+1} - x^*\|^2 > q^t t^2\right) \leq \frac{\mathbb{E}[\|x_{t+1} - x^*\|^2]}{q^t t^2}.$$

Thus, we get

$$\mathbb{P}\left(q^{t(\epsilon-1)}\|x_{t+1} - x^*\|^2 > q^{t\epsilon}t^2\right) \leq \|r_1\| t^{-2}.$$

Obviously, $q^{t\epsilon}t^2 \to 0$ as $t \to \infty$, and $\sum_{t=1}^\infty \|r_1\| t^{-2} < \infty$. Applying Lemma 18 with $\varepsilon_t = q^{t\epsilon}t^2$, we know that for any $\epsilon \in (0,1)$,

$$\lim_{t\to\infty} (1 - \lambda_r)^{t(\epsilon-1)} \|x_{t+1} - x^*\|^2 = 0 \quad \textit{almost surely.}$$

7. Simulations and Discussions

In this section we provide some numerical simulations and further discussions on our error analysis.

To illustrate our derived convergence rates and compare with the existing literature, we carry out numerical simulations corresponding to Example 2 with the same data distributions as in (Needell, 2010): $m = 200, d = 100$, $A \in \mathbb{R}^{200\times 100}$ is a Gaussian matrix with each entry drawn independently from the standard normal distribution $N(0,1)$, and $y \in \mathbb{R}^{100}$ is a Gaussian noise with each component drawn independently from the normal distribution with mean 0 and standard deviation 0.02. The measurement vectors $\{\psi_t = \frac{1}{\|\varphi_t\|}\varphi_t\}$ are drawn from the normalized rows of A as in Example 2 and $\{\widetilde{y}_t = y_t/\|\varphi_t\|\}$ with mean $x^* = 0$. We conduct 100 trials for each choice of the relaxation parameter sequences $\eta_t = 1, \eta_t = 1/\sqrt{t}, \eta_t = 1/t$. In each trial, algorithm (4) is run 100 times with random Gaussian initial vectors of norm $\|x_1\| = 0.02$. Figure 1 depicts the error $\|x_{t+1} - x^*\|$ for $t = 1, \ldots, 1500$ (averaged with 100 trials and 100 initial vectors). The black line is a plot with the constant relaxation parameter sequence $\eta_t = 1$, which verifies the divergence of the algorithm, as proved in (Needell, 2010). The blue line is a plot with $\eta_t = 1/\sqrt{t}$, which hints a slow convergence of the algorithm. The red line is a plot with $\eta_t = 1/t$, which confirms a faster convergence. The above simulations are consistent with our error analysis.

In this paper, a learning theory approach to the relaxed randomized Kaczmarz algorithm is presented. It yields new results and observations including a necessary and sufficient condition (9), stated in Theorem 3, for the convergence in expectation when the sampling process is noisy or nonlinear. For noise-free and linear sampling processes (that is, $\mathcal{E}(f_\mathcal{H}) = 0$), we can see from Remark 11 with $\eta_t \equiv 1$ that $\mathbb{E}_{z_1,\ldots,z_T}[\|x_{T+1}-x^*\|^2] \leq \|x_1 - x^*\|^2(1-\lambda_r)^T$. This exponential convergence result was proved in (Strohmer and Vershynin, 2009) for Example 2 under the restriction that the matrix A has full column rank, where the number

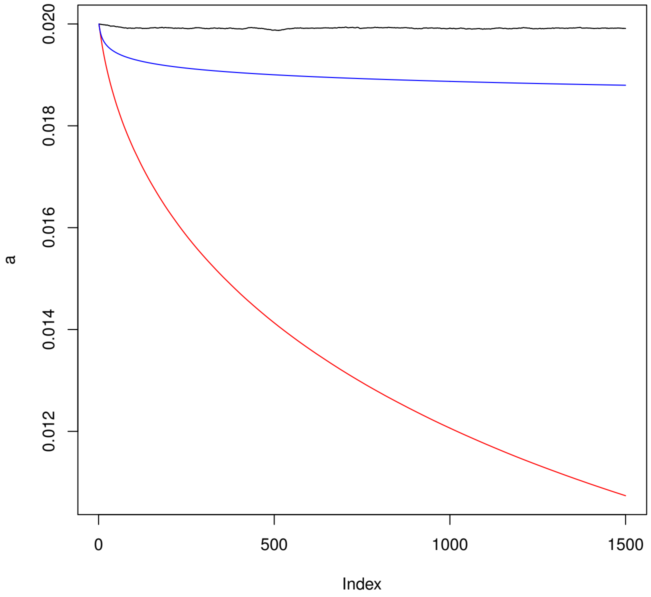

Figure 1: Error of the relaxed randomized Kaczmarz algorithm with $\eta_t = 1$ (black line), $\eta_t = 1/\sqrt{t}$ (blue line), and $\eta_t = 1/t$ (red line)

$1 - \lambda_r$ is replaced by a quantity involving $\|A^{-1}\| = \inf\{M : M\|Ax\| \geq \|x\| \text{ for all } x\}$. Our result is more general (valid for underdetermined systems with $\|A^{-1}\| = \infty$).

In the framework of Kaczmarz algorithms, we consider online learning algorithms associated with the least squares loss. It would be interesting to extend our study to algorithms associated with more general loss functions (Ying and Zhou, 2006) such as hinge loss, and to consider error analysis without requiring the approximation error (Ying and Zhou, 2006) tending to zero.

Acknowledgments

The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 104113]. The authors would like to thank the referees for constructive suggestions and Dr. Xin Guo for helping with the numerical simulations. The corresponding author is Ding-Xuan Zhou.

References

X. Chen and A. Powell. Almost sure convergence for the Kaczmarz algorithm with random measurements. *Journal of Fourier Analysis and Applications*, 18:1195–1214, 2012.

T. Hu, J. Fan, Q. Wu, and D. X. Zhou. Regularization schemes for minimum error entropy principle. *Analysis and Applications*, 13:437–455, 2015.

J. Johnston. *Econometric Methods*. McGraw Hill, New York, 1963.

S. Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres A*, 35:355–357, 1937.

A. Klenke. *Probability Theory: A Comprehensive Course*. Springer-Verlag, London, 2008.

D. Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT. Numerical Mathematics*, 50:395–403, 2010.

I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *Annals of Probability*, 22:1679–1706, 1994.

Y. Ying and M. Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8:561–596, 2008.

S. Smale and Y. Yao. Online learning algorithms. *Foundations of Computational Mathematics*, 6:145–170, 2005.

S. Smale and D. X. Zhou. Online learning with Markov sampling. *Analysis and Applications*, 7:87–113, 2009.

T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15:262–278, 2009.

P. Tarrés and Y. Yao. Online learning as stochastic approximations of regularization paths. *IEEE Transactions on Information Theory*, 60:5716–5735, 2014.

V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

Y. Ying and D. X. Zhou. Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52:4775–4788, 2006.

A. Zouzias and N. M. Freris. Randomized extended Kaczmarz for solving least squares. *SIAM Journal on Matrix Analysis and Applications*, 34:773–793, 2013.