Rates of Convergence of Randomized Kaczmarz Algorithms in Hilbert Spaces

Xin Guo[*] Junhong Lin[†] Ding-Xuan Zhou[‡]

Abstract

Recently, the Randomized Kaczmarz algorithm (RK) draws much attention because of its low computational complexity and less requirement on computer memory. Many existing results on analysis focus on the behavior of RK in Euclidean spaces, and typically derive exponential converge rates with the base tending to one, as the condition number of the system increases. The dependence on the condition number largely restricts the application of these estimates. There are also results using relaxation (i.e., small step-sizes tending to zero as the sample size increases) to achieve polynomial convergence rates of RK in Hilbert spaces. In this paper, we prove the weak convergence (with polynomial rates) of RK with constant step-size in Hilbert spaces. As a consequence, the strong convergence of RK in Euclidean spaces is obtained with condition number-free polynomial rates. In the setting with noisy data, we study the relaxation technique and obtain a strong convergence of RK in Hilbert spaces, with the rates arbitrarily close to the minimax optimal ones. We apply the analysis to reproducing kernel-based online gradient descent learning algorithms and improve the state-of-the-art convergence estimate in the literature.

Keywords: randomized Kaczmarz algorithm, Hilbert space, weak convergence, online gradient descent learning algorithm, relaxation

1 Introduction

Let $(H, \langle \cdot, \cdot \rangle, \| \cdot \|)$ be a real separable Hilbert space. Let $x^0 \in H$ be an unknown vector. Consider a sequence $\{(\xi_i, y_i)\}$ of observations, where for each i, $\xi_i \in H$ with $\|\xi_i\| = 1$,

$$y_i = \langle \xi_i, x^0 \rangle + \epsilon_i \in \mathbb{R}, \tag{1}$$

and ϵ_i models the observational noise. In this paper, we study the randomized Kaczmarz algorithm (RK) for recovering x^0. The algorithm is defined with $x_1 = 0 \in H$ and then iteratively by

$$x_{k+1} = x_k + \eta_k (y_k - \langle \xi_k, x_k \rangle) \xi_k, \quad k \geq 1, \tag{2}$$

where $\eta_k \in (0, 2)$ is the step-size of the k-th step. The iteration (2) is an instance of stochastic gradient descent along the negative gradient of least squares, $-\nabla_x \frac{1}{2}(y_k - \langle \xi_k, x \rangle)^2|_{x=x_k}$. It is expected that after sufficient steps of iteration (i.e., when k is large enough), x_{k+1} would be close to x^0,

$$\lim_{k\to\infty} r_{k+1} = 0, \quad \text{where } r_{k+1} := x_{k+1} - x^0, \text{ thus } r_1 = -x^0. \tag{3}$$

[*] School of Mathematics and Physics, The University of Queensland, Brisbane, 4072, Australia. xin.guo@uq.edu.au

[†] Center for Data Science, Zhejiang University, 310027, Hangzhou, China. junhong@zju.edu.cn

[‡] School of Data Science and Department of Mathematics, City University of Hong Kong, Hong Kong. mazhou@cityu.edu.hk

The classical Kaczmarz algorithm [27] is designed on a Euclidean space $H = \mathbb{R}^d$, for solving a system of linear equations $Ax^0 = y$. Here, A is a matrix of size $m \times d$, $x^0 \in \mathbb{R}^d$ is an unknown vector, and $y = (y_1, \dots, y_m)^T \in \mathbb{R}^m$. Each row A_k of A is assumed normalized (otherwise, one normalizes A_k, together with the corresponding coordinate y_k of y, by scaling both of them with the factor $\|A_k\|^{-1}$). In [27], the Kaczmarz algorithm takes the form (2), where the constant step-size $\eta_k \equiv 1$ is used, the k-th row of A is employed as ξ_k, and the k-th coordinate of y is just used as y_k. After k exceeds m, the iteration continues, with the rows of A and the coordinates of y cyclically reused. Kaczmarz [27] shows that if the rows of A span the whole space $H = \mathbb{R}^d$, then Algorithm (2) converges. However, the speed of the convergence (3) is not guaranteed, and may depend on the order of the rows of the coefficient matrix A. See [43, 9] and the reference therein. RK is introduced to remove this dependence. In particular, Strohmer and Vershynin [43] show that if the random measurement vectors $\{\xi_k\}$ are independently drawn from the rows of A, and each row is drawn with probability proportional to the square of its norm (before the scaling we mentioned above), then

$$\mathbb{E}\left[\|r_{k+1}\|^2\right] \le \left(1 - \frac{1}{\|A\|_{\mathsf{F}}^2 \|A^{-1}\|_{\mathbb{R}^m \to \mathbb{R}^d}^2}\right)^k \|x^0\|^2, \tag{4}$$

where A^{-1} is the left inverse (assumed to exist), and $\|A\|_{\mathsf{F}}$ is the Frobenius norm of A. The quantity $\|A\|_{\mathsf{F}} \|A^{-1}\|_{\mathbb{R}^m \to \mathbb{R}^d}$, referred to as the scaled condition number [43], is bounded from below by the condition number $\kappa(A)$ of A (here $\kappa(A)$ is the ratio of the maximum singular value of A to the minimum singular value of A). Estimate (4) shows that RK converges exponentially. The case with general distributions of the vectors $\xi_i \in \mathbb{R}^d$ is studied by Chen and Powell [9], where almost sure exponential convergence is also derived.

It is easy to see that as the size of A increases, the base of the exponential rate (4) approaches 1. Then (4) only guarantees a very slow convergence speed. Furthermore, it is shown in [43] with an example that the bound (4) is sharp. So, the analysis in [43] does not apply to Hilbert spaces, and leaves a gap between the theory and applications, since a linear system with large scale is a typical scenario where the Kaczmarz algorithm is employed to outperform direct solvers. In this paper, we provide new analysis of RK (2) that works in Hilbert spaces. Our first motivation is to provide a condition number-free convergence estimate of (2) for large scale problems.

Another special case of Algorithm (2) is the online gradient descent method extensively studied for nonlinear regression learning [42, 49, 47, 48, 28, 8, 45, 21, 39, 23]. Let K be a Mercer kernel on a metric space X. That is, $K : X \times X \to \mathbb{R}$ is a continuous and symmetric function which is positive semi-definite (i.e., $\sum_{i,j=1}^m c_i c_j K(x_i, x_j) \ge 0$ for any integer $m \ge 1$, any $x_1, \dots, x_m \in X$ and $c_1, \dots, c_m \in \mathbb{R}$). Write $K_u(v) = K(u, v)$. The function space spanned by $\{K_u : u \in X\}$ with inner product induced by $\langle K_u, K_v \rangle_K := K(u, v)$ for any $u, v \in X$, after completion, is a (reproducing kernel) Hilbert space. Denote $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K, \|\cdot\|_K)$ this Hilbert space of functions. Consider the regression problem with observations $\{(u_i, y_i)\}_i \subset X \times \mathbb{R}$. The online gradient descent learning algorithm is defined with $f_1 = 0$, and then iteratively by

$$f_{k+1} = f_k + \eta_k (y_k - f_k(u_k)) K_{u_k}, \quad k \ge 1. \tag{5}$$

In the literature, a more general form $f_{k+1} = f_k + \eta_k[(y_k - f_k(u_k))K_{u_k} + \lambda f_k]$ of iteration is also studied with the parameter $\lambda \ge 0$ inspired by the Tikhonov regularization. Assume $K(u, u) = 1$ for any $u \in X$ (this is true for the widely used Gaussian kernel; while for a general kernel $\tilde{K}$, one could normalize it by defining $K(u, v) = \tilde{K}(u, v)/\sqrt{\tilde{K}(u, u)\tilde{K}(v, v)}$ to guarantee this assumption. The normalization is by itself a usual practice [41, 40]). Then the online algorithm (5) is unified to the form (2) by letting $\xi_k = K_{u_k}$ (where $K(u_k, u_k) = 1$ implies $\|\xi_k\|_K = 1$), and noting the reproducing property $f(u) = \langle f, K_u \rangle_K$ for any $f \in \mathcal{H}_K$ and $u \in X$.

The second motivation of this paper is to put Algorithm (5) under the general framework (2), of which our new analysis improves the state-of-the-art convergence estimate of (5) in the literature.

In this paper, we study RK with the following three settings of step-sizes.

1. **RK-CS**, RK with a *universal* constant step-size. In this setting, we use $\eta_k \equiv \eta \in (0, 2)$ in (2) for $k = 1, 2, \ldots$. So, the step-size is set as an absolute constant independent of the step count k. This is for studying the ideal scenario where observational noise could be ignored ($\sigma^2 = 0$). In particular, $\eta_k \equiv 1$ corresponds to the vanilla RK.

2. **RK-VS**, RK with *vanishing step-sizes*. Here, one lets $\eta_k \in (0, 1)$ decrease as k increases, for $k = 1, 2, \ldots$. The iteration continues to update the output vector (or the output function for kernel-based learning) with new data points. Same as RK-CS, in this setting Algorithm (2) does not have to terminate. For example, it is proved in [30] that in Euclidean spaces, (2) converges strongly if and only if $\eta_k \to 0$ and $\sum_k \eta_k = \infty$.

3. **RK-FH**, RK with *finite horizon* setting, i.e., with a constant step-size that depends on the known total sample size. In this setting, we assume the access to the size $m < \infty$ of training data. Algorithm (2) is designed to terminate after m iterations. Here we define a constant step-size $\eta_k = \eta(m) \in (0, 1)$ for $1 \le k \le m$, which depends (usually decreasingly) on m. For example, we shall study the setting (15) below of step-size in Theorem 2.6.

Note that Setting RK-CS is designed for the ideal noiseless model to explore the intrinsic behavior of RK. While Settings RK-VS and RK-FH are designed for real data that may be contaminated with noise.

The setting of step-sizes $0 < \eta_k < 2$ is also referred to as relaxation [7, 17, 21]. Many works in the literature suggest that with the presence of noise, RK-FH provides faster convergence than RK-VS does, though the former one is not truly online and has to terminate after finite iterations. We do not consider the design $1 < \eta_k < 2$ for RK-VS and RK-FH, and leave the discussion as future work.

The rest of the paper is organized as follows. In Section 2, we present our main results and some comparisons with the literature. Some key assumptions and technical notations are also discussed. Section 3 is devoted to RK-CS while Section 4 develops the theory for RK-VS and RK-FH.

2 Main Results and Discussions

We organize this section along the three settings of step-sizes. Theorem 2.2 through Example 2.4 study the noise-free setting with RK-CS, and the application to the classical RK in Euclidean spaces. Theorem 2.5 provides error bounds for the weak convergence of RK-VS. Theorem 2.6 through Corollary 2.10 provide convergence analysis for RK-FH with applications to kernel-based online learning.

Write $\mathbb{S} := \{x \in H : \|x\| = 1\}$ the unit sphere in H. Let ρ be a Borel probability measure on $\mathbb{S}$. Define

$$L = \mathbb{E}_{\xi \sim \rho}[\xi \otimes \xi]. \tag{6}$$

Here for any $u, v \in H$, we define $u \otimes v$ as the linear operator on H with $(u \otimes v)x := \langle v, x \rangle u$ for any $x \in H$. It is well known that L is positive semi-definite (thus self-adjoint), and of trace class (thus Hilbert-Schmidt and compact). In particular, for $s > 0$, L^s is defined through the spectral expansion of L, and $L^{-s}x$ denotes the preimage vector in the closure of $L(H)$ for any $x \in L^s(H)$. The existence of $L^{-s}x$ follows from that L is bounded and positive semi-definite. One proves the uniqueness of $L^{-s}x$ with the argument in the proof of Lemma B.1.

We provide a detailed treatment of these properties at the end of Section 1 for completeness. We use the following definition to specify the probability model of (1).

Definition 2.1 (Model $\mathbb{M}(\mathbb{S}, \rho, x^0, m, \sigma^2)$)**.** *We say that a sequence $\{(\xi_i, y_i)\}_{i=1}^m$ is drawn from Model $\mathbb{M}(\mathbb{S}, \rho, x^0, m, \sigma^2)$, if (i) $\{\xi_i\}_{i=1}^m$ are all drawn from $(\mathbb{S}, \rho)$; (ii) for each $1 \le i \le m$, y_i is defined by (1), where $\epsilon_i \in \mathbb{R}$ is a random variable with $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2 < \infty$; and (iii) the random variables $\xi_1, \dots, \xi_m, \epsilon_1, \dots, \epsilon_m$ are independent. We define Model $\mathbb{M}(\mathbb{S}, \rho, x^0, \infty, \sigma^2)$ in the same way except for infinite sequences $\{(\xi_i, y_i)\}_{i=1}^\infty$.*

Analysis under Setting RK-CS. First, we show that in the noise-free setting, and under the regularity assumption $x^0 \in L^{s_1}(H)$ on the unknown vector x^0, or the assumption $x' \in L^{s_2}(H)$ on some measurement vector x' (where $s_1, s_2 \in (0, 1/4]$), the expected weak error $\mathbb{E}[\langle r_{k+1}, x' \rangle^2]$ converges to zero in polynomial speed. We also give an example where the observations are contaminated with just arbitrarily small noise (i.e. $\sigma^2 > 0$ for Model $\mathbb{M}(\mathbb{S}, \rho, x^0, \infty, \sigma^2)$) yet one already has $\lim_{k\to\infty} \mathbb{E}[\|r_{k+1}\|^2] = \infty$. These results suggest that RK in Hilbert spaces (and thus in high dimensional Euclidean spaces) has a nature of weak convergence instead of strong convergence. This is consistent with the difficulties observed in the literature on developing strong convergence of RK in Hilbert spaces.

Theorem 2.2. *Consider an infinite sequence $\{(\xi_i, y_i)\}_{i=1}^\infty$ from Model $\mathbb{M}(\mathbb{S}, \rho, x^0, \infty, \sigma^2 = 0)$. Assume $x^0 \in L^{s_1}(H)$ with $s_1 \in (0, 1/4]$. Let $x' \in L^{s_2}(H)$ with $s_2 \in (0, 1/4]$. Setting $\eta_k \equiv \eta \in (0, 2)$ in Algorithm (2), we have*

$$\mathbb{E}\left[\langle x', r_{k+1} \rangle^2\right] \le \|L^{-s_2}x'\|^2 \|L^{-s_1}x^0\|^2 C_{\eta, 2s_1+2s_2} (\eta(2-\eta)k)^{-2s_1-2s_2}, \quad k \ge 1, \tag{7}$$

where $C_{\eta,s}$ for $0 < s \le 1$ is defined in (34) in the proof. In particular, $C_{\eta,s} = 1$ when $0 < \eta \le 1$.

If instead of the fixed vector x', a random measurement vector $\xi' \sim (\mathbb{S}, \rho)$ is employed, then

$$\mathbb{E}\left[\langle \xi', r_{k+1} \rangle^2\right] \le \sqrt{\mathrm{Tr}(L)} \|L^{-s_1}x^0\|^2 C_{\eta, 2s_1+1/2} (\eta(2-\eta)k)^{-2s_1-\frac{1}{2}}, \quad k \ge 1. \tag{8}$$

Note that in this paper, all the expectations $\mathbb{E}[\cdot]$ are taken with respect to all the randomness specified in the context, unless otherwise specified. For example, in (7), the expectation is taken with respect to the randomness of x_{k+1}, coming from the sample subset $\{(\xi_i, y_i)\}_{i=1}^k$ through (2); In (8), one takes the expectation with respect to both x_{k+1} and ξ'.

Here we briefly explain the ideas of proving Theorem 2.2, and leave the detailed treatments in Section 3. First, we rewrite the expected error as $\mathbb{E}[\langle x', r_{k+1} \rangle^2] = \langle x' \otimes x', \mathcal{R}_{k+1} \rangle_{\mathsf{HS}}$, where $\mathcal{R}_{k+1} = \mathbb{E}[r_{k+1} \otimes r_{k+1}]$ is an operator, and $\langle \cdot, \cdot \rangle_{\mathsf{HS}}$ denotes the Hilbert-Schmidt inner product. From the noiseless assumption, $\mathcal{R}_{k+1} = Q_\eta(\mathcal{R}_k) = Q_\eta^k(\mathcal{R}_1) = Q_\eta^k(x^0 \otimes x^0)$. The main innovation in our analysis is that we manage to factor out some powers of $(I - Q_\eta)$ from $x' \otimes x'$ and $x^0 \otimes x^0$ respectively, to join Q_η^k and form a vanishing operator $Q_\eta^k(I - Q_\eta)^{2s_1+2s_2}$, which leads to the decay of error $\mathbb{E}[\langle x', r_{k+1} \rangle^2]$. When x' is replaced by the random vector ξ', we use $\mathbb{E}[\xi' \otimes \xi'] = L$, of which we transform a half $L^{1/2}$ into $(I - Q_\eta)^{1/2}$ and obtain the corresponding vanishing operator $Q_\eta^k(I - Q_\eta)^{2s_1+\frac{1}{2}}$.

As an application, consider the linear equation $Ax^0 = b$ with $A \in \mathbb{R}^{m \times d}$. Assume the rows $A_1, \dots, A_m$ of A are normalized $\|A_1\| = \ldots = \|A_m\| = 1$ For this example let ρ be the uniform distribution on the rows of A, $\rho(A_i) = \frac{1}{m}$ for $1 \le i \le m$. This is consistent with the model in [43] and reduces the definition (6) to $L = \sum_{i=1}^m \frac{1}{m} A_i \otimes A_i = \frac{1}{m} A^T A$. Let $y_k = \langle \xi_k, x^0 \rangle$. One applies the

RK iteration with $\eta_k \equiv \eta \in (0,2)$. Theorem 2.2 implies the following Corollary 2.3, which provides condition number-free polynomial convergence rates for RK in Euclidean spaces.

Corollary 2.3. *For the above setting of linear equation $Ax^0 = b$ with $A \in \mathbb{R}^{m\times d}$, if $x^0 = \left(\frac{1}{m}A^TA\right)^{s_1} x^0_*$ for some $s_1 \in (0, 1/4]$ and $x^0_* \in \mathbb{R}^d$, and $x' = \left(\frac{1}{m}A^TA\right)^{s_2} x'_*$ for some $s_2 \in (0, 1/4]$ and $x'_* \in \mathbb{R}^d$, then*

$$\mathbb{E}\left[\langle x', r_{k+1}\rangle^2\right] \le \|x^0_*\|^2\|x'_*\|^2 C_{\eta,2s_1+2s_2}(\eta(2-\eta)k)^{-2s_1-2s_2}. \tag{9}$$

In particular, when A^TA is invertible,

$$\mathbb{E}\left[\|r_{k+1}\|^2\right] \le d\|x^0_*\|^2 C_{\eta,2s_1}(\eta(2-\eta)k)^{-2s_1}. \tag{10}$$

□

Bound (9) is a direct appliation of Theorem 2.2. To see (10), let $\{e_i\}_{i=1}^d$ be an orthonormal basis of $\mathbb{R}^d$. One applies (9) with $x'_* = (\frac{1}{m}A^TA)^{-s_2}e_i$ for $0 < s_2 \le 1/4$ to obtain

$$\mathbb{E}\left[\langle e_i, r_{k+1}\rangle^2\right] \le \|x^0_*\|^2\|(\frac{1}{m}A^TA)^{-s_2}e_i\|^2 C_{\eta,2s_1+2s_2}(\eta(2-\eta)k)^{-2s_1-2s_2}. \tag{11}$$

Recall $s_1 > 0$. One takes the limit of (11) as $s_2 \downarrow 0$, and adds up the limits for $i = 1, \dots, d$, to obtain (10).

The following example shows that in general, a strong convergence of RK-CS in Hilbert spaces cannot be expected in practice with noisy data. Combined with the theorem above, this example reveals the weak convergence nature of RK-CS.

Example 2.4. *Let $\{e_i\}_{i=1}^\infty$ be an orthonormal basis of H. Let $q_1 \ge q_2 \ge \ldots > 0$ and $\sum_{i=1}^\infty q_i = 1$. Let ρ be a discrete probability distribution such that $\rho(e_i) = q_i$. Assume $\sigma^2 > 0$ for Model $\mathbb{M}(\mathbb{S}, \rho, x^0, \infty, \sigma^2)$ and set $\eta_k \equiv \eta \in (0, 1]$ for Algorithm (2). Then $\lim_{k\to\infty} \mathbb{E}\left[\|r_{k+1}\|^2\right] = \infty$, and this limit is independent of x^0 (in particular, it holds true even for $x^0 = 0$). Meanwhile, if $x^0 \in L^{s_1}(H)$ with $s_1 > 0$ and $x' \in L^{s_2}(H)$ with $s_2 > 0$, one has that $\limsup_{k\to\infty} \mathbb{E}[\langle x', r_{k+1}\rangle^2] = O(\eta)$ as $\eta \to 0^+$.*

The proofs of the claims of Example 2.4 are postponed to Appendix A.

It remains an interesting open question whether there is any strong convergence of RK-CS in general Hilbert spaces. We note that Griebel and Oswald [19, 20] provide analysis for strong convergence of the *Schwarz Iterative Methods* for elliptic variational problems with coercive Hermitian forms in Hilbert spaces. In fact, the setting of finite or countable space splitting in [19, 20] well covers the special case of Example 2.4 (with $\sigma^2 = 0$), and yields the strong convergence. So, the question we raise here is partly affirmatively answered. Nonetheless, the analysis in [19, 20] does not apply to the general scenarios where the support of ρ is not included in the union of countable finite-dimensional subspaces, for example, the most interesting application in online learning (5) with Gaussian or Sobolev kernels. In addition, Example 2.4 shows that the strong convergence, even if it exists, could be spoiled by noise.

Berthier, Bach and Gaillard [5] derive an elegant strong convergence bound $\mathbb{E}[\|r_{k+1}\|^2] \le O(k^{-2s_1})$ and a weak convergence bound $\min_{1\le j\le k} \mathbb{E}[\langle \xi', r_{j+1}\rangle^2] \le O(k^{-1-2s_1})$ of (2) applicable to the noiseless model $\mathbb{M}(\mathbb{S}, \rho, x^0, \infty, \sigma^2 = 0)$ with setting RK-CS and the assumption

$$\left\|L^{-s_1}\xi\right\| \le C(s_1), \quad \text{almost surely for } \xi \sim \rho, \tag{12}$$

where $s_1 \ge 0$ is the index we use in Theorem 2.2 and $C(s_1) < \infty$ is a constant. The assumption (12) with $s_1 > 0$ is nonetheless strong and even in the simple model we study in Example 2.4 implies that L has only a finite rank. In particular, we have removed the assumption (12) in this paper. Another improvement of our analysis is that the weak convergence bound in [5] is provided for the *smallest error* along the whole path of iterations, while Theorem 2.2 provides error bound for the *last iteration.* Note that due to the assumption (12), the rates in (7) are not directly comparable with the results in [5]. The analysis in Varre et al. [44] is also applicable to the noiseless model $\mathbb{M}(\mathbb{S}, \rho, x^0, \infty, \sigma^2 = 0)$. The expected weak error $\mathbb{E}[\langle \xi', r_{k+1}\rangle^2]$ with random measurement $\xi' \sim (\mathbb{S}, \rho)$ is bounded as $O(k^{-1}\log k)$ in [44, Theorem 1] by adopting the finite horizon setting (fixing $\eta \log m$ as a constant). Compared with this rate, our bound (8) does not require the finite horizon setting, and achieves a better rate $O(k^{-1})$ when $s_1 = 1/2$. The bound $O(k^{-1})$ is further derived in [44, Theorem 2] without the finite horizon setting, but under another assumption that $cL - \mathbb{E}[\langle \xi', \log(L^{-1})\xi'\rangle \xi' \otimes \xi']$ is a positive operator for some finite constant $c > 0$.

Analysis under Setting RK-VS. Suppose we have an infinite sequence $\{(\xi_i, y_i)\}_{i=1}^{\infty}$ of observations from Model $\mathbb{M}(\mathbb{S}, \rho, x^0, \infty, \sigma^2)$ with $\sigma^2 \geq 0$, based on which we apply Algorithm (2) for the sequence $\{x_k\}_{k=2}^{\infty}$ of output vectors. Suppose the step-sizes are set as $\eta_k = \eta_1 k^{-\omega}$ for $k \geq 2$ and $0 < \eta_1 \leq \sqrt{1-(2\omega)^{-1}}$. Theorem 2.5 gives the rate of weak convergence of $\{x_k\}$ to x^0.

Let $\mathscr{J}_1(H)$ denote the space of all the trace-class operators on H.

Theorem 2.5. *Assume $x^0 \in L^{s_1}(H)$ for some $s_1 > 0$. Let $x' \in L^{s_2}(H)$ be a measurement vector with $s_2 \geq 1/4$, and $s_1 + s_2 > 1/2$. Set $\omega = \frac{2s_1+2s_2}{1+2s_1+2s_2}$. For any $k \geq 1$, we have*

$$\mathbb{E}\left[\langle x', r_{k+1}\rangle^2\right] \leq C_1 (k+1)^{\max\{-\omega, -(2s_2+\frac{1}{2})(1-\omega)\}}, \tag{13}$$

where C_1 is a constant independent of k, and it will be specified in (62) at the end of the proof.

If we replace x' by a random vector $\xi' \sim (\mathbb{S}, \rho)$, assume $L^{2-4s_0} \in \mathscr{J}_1(H)$ for some $s_0 \in [1/4, 1/2)$, $s_1 + s_0 > 1/2$, and set $\omega = \frac{2s_1+2s_0}{1+2s_1+2s_0}$, then for $k \geq 1$,

$$\mathbb{E}\left[\langle \xi', r_{k+1}\rangle^2\right] \leq C_1' (k+1)^{\max\{-\omega, -(2s_0+\frac{1}{2})(1-\omega)\}}, \tag{14}$$

where C_1' is a constant independent of k and it will be specified in (63) at the end of the proof.

Guo and Shi [24] provide rates of strong convergence for RK-VS in reproducing kernel Hilbert spaces.

Analysis under Setting RK-FH. Theorem 2.6 below provides the rates of strong convergence of relaxed RK.

Theorem 2.6. *Consider a finite sample $\{(\xi_i, y_i)\}_{i=1}^{m}$ from Model $\mathbb{M}(\mathbb{S}, \rho, x^0, m, \sigma^2)$ with $\sigma^2 \geq 0$ and $2 \leq m < \infty$. Assume $x^0 \in L^{s_1}(H)$ for some $s_1 > 0$, and $L^{s_*} \in \mathscr{J}_1(H)$ for some $0 < s_* \leq 1$. Set for $1 \leq k \leq m$ that*

$$\eta_k \equiv \eta = \begin{cases} m^{-(2s_1+s_*)/(1+2s_1+s_*)}, & \text{if } s_1 + s_* \geq 1, \\ m^{-(1+s_1)/(2+s_1)}, & \text{if } s_1 + s_* < 1. \end{cases} \tag{15}$$

Then,

$$\mathbb{E}\left[\|r_{m+1}\|^2\right] \leq C_2 \begin{cases} m^{-2s_1/(1+2s_1+s_*)}, & \text{if } s_1 + s_* \geq 1, \\ m^{-2s_1/(2+s_1)}, & \text{if } s_1 + s_* < 1, \end{cases} \tag{16}$$

where C_2 is independent of m, and it will be specified in (48) at the end of the proof.

As a direct application of Theorem 2.6, now we consider Algorithm (5), which is a kernel-based online regression learning algorithm without regularization. Let X be a compact metric space with a

Borel probability measure μ. Let K be a Mercer kernel on X with $K(x,x)=1$ for any $x \in X$, and $(\mathcal{H}_K, \langle\cdot,\cdot\rangle_K, \|\cdot\|_K)$ be the induced reproducing kernel Hilbert space. The following definition specifies the probability model on which we develop the convergence analysis for Algorithm (5).

Definition 2.7 (Model $\tilde{\mathbb{M}}(X,\mu,f^0,m,\sigma^2)$)**.** *We say that a sequence $\{(u_i,y_i)\}_{i=1}^m$ is drawn from Model $\tilde{\mathbb{M}}(X,\mu,f^0,m,\sigma^2)$, if (i) $\{u_i\}_{i=1}^m$ are all drawn from (X,μ); (ii) for each $1 \le i \le m$, $y_i = f^0(u_i)+\epsilon_i$, where ϵ_i is a random variable with $\mathbb{E}[\epsilon_i]=0$ and $\mathrm{Var}(\epsilon_i)=\sigma^2<\infty$; and (iii) the random variables $u_1,\ldots,u_m,\epsilon_1,\ldots,\epsilon_m$ are independent.*

For Model $\tilde{\mathbb{M}}(X,\mu,f^0,m,\sigma^2)$, the relation $\xi = K_x$ pushes the distribution μ on X to the distribution ρ on $\mathbb{S} \subset H = \mathcal{H}_K$, and reduces the definition (6) to a particular form $L = L_K := \mathbb{E}_{x\sim\mu}[K_x \otimes K_x]$. The following corollary for Algorithm (5), as a direct application of Theorem 2.6, improves the analysis in [48].

Corollary 2.8. *Suppose f_{m+1} is the output function of Algorithm (5) after $2 \le m < \infty$ iterations, with data $\{(u_i,y_i)\}_{i=1}^m$ drawn from Model $\tilde{\mathbb{M}}(X,\mu,f^0,m,\sigma^2)$. Assume $\sigma^2 \ge 0$, $f^0 \in L_K^{s_1}(\mathcal{H}_K)$ for some $s_1>0$, and $L_K^{s_*} \in \mathcal{J}_1(\mathcal{H}_K)$ for some $0<s_*\le 1$. Set η_k as (15). Then,*

$$\mathbb{E}\left[\left\|f_{m+1}-f^0\right\|_K^2\right] \le C_2 \begin{cases} m^{-2s_1/(1+2s_1+s_*)}, & \text{if } s_1+s_* \ge 1, \\ m^{-2s_1/(2+s_1)}, & \text{if } s_1+s_*<1, \end{cases} \tag{17}$$

where C_2 is defined in Theorem 2.6. □

When the eigenvalues $\{\lambda_i(L_K)\}$ (arranged in non-increasing order) of L_K decay polynomially $\lambda_i(L_K) \sim i^{-1/s_*}$ for some $0<s_*<1$ (i.e., $c_1 i^{-1/s_*} \le \lambda_i(L_K) \le c_2 i^{-1/s_*}$ for some positive numbers c_1 and c_2, and every i), the minimax optimal rates $O(m^{-2s_1/(1+2s_1+s_*)})$ of strong convergence for recovering x^0 is derived in [22, 6]. Note that

$$L^{s_*} \in \mathcal{J}_1(H) \overset{\mathsf{A}}{\Longrightarrow} \lambda_i(L) = O(i^{-1/s_*}) \overset{\mathsf{B}}{\Longrightarrow} L^{s_*+\epsilon} \in \mathcal{J}_1(H), \quad \text{for any } 0<s_*\le 1, \epsilon>0. \tag{18}$$

In fact, since $\{\lambda_i(L)\}$ is non-increasing, one has $i\lambda_i(L)^{s_*} \le \mathrm{Tr}(L^{s_*})$, which yields Implication A in (18). Implication B is verified by $\mathrm{Tr}(L^{s_*+\epsilon}) \le \sum \lambda_i(L)^{s_*+\epsilon} \le \sum \left(\frac{1}{i}\right)^{(s_*+\epsilon)/s_*} < \infty$. Therefore, when $s_1+s_*\ge 1$, Bound (16) is arbitrarily close to the above minimax optimal rate. Also, by adopting the capacity assumption $L^{s_*} \in \mathcal{J}_1(H)$, Bound (16) improves the learning rate $O(m^{-2s_1/(2s_2+2)})$ obtained in [48, Theorem 6].

Now we study the weak convergence of RK in the finite horizon setting, RK-FH.

Theorem 2.9. *Assume $x^0 \in L^{s_1}(H)$ for some $s_1>0$. Let $x' \in L^{s_2}(H)$ be a measurement vector with $s_2>0$. Let x_{m+1} be the output of Algorithm (2) with data $\{(\xi_i,y_i)\}_{i=1}^m$ drawn from Model $\mathbb{M}(\mathbb{S},\rho,x^0,m,\sigma^2)$ with $2\le m<\infty$ and $\sigma^2\ge 0$, and the step-sizes $\eta_k \equiv m^{-\omega}$ with*

$$\omega = \begin{cases} \frac{1+4s_1}{3+4s_1}, & \text{if } 0<s_2<1/4, \\ \frac{2s_1+2s_2}{1+2s_1+2s_2}, & \text{if } s_2 \ge 1/4. \end{cases} \tag{19}$$

We assume $\omega>1/2$ (which means to assume $s_1>1/4$ when $0<s_2<1/4$, and to assume $s_1+s_2>1/2$ when $s_2\ge 1/4$). Then

$$\mathbb{E}\left[\langle x', r_{m+1}\rangle^2\right] \le C_3 \begin{cases} m^{-(4s_1+4s_2)/(3+4s_1)}, & \text{if } 0<s_2<1/4, \\ m^{-(1+4s_1)/(3+4s_1)}\log(m+1), & \text{if } s_2=1/4, \\ m^{-(2s_1+2s_2)/(1+2s_1+2s_2)}, & \text{if } s_2>1/4, \end{cases} \tag{20}$$

where C_3 is a constant independent of m, and it will be specified in (58) at the end of the proof.

If we replace x' by the random vector $\xi' \sim (\mathbb{S}, \rho)$, assume $L^{2-4s_0} \in \mathcal{J}_1(H)$ for some $s_0 \in [1/4, 1/2)$ with $s_1 + s_0 > 1/2$, and set $\eta_k \equiv m^{-\omega'}$ with $\omega' = \frac{2s_1+2s_0}{1+2s_1+2s_0}$, then

$$\mathbb{E}\left[\langle \xi', r_{m+1}\rangle^2\right] \leq C_3' \begin{cases} m^{-(1+4s_1)/(3+4s_1)} \log(m+1), & \text{if } s_0 = 1/4, \\ m^{-(2s_1+2s_0)/(1+2s_1+2s_0)}, & \text{if } s_0 \in (1/4, 1/2), \end{cases} \tag{21}$$

where C_3' is a constant independent of m, and it will be specified in (59) at the end of the proof.

The following corollary is a direct application of Theorem 2.9 to Algorithm (5). Denote $\|f\|_\mu^2 = \int_X f(x)^2 d\mu(x)$ for all square μ-integrable functions f on X.

Corollary 2.10. *Assume $f^0 \in L_K^{s_1}(\mathcal{H}_K)$ for some $s_1 > 0$ and $L_K^{2-4s_0} \in \mathcal{J}_1(\mathcal{H}_K)$ for some $s_0 \in [1/4, 1/2)$. Suppose f_{m+1} is the output function of Algorithm (5) with data $\{(u_i, y_i)\}_{i=1}^m$ drawn from Model $\tilde{\mathbb{M}}(X, \mu, f^0, m, \sigma^2)$ with $2 \leq m < \infty$ and $\sigma^2 \geq 0$, and step-sizes $\eta_k \equiv m^{-(2s_1+2s_0)/(1+2s_1+2s_0)}$. Then*

$$\mathbb{E}\left[\|f_{m+1} - f^0\|_\mu^2\right] \leq C_3' \begin{cases} m^{-(1+4s_1)/(3+4s_1)} \log(m+1), & \text{if } s_0 = 1/4, \\ m^{-(2s_1+2s_0)/(1+2s_1+2s_0)}, & \text{if } s_0 \in (1/4, 1/2), \end{cases} \tag{22}$$

where C_3' is defined in Theorem 2.9. □

The estimates (10), (16), and (22) (we exclude (17) from the discussion since it is just a direct application of (16)), though all on squared norms of error vectors (functions), are derived in different ways. In $\mathbb{R}^d$, (10) is obtained by applying (9) for d times to basis vectors respectively, $\mathbb{E}[\|r_{k+1}\|^2] = \sum_{j=1}^d \mathbb{E}[\langle e_j, r_{k+1}\rangle^2]$, and we let $s_2 \downarrow 0$ to avoid involving large factors related to the condition number of $\frac{1}{m}A^T A$. Bound (16) is derived via $\mathbb{E}[\|r_{m+1}\|^2] = \mathbb{E}\mathrm{Tr}(r_{m+1} \otimes r_{m+1}) = \mathrm{Tr}(\mathcal{R}_{m+1})$, and the key error decomposition is given in (39). Bound (22) is derived from (21) through the reproducing property $\int_X (f_{m+1}(x) - f^0(x))^2 d\mu(x) = \mathbb{E}_{x\sim\mu}[\langle K_x, f_{m+1} - f^0\rangle_K^2]$.

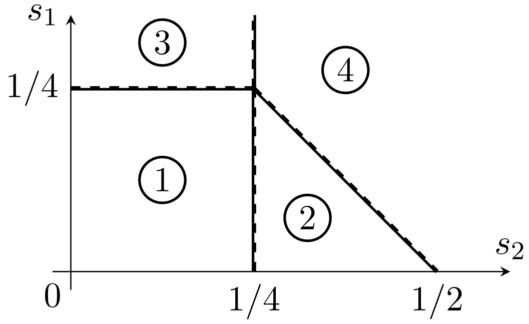

Figure 1: To compare Theorems 2.2, 2.9 and 2.5, we divide the quadrant of $s_1, s_2 > 0$ into four areas: Area 1 ($0 < s_1, s_2 \leq 1/4$), Area 2 ($s_1 > 0$, $s_2 > 1/2$, and $s_1 + s_2 \leq 1/2$), Area 3 ($0 < s_2 < 1/4$ and $s_1 > 1/4$), and Area 4 ($s_1 > 0$, $s_2 \geq 1/4$, and $s_1 + s_2 > 1/2$). Detailed discussion is provided in Remark 2.11.

Remark 2.11. *We compare the convergence estimates (7), (20) and (13) despite their different algorithm designs and the assumptions on noise levels (i.e., $\sigma^2 = 0$ and $\sigma^2 \geq 0$). The space of s_1 (index of the target vector regularity) and s_2 (index of the measurement vector regularity) is divided into four areas as visualized in Figure 1. Bound (7) in Theorem 2.2 covers Area 1. When $s_1 > 1/4$,* *since L is bounded, $x^0 \in L^{s_1}(H)$ implies $x^0 \in L^{1/4}(H)$. Similarly, when $s_2 > 1/4$, $x' \in L^{s_2}(H)$ implies $x' \in L^{1/4}(H)$. So, for Areas 2, 3, and 4, Bound (7) still holds true but saturates at the boundary of Area 1. It is an interesting question whether this saturation is intrinsic to the algorithm or just because of the limitation of our analysis. Bound (20) in Theorem 2.9 covers Areas 3 and 4, while Bound (13) in Theorem 2.5 covers only Area 4. In Area 4, despite the difference that RK-FH has to stop after m iterations while RK-VS does not have to terminate, we simply set $k = m$ and consider $s_2 > 1/4$ to find that (20) provides rate $O(m^{-\omega})$ (where $\omega = \frac{2s_1+2s_2}{1+2s_1+2s_2}$) and (13) provides rate $O(m^{\max\{-\omega,-(2s_2+\frac{1}{2})(1-\omega)\}})$, while $-\omega < -(2s_2 + \frac{1}{2})(1-\omega)$ (i.e., the rate in (13) is slower than that in (20)) if and only if $s_1 > 1/4$. It would be interesting if the analysis of RK-VS and RK-FH can be improved to cover all the four areas.*

We point out that because of observational noise, the best provable convergence rates for settings RK-VS and RK-FH are attained with necessary dependence of the step-sizes on the capacity of hypothesis space, and the regularity of the unknown vector x^0. This dependence is also reported in some related works [24, 48]. Nonetheless, in this paper we also provide analysis for general designs of step-sizes, in Theorems 4.4 and 4.6, for strong and weak convergences, respectively.

We provide brief discussions on Bounds (8), (21) and (14), compared with (7), (20) and (13), respectively. Recall that $\mathcal{J}_1(H)$ denotes the space of trace class operators, and $\mathcal{J}_2(H)$ denotes the space of Hilbert-Schmidt operators. For (21), the regularity $L^{2-4s_0} \in \mathcal{J}_1(H)$ is used to bound L^{1-2s_0} in $\mathcal{J}_2(H)$ for $\mathbb{E}[\langle \xi', r_{m+1}\rangle^2] = \mathrm{Tr}(L\mathcal{R}_{m+1}) = \mathrm{Tr}(L^{1-2s_0}(L^{s_0}\mathcal{R}_{m+1}L^{s_0}))$. While the regularity $x' \in L^{s_2}(H)$ is exploited similarly for (20), $\mathbb{E}[\langle x', r_{k+1}\rangle^2] = \mathrm{Tr}([(L^{-s_2}x') \otimes (L^{-s_2}x')](L^{s_2}\mathcal{R}_{k+1}L^{s_2}))$. This explains the similar roles s_2 and s_0 play in (20) and (21) respectively. The relation between Bounds (13) and (14) are similar, because Theorems 2.9 and 2.5 are both corollaries of Theorem 4.6. Nonetheless, the assumption $L^{2-4s_0} \in \mathcal{J}_1(H)$ does not help to improve (8) because of the saturation we mentioned in Remark 2.11. See the proof of Theorem 2.2 for details.

There are research works studying general formulations of (randomized) Kaczmarz algorithms [18], mini-batch iteration [29, 10], and shrinkage and sparsity [38]. In Lin and Zhou [30], RK is studied from the learning theory perspective. Kaczmarz algorithms belong to a broad class of algorithms called the *Method of Alternating Projections* (MAP). See Wiener [46] and Halperin [25]. See also [15, Chapter 3] and [17, Chapter 7] for comprehensive treatments, and [35, 11]. Kaczmarz algorithms are closely related to the *Schwarz Iterative Methods*, and the *Alternating Directions Method* (ADM). See [19, 20, 36, 12, 17]. In computerized tomography, (2) belongs to a large family of algorithms called *Algebraic Reconstruction Techniques* (ART). See [26, Chapter 11]. The iteration (2) could be seen as a special case of the general *Projecting onto Convex Sets* (POCS) algorithm. See [3], [15, Chapter 5], and the reference therein. There is a large literature studying the acceleration of iterative methods. For example, the seminal Nesterov accelerated gradient descent [34], the *Accelerated Randomized Kaczmarz* (Liu and Wright, [31]) which is designed in Euclidean spaces and needs finite condition number of the coefficient matrix A, and the acceleration by *averaging* (Dieuleveut and Bach, [13]) which achieves optimal convergence rates. See also [33, 1, 2, 16, 24]. It would be interesting to apply our analysis to stochastic gradient descent algorithms in deep learning [51, 52].

Discussions on Some Notations and Assumptions. Recall $L = \mathbb{E}_{\xi\sim\rho}(\xi \otimes \xi)$ as defined in (6). Obviously for any $u, v \in H$, $u \otimes v$ is a bounded linear operator. Furthermore, $u \otimes v$ is a Hilbert-Schmidt operator. In this paper, we let $\mathcal{J}_2(H)$ denote the space of all the Hilbert-Schmidt operators on H, with inner product $\langle\cdot,\cdot\rangle_{\mathsf{HS}}$ and norm $\|\cdot\|_{\mathsf{HS}}$. Since H is separable, so is $\mathcal{J}_2(H)$. From $\|\xi\| = 1$, one has $\|\xi \otimes \xi\|_{\mathsf{HS}} = 1$, so the mean $\mathbb{E}_{\xi\sim\rho}[\xi \otimes \xi]$ is well defined by the Bochner's integral and $\|L\|_{\mathsf{HS}} \le \mathbb{E}_{\xi\sim\rho}[\|\xi \otimes \xi\|_{\mathsf{HS}}] = 1$. So, L is Hilbert-Schmidt, and is thus compact. See [14, Section

XI.6] for a comprehensive treatment of Hilbert-Schmidt operators, and [50, Section V.5] for a neat introduction of Bochner's integrals.

Let $\mathcal{I}$ be a set of indices: when $\dim(H) = d < \infty$, write $\mathcal{I} := \{1, 2, \dots, d\}$, and when $\dim(H) = \infty$, let $\mathcal{I}$ denote the set of all the positive integers.

For any $u, v \in H$, $\langle Lu, v\rangle = \mathbb{E}_{\xi\sim\rho}[\langle u, \xi\rangle \langle \xi, v\rangle] = \langle u, Lv\rangle$, so L is self-adjoint. In particular, $\langle Lu, u\rangle \geq 0$, so L is positive semi-definite. Therefore, we can write $\{(\lambda_i, \phi_i)\}_{i\in\mathcal{I}}$ as the eigensystem of L, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, and $\{\phi_i\}_{i\in\mathcal{I}}$ form an orthonormal basis of H. So $\sum_{i\in\mathcal{I}} \langle \phi_i, \xi\rangle^2 = \|\xi\|^2 = 1$ almost surely for $\xi \sim (\mathbb{S}, \rho)$. We have

$$\sum_{i\in\mathcal{I}} \lambda_i = \sum_{i\in\mathcal{I}} \langle \phi_i, L\phi_i\rangle = \mathbb{E}_{\xi\sim\rho}\left[\sum_{i\in\mathcal{I}} \langle \phi_i, \xi\rangle^2\right] = 1.$$

Therefore $L \in \mathcal{J}_1(H)$ and $\mathrm{Tr}(L) = 1$. Recall that $\mathcal{J}_1(H)$ denotes the set of trace class operators on H. For the notion of trace class operators, see [14, Section XI.9]. For $s \in \mathbb{R}$, define

$$L^s = \sum_{\lambda_i > 0} \lambda_i^s \phi_i \otimes \phi_i.$$

So, L^0 is the orthogonal projection onto the closure of $L(H)$, and when $s < 0$, L^s may only be defined on a subspace of H.

3 Polynomial Rates of Convergence in the Noise-free Setting

In the following we let I denote the identity operator, of which the domain is inferred from the context. For bounded self-adjoint operators A and B on a Hilbert space, we write $A \preceq B$ (or $B \succeq A$) if $B - A$ is positive semi-definite.

The main target of this section is to prove Theorem 2.2. The idea is to expand $\mathbb{E}[\langle x', r_{k+1}\rangle^2]$ as (for details see (35))

$$\left\langle (I - Q_\eta)^{-2s_2}(x' \otimes x'), \left[Q_\eta^k (I - Q_\eta)^{2s_1+2s_2}\right] (I - Q_\eta)^{-2s_1}(x^0 \otimes x^0)\right\rangle_{\mathsf{HS}},$$

where Q_η is a bounded linear operator on $\mathcal{J}_2(H)$ defined by (27) below. Then, we show the boundedness of $(I - Q_\eta)^{-2s_2}(x' \otimes x')$ and $(I - Q_\eta)^{-2s_1}(x^0 \otimes x^0)$ by the regularity assumptions of x' and x^0, respectively. Finally, we show the decay of the operator $Q_\eta^k (I - Q_\eta)^{2s_1+2s_2}$ by applying the spectral theorem with the polynomial $x^k(1 - x)^{2s_1+2s_2}$ on $\min\{0, 1 - \eta\} \leq x \leq 1$, as $k \to \infty$.

Now we consider the convergence of Algorithm (2) in noise-free setting. That is, we assume $\sigma^2 = 0$. So, the uncertainty of Algorithm (2) all comes from the randomness of ξ_k's.

Recall $r_k = x_k - x^0$, as the error vector after the $(k-1)$-th iteration. For $k \geq 1$,

$$r_{k+1} = x_k - x^0 + \eta_k(y_k - \langle \xi_k, x_k\rangle)\xi_k = r_k - \eta_k \left\langle \xi_k, x_k - x^0\right\rangle \xi_k = (I - \eta_k P_k) r_k, \tag{23}$$

where $P_k = \xi_k \otimes \xi_k$. We repeat the iteration (23) to give

$$r_{k+1} = (I - \eta_k P_k) \cdots (I - \eta_1 P_1)(-x^0). \tag{24}$$

Let $\xi \in \mathbb{S}$ and write $P = P_\xi = \xi \otimes \xi$ the associated rank-one orthogonal projection. We claim that

$$\|(I - \eta P)u\| \leq \|u\|, \quad \text{for any } u \in H \text{ and } 0 < \eta < 2. \tag{25}$$

To see this, write $u = u_\xi + u_\perp$ with $u_\xi = Pu$ and $u_\perp = (I - P)u$. Since $|1 - \eta| < 1$, $\|(I - \eta P)u\|^2 = \|u_\perp + (1 - \eta)u_\xi\|^2 = \|u_\perp\|^2 + (1 - \eta)^2\|u_\xi\|^2 \leq \|u\|^2$.

For any $B \in \mathcal{J}_2(H)$, recall that $\|B\|_{\mathsf{HS}}^2 = \sum_{i\in\mathcal{I}} \|Be_i\|^2$, where $\{e_i\}_{i\in\mathcal{I}}$ is any orthonormal basis of H. From (25), we use $e_1 = \xi$ to obtain that for any $0 < \eta < 2$,

$$\begin{aligned}\|(I-\eta P)B(I-\eta P)\|_{\mathsf{HS}}^2 &= (1-\eta)^2 \|(I-\eta P)Be_1\|^2 + \sum_{i\in\mathcal{I}\setminus\{1\}} \|(I-\eta P)Be_i\|^2 \\ &\le \sum_{i\in\mathcal{I}} \|(I-\eta P)Be_i\|^2 \le \sum_{i\in\mathcal{I}} \|Be_i\|^2 = \|B\|_{\mathsf{HS}}^2. \end{aligned} \tag{26}$$

Therefore, for any $0 < \eta < 2$, we can define the linear operator $Q_\eta : \mathcal{J}_2(H) \to \mathcal{J}_2(H)$ through Bochner's integral,

$$Q_\eta(B) = \mathbb{E}_{\xi\sim\rho}\left[(I-\eta P)B(I-\eta P)\right], \quad B \in \mathcal{J}_2(H). \tag{27}$$

The estimate (26) shows that

$$\|Q_\eta\|_{\mathcal{J}_2(H)\to\mathcal{J}_2(H)} \le 1. \tag{28}$$

For any $B_1 \in \mathcal{J}_2(H)$,

$$\begin{aligned}\langle Q_\eta(B), B_1\rangle_{\mathsf{HS}} &= \mathbb{E}\mathrm{Tr}\left((I-\eta P)B(I-\eta P)B_1^T\right) \\ = \mathrm{Tr}&\left(B\left[\mathbb{E}(I-\eta P)B_1(I-\eta P)\right]^T\right) = \langle B, Q_\eta(B_1)\rangle_{\mathsf{HS}}. \end{aligned} \tag{29}$$

So Q_η is self-adjoint.

Denote $\sigma(Q_\eta)$ the spectrum of Q_η. The following lemma provides an estimate for $\sigma(Q_\eta)$.

Lemma 3.1. *When $0 < \eta \le 1$, $\sigma(Q_\eta) \subset [(1-\eta)^2, 1]$. When $1 < \eta < 2$, $\sigma(Q_\eta) \subset [1-\eta, 1]$.*

Proof. The upper bound 1 is guaranteed by (28).

When $0 < \eta \le 1$, we need only to prove that $Q_\eta \succeq (1-\eta)^2 I$. Let $\{e_i\}_{i\in\mathcal{I}}$ be an orthonormal basis of H with $e_1 = \xi$. Then,

$$\begin{aligned}\mathrm{Tr}(BB^T - PBPB^TP) &= \sum_{i\in\mathcal{I}} \left\langle e_i, (BB^T - PBPB^TP)e_i\right\rangle \\ &= \left\|B^Te_1\right\|^2 - \left\|PB^Te_1\right\|^2 + \sum_{i\in\mathcal{I}\setminus\{1\}} \left\|B^Te_i\right\|^2 \ge 0.\end{aligned}$$

Recall that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ for any $A, B \in \mathcal{J}_2(H)$, and that $P = P^2$. We have

$$\begin{aligned}&\mathrm{Tr}\left((I-\eta P)B(I-\eta P)B^T - (1-\eta)^2BB^T\right) \\ =&\mathrm{Tr}\left((2\eta-\eta^2)BB^T - \eta PBB^TP - \eta BPB^T + \eta^2 PBPB^TP\right) \\ =&\eta\mathrm{Tr}\left((I-P)B(I-P)B^T\right) + \eta(1-\eta)\mathrm{Tr}\left(BB^T - PBPB^TP\right) \\ =&\eta\|(I-P)B(I-P)\|_{\mathsf{HS}}^2 + \eta(1-\eta)\mathrm{Tr}(BB^T - PBPB^TP) \ge 0.\end{aligned}$$

We take expectations of the above inequality to have $Q_\eta \succeq (1-\eta)^2 I$.

Similarly when $1 < \eta < 2$,

$$\begin{aligned}&\mathrm{Tr}\left((I-\eta P)B(I-\eta P)B^T + (\eta-1)BB^T\right) \\ =&\eta\mathrm{Tr}\left(BB^T - PBB^TP - BPB^T + PBPB^TP\right) + \eta(\eta-1)\mathrm{Tr}(PBPB^TP) \\ =&\eta\left\|(I-P)B(I-P)\right\|_{\mathsf{HS}}^2 + \eta(\eta-1)\left\|PBP\right\|_{\mathsf{HS}}^2 \ge 0.\end{aligned}$$

So $Q_\eta \succeq (1-\eta)I$. □

We now provide some characteristics of the structure of Q_η. Define operators R_L, S_L, and R_*, all from $\mathcal{J}_2(H)$ to $\mathcal{J}_2(H)$, by $R_L(B) = L^{1/2}BL^{1/2}$, $S_L(B) = (LB + BL)/2$, and $R_*(B) = \mathbb{E}[PBP]$, respectively. Recall that $L = \mathbb{E}P$. Since $\|L\|_{H\to H}$ is finite, R_L and S_L are all bounded operators. Moreover, $\|P\|_{H\to H} = 1$ implies that R_* is bounded. One can use the arguments in (29) to show that R_L, S_L, and R_* are all self-adjoint. For any $B \in \mathcal{J}_2(H)$, $\langle R_L(B), B\rangle_{\mathsf{HS}} = \left\|L^{1/4}BL^{1/4}\right\|_{\mathsf{HS}}^2 \geq 0$, $\langle S_L(B), B\rangle_{\mathsf{HS}} = (\left\|L^{1/2}B\right\|_{\mathsf{HS}}^2 + \left\|BL^{1/2}\right\|_{\mathsf{HS}}^2)/2 \geq 0$, and $\langle R_*(B), B\rangle_{\mathsf{HS}} = \mathbb{E}\left[\|PBP\|_{\mathsf{HS}}^2\right] \geq 0$. So R_L, S_L, and R_* are all positive semi-definite. It is straightforward to see that

$$Q_\eta = I - 2\eta S_L + \eta^2 R_*.$$

Recall that $\{\phi_i \otimes \phi_j\}_{i,j\in\mathcal{I}}$ is an orthonormal basis of $\mathcal{J}_2(H)$. In particular, the eigensystem of R_L is $\left\{(\sqrt{\lambda_i\lambda_j}, \phi_i \otimes \phi_j)\right\}_{i,j\in\mathcal{I}}$, and the eigensystem of S_L is $\{((\lambda_i + \lambda_j)/2, \phi_i \otimes \phi_j)\}_{i,j\in\mathcal{I}}$. Since R_L and S_L share the same set of eigenvectors, they commute. Thanks to the inequality $\sqrt{xy} \leq (x + y)/2$ for $x, y \geq 0$, we have $R_L \preceq S_L$. For $s > 0$ and $B \in \mathcal{J}_2(H)$, the spectral decomposition implies $R_L^s(B) = L^{s/2}BL^{s/2}$.

We claim that $R_* \preceq S_L$. In fact, for any $B \in \mathcal{J}_2(H)$, one has that

$$\langle B, R_*(B)\rangle_{\mathsf{HS}} = \mathbb{E}\mathrm{Tr}(B^TPBP) = \mathbb{E}\mathrm{Tr}(PB^TPPBP).$$

A simple calculation gives that,

$$\mathrm{Tr}(PB^TPPBP) \leq \begin{cases} \mathrm{Tr}(B^TPPB), \\ \mathrm{Tr}(PB^TBP), \end{cases}$$

which implies $R_* \preceq S_L$.

We are now ready to prove Theorem 2.2.

Proof of Theorem 2.2. Since $\eta \in (0, 2)$, by the relations $R_* \preceq S_L$ and $R_L \preceq S_L$, one has

$$(2\eta - \eta^2)R_L \preceq (2\eta - \eta^2)S_L + \eta^2(S_L - R_*) = I - Q_\eta. \tag{30}$$

By the Löwner-Heinz inequality [32, 37], for any $s \in (0, \frac{1}{2}]$, one has $(\eta(2 - \eta)R_L)^{2s} \preceq (I - Q_\eta)^{2s}$, which implies (we group the details into Lemma B.1 in Appendix B for completeness) that for any $B \in \mathcal{J}_2(H)$, a preimage vector $(I - Q_\eta)^{-s}[R_L^s(B)]$ exists and satisfies

$$\|(I - Q_\eta)^{-s}R_L^s(B)\|_{\mathsf{HS}} \leq (\eta(2 - \eta))^{-s}\|B\|_{\mathsf{HS}}. \tag{31}$$

Now we apply the spectral mapping theorem (see, e.g., Yosida [50, Section XI.5]) to estimate $\left\|Q_\eta^k(I - Q_\eta)^s\right\|_{\mathcal{J}_2(H)\to\mathcal{J}_2(H)}$, for any integer $k \geq 1$ and any $0 < s \leq 1$. When $0 < \eta \leq 1$, recall $\sigma(Q_\eta) \subset [0, 1]$. Since the function $\tau^k(1 - \tau)^s$ of $\tau \in [0, 1]$ achieves its maximum $\frac{k^ks^s}{(k+s)^{k+s}} \leq k^{-s}s^s \leq k^{-s}$ at $\tau = \frac{k}{k+s}$, we have

$$\left\|Q_\eta^k(I - Q_\eta)^s\right\|_{\mathcal{J}_2(H)\to\mathcal{J}_2(H)} \leq k^{-s}, \quad \text{for any } 0 < \eta \leq 1. \tag{32}$$

When $1 < \eta < 2$, $\sigma(Q_\eta)$ extends below zero, and is bounded from below by $1-\eta$. On $[1-\eta, 0]$, the function $|\tau^k(1 - \tau)^s|$ is decreasing and achieves its maximum $(\eta - 1)^k\eta^s$ at the left end. Extending $k^s(\eta - 1)^k\eta^s$ as a function of k on $(0, \infty)$, one sees that this function tends to zero as $k \downarrow 0$ and as $k \to \infty$. Also, $k^s(\eta - 1)^k\eta^s$ achieves its maximum $e^{-s}(-\eta s/\log(\eta - 1))^s$ at $k = -s/\log(\eta - 1)$. Recall that for $\tau \in [0, 1]$, $0 \leq \tau^k(1 - \tau)^s \leq k^{-s}$. Therefore we bound the norm of $Q_\eta^k(I - Q_\eta)^s$ by $k^{-s}\max\{1, e^{-s}(-\eta s/\log(\eta - 1))^s\}$. This bound, together with (32), yields that

$$\left\|Q_\eta^k(I - Q_\eta)^s\right\|_{\mathcal{J}_2(H)\to\mathcal{J}_2(H)} \leq C_{\eta,s}k^{-s}, \quad \text{for any } 0 < \eta < 2, \tag{33}$$

where

$$C_{\eta,s} := \begin{cases} 1, & \text{if } 0 < \eta \leq 1, \\ \max\{1, e^{-s}(-\eta s/\log(\eta-1))^s\}, & \text{if } 1 < \eta < 2. \end{cases} \tag{34}$$

The assumption $x^0 \in L^{s_1}(H)$ implies $x^0 \otimes x^0 = R_L^{2s_1}((L^{-s_1}x^0) \otimes (L^{-s_1}x^0))$. Similarly, $x' \otimes x' = R_L^{2s_2}((L^{-s_2}x') \otimes (L^{-s_2}x'))$. Consider the error decomposition

$$\begin{aligned} \mathbb{E}\left[\langle x', r_{k+1}\rangle^2\right] =& \mathbb{E}\left\langle x', (I-\eta P_k)\cdots(I-\eta P_1)(x^0 \otimes x^0)(I-\eta P_1)\cdots(I-\eta P_k)x'\right\rangle \\ =& \mathrm{Tr}[(x' \otimes x')Q_\eta^k(x^0 \otimes x^0)] = \left\langle \Upsilon_1, Q_\eta^k(I-Q_\eta)^{2s_1+2s_2}(\Upsilon_2)\right\rangle_{\mathsf{HS}}, \end{aligned} \tag{35}$$

where

$$\Upsilon_1 = (I-Q_\eta)^{-2s_2}(x' \otimes x') = (I-Q_\eta)^{-2s_2}R_L^{2s_2}((L^{-s_2}x') \otimes (L^{-s_2}x')),$$

and

$$\Upsilon_2 = (I-Q_\eta)^{-2s_1}(x^0 \otimes x^0).$$

One uses (31) to obtain $\|\Upsilon_1\|_{\mathsf{HS}} \leq (\eta(2-\eta))^{-2s_2}\|(L^{-s_2}x') \otimes (L^{-s_2}x')\|_{\mathsf{HS}} = (\eta(2-\eta))^{-2s_2}\|L^{-s_2}x'\|^2$ and $\|\Upsilon_2\|_{\mathsf{HS}} \leq (\eta(2-\eta))^{-2s_1}\|L^{-s_1}x^0\|^2$. We combine (35) and (33) to obtain

$$\mathbb{E}\left[\langle x', r_{k+1}\rangle^2\right] \leq \|L^{-s_2}x'\|^2\|L^{-s_1}x^0\|^2 C_{\eta,2s_1+2s_2}(\eta(2-\eta)k)^{-2s_1-2s_2},$$

which is (7).

If the random vector $\xi' \sim (\mathbb{S}, \rho)$ is used, since ξ' is independent of the whole sample,

$$\mathbb{E}\left[\langle \xi', r_{k+1}\rangle^2\right] = \mathbb{E}\mathrm{Tr}[(\xi' \otimes \xi')Q_\eta^k(x^0 \otimes x^0)] = \mathrm{Tr}[LQ_\eta^k(x^0 \otimes x^0)].$$

Note that $L = R_L^{1/2}(L^{1/2})$, and $\|L^{1/2}\|_{\mathsf{HS}}^2 = \mathrm{Tr}(L)$. The above error is further estimated by

$$\mathbb{E}\left[\langle \xi', r_{k+1}\rangle^2\right] = \left\langle \Upsilon_1', Q_\eta^k(I-Q_\eta)^{2s_1+\frac{1}{2}}(\Upsilon_2)\right\rangle_{\mathsf{HS}},$$

where

$$\Upsilon_1' = (I-Q_\eta)^{-1/2}R_L^{1/2}(L^{1/2}).$$

So

$$\begin{aligned} \mathbb{E}\left[\langle \xi', r_{k+1}\rangle^2\right] &\leq \|L^{1/2}\|_{\mathsf{HS}}\|L^{-s_1}x^0\|^2 C_{\eta,2s_1+1/2}(\eta(2-\eta)k)^{-2s_1-\frac{1}{2}} \\ &= \sqrt{\mathrm{Tr}(L)}\|L^{-s_1}x^0\|^2 C_{\eta,2s_1+1/2}(\eta(2-\eta)k)^{-2s_1-\frac{1}{2}}, \end{aligned}$$

which proves (8). □

4 Error Analysis in the General Setting with Noise

In this section we study Algorithm (2) in the general setting with observational noise. That is, we assume $\sigma^2 = \mathrm{Var}(\epsilon_i) \geq 0$ for Model $\mathbb{M}(\mathbb{S}, \rho, x^0, m, \sigma^2)$. The key technique of our analysis is inspired by (36) which does a one-step reduction of the expected error $\mathbb{E}\left[\|r_{k+1}\|^2\right]$. We lift the trace operator in (36) to obtain (37), which is further used iteratively to derive (39). The operator W_η is introduced in

(38) to replace Q_η for a better algebraic property (as provided in Lemma 4.1). Theorem 4.6 parallels Theorem 4.4 for studying weak convergence.

Recall that we use $r_{k+1} = x_{k+1} - x^0$ to denote the error vector after the k-th iteration of Algorithm (2). For $k \geq 1$, because of noise, the iteration step of Algorithm (2) is now written as

$$r_{k+1} = x_{k+1} - x^0 = x_k - x^0 + \eta_k \left(\left\langle \xi_k, x^0 \right\rangle + \epsilon_k - \langle \xi_k, x_k \rangle \right) \xi_k = (I - \eta_k P_k) r_k + \eta_k \epsilon_k \xi_k.$$

Since r_k, ξ_k, and ϵ_k are independent, and $\mathbb{E}\epsilon_k = 0$, recall the operator Q_η defined in (27), we have

$$\begin{aligned}\mathbb{E}\|r_{k+1}\|^2 =& \mathbb{E}\mathrm{Tr}(r_{k+1} \otimes r_{k+1}) = \mathrm{Tr}\mathbb{E}(r_{k+1} \otimes r_{k+1}) \\ =& \mathrm{Tr}Q_{\eta_k}(\mathbb{E}(r_k \otimes r_k)) + 2\eta_k(\mathbb{E}\epsilon_k)\mathbb{E}\mathrm{Tr}[((I - \eta_k P_k) r_k) \otimes \xi_k] \\ &+ \eta_k^2 (\mathbb{E}\epsilon_k^2) \mathrm{Tr}\mathbb{E}(\xi_k \otimes \xi_k) \\ =& \mathrm{Tr}Q_{\eta_k}(\mathbb{E}(r_k \otimes r_k)) + \eta_k^2 \sigma^2 \mathrm{Tr}(L). \end{aligned} \tag{36}$$

For $k \geq 1$, we write $\mathcal{R}_k = \mathbb{E}(r_k \otimes r_k)$. Then $\mathbb{E}\|r_{k+1}\|^2 = \mathrm{Tr}(\mathcal{R}_{k+1})$, and similar as Equation (36), one has

$$\mathcal{R}_{k+1} = Q_{\eta_k}(\mathcal{R}_k) + \sigma^2 \eta_k^2 L. \tag{37}$$

In particular, $\mathcal{R}_1 = (-x^0) \otimes (-x^0) = x^0 \otimes x^0$. Recall that $Q_\eta = I - 2\eta S_L + \eta^2 R_*$, where in general S_L does not commute with R_*, so Q_η may not commute with $Q_{\eta'}$ when $\eta \neq \eta'$. To overcome the difficulty, we use the operator W_η defined by

$$W_\eta(B) = Q_\eta(B) + \eta^2 (R_L^2 - R_*)(B) = (I - \eta L) B (I - \eta L) \tag{38}$$

for any $B \in \mathcal{J}_2(H)$ and $\eta \in [0, 1]$. Then $W_\eta = I - 2\eta S_L + \eta^2 R_L^2$, and we summarize some properties of W_η which will be used later.

Lemma 4.1. *For $\eta \in [0,1]$, the operator W_η on $\mathcal{J}_2(H)$ has the following properties.*

(1) The eigensystem of W_η is $\{(1 - \eta\lambda_i)(1 - \eta\lambda_j), \phi_i \otimes \phi_j\}_{i,j \in \mathcal{I}}$.

(2) W_η is positive semi-definite.

(3) For any positive semi-definite operator $B \in \mathcal{J}_2(H)$, $W_\eta(B)$ is positive semi-definite.

(4) For any positive semi-definite operator $B \in \mathcal{J}_1(H)$, $\mathrm{Tr}(W_\eta(B)) \leq \mathrm{Tr}(B)$.

Proof. The eigensystem is derived from the expansion $W_\eta = I - 2\eta S_L + \eta^2 R_L^2$. Since $\lambda_i \geq 0$ and $\sum_i \lambda_i = \mathrm{Tr}(L) = \mathbb{E}\mathrm{Tr}(P) = 1$, $0 \leq (1 - \eta\lambda_i)(1 - \eta\lambda_j) \leq 1$ for any i, j. Therefore W_η is positive semi-definite. For $B \in \mathcal{J}_2(H)$, $W_\eta(B) = (I - \eta L) B (I - \eta L)$, so whenever B is positive semi-definite, so is $W_\eta(B)$. If $B \in \mathcal{J}_1(H)$ is positive semi-definite, we have

$$\mathrm{Tr}(W_\eta(B)) = \sum_{i \in \mathcal{I}} \langle \phi_i, (I - \eta L) B (I - \eta L) \phi_i \rangle = \sum_{i \in \mathcal{I}} (1 - \eta\lambda_i)^2 \langle \phi_i, B\phi_i \rangle.$$

Since $\langle \phi_i, B\phi_i \rangle \geq 0$ for any i, $\mathrm{Tr}(W_\eta(B)) \leq \mathrm{Tr}(B)$. The proof is thus complete. □

Consider the following error decomposition,

$$\begin{aligned}\mathcal{R}_{k+1} =& W_{\eta_k}(\mathcal{R}_k) + \eta_k^2((R_* - R_L^2)(\mathcal{R}_k) + \sigma^2 L) \\ =& W_{\eta_k} W_{\eta_{k-1}}(\mathcal{R}_{k-1}) + \eta_{k-1}^2 W_{\eta_k}((R_* - R_L^2)(\mathcal{R}_{k-1}) + \sigma^2 L) \\ &+ \eta_k^2((R_* - R_L^2)(\mathcal{R}_k) + \sigma^2 L) \\ =& W_{\eta_k} W_{\eta_{k-1}} W_{\eta_{k-2}}(\mathcal{R}_{k-2}) + \eta_{k-2}^2 W_{\eta_k} W_{\eta_{k-1}}((R_* - R_L^2)(\mathcal{R}_{k-2}) + \sigma^2 L) \\ &+ \eta_{k-1}^2 W_{\eta_k}((R_* - R_L^2)(\mathcal{R}_{k-1}) + \sigma^2 L) + \eta_k^2((R_* - R_L^2)(\mathcal{R}_k) + \sigma^2 L). \end{aligned}$$

Note that $W_\eta = I - 2\eta S_L + \eta^2 R_L^2$, of which the eigenvectors are $\{(\phi_i \otimes \phi_j)\}_{i,j\in\mathcal{I}}$, as we discussed above. So the operators $W_{\eta_1}, \cdots, W_{\eta_k}$ commute. One carries on the above iterations to give

$$\mathcal{R}_{k+1} = \left(\prod_{l=1}^{k} W_{\eta_l}\right)(\mathcal{R}_1) + \sum_{j=1}^{k} \eta_j^2 \left(\prod_{l=j+1}^{k} W_{\eta_l}\right)\left[(R_* - R_L^2)(\mathcal{R}_j) + \sigma^2 L\right]. \tag{39}$$

Here we abuse the notation a little bit by defining $\prod_{l=k+1}^{k} W_{\eta_l}$ (of which the starting index $k+1$ is greater than the ending index k) as identity.

When $\sigma^2 = 0$, Equation (37) becomes $\mathcal{R}_{k+1} = Q_{\eta_k}(\mathcal{R}_k)$, which implies $\mathbb{E}\left[\|r_{k+1}\|^2\right] = \mathbb{E}\left[\|(I - \eta_k P_k) r_k\|^2\right]$. Since P_k is a random orthogonal projection, one always has $\mathbb{E}\left[\|r_{k+1}\|^2\right] \leq \mathbb{E}\left[\|r_k\|^2\right]$, and therefore the expected error is at least uniformly bounded. When $\sigma^2 > 0$, the following lemma shows that $\mathbb{E}\left[\|r_k\|^2\right]$ is still uniformly bounded under the assumption that the sum of the squared step-sizes is finite. This lemma therefore suggests the important role the step-sizes playing in controlling the error introduced by observational noise.

Lemma 4.2. *Let m be a positive integer, or infinity. Assume $\sigma^2 \geq 0$ for Model $\mathbb{M}(\mathbb{S}, \rho, x^0, m, \sigma^2)$. If $\eta_{2,m} := \sum_{j=1}^{m} \eta_j^2 < 1$, then for $1 \leq k \leq m$ (or for $1 \leq k < \infty$ when $m = \infty$), one has*

$$\mathrm{Tr}(\mathcal{R}_k) \leq \frac{\|x^0\|^2 + \sigma^2 \mathrm{Tr}(L)\eta_{2,m}}{1 - \eta_{2,m}}. \tag{40}$$

Proof. Recall that $\mathcal{R}_1 = x^0 \otimes x^0$, so $\mathrm{Tr}(\mathcal{R}_1) = \|x^0\|^2$ and the estimate (40) holds true. Now suppose for some $1 \leq k < m$, the estimate (40) holds true for all $l = 1, \ldots, k$. To finish the proof with induction, we need only to verify the inequality (40) for $l = k+1$.

We study the trace of the right-hand side of (39). Let $B \in \mathcal{J}_1(H)$ be positive semi-definite. Lemma 4.1 gives

$$0 \leq \mathrm{Tr}(W_\eta(B)) \leq \mathrm{Tr}(B). \tag{41}$$

One repeats (41) to give

$$\mathrm{Tr}\left[\left(\prod_{l=1}^{k} W_{\eta_l}\right)(\mathcal{R}_1)\right] \leq \mathrm{Tr}(\mathcal{R}_1) = \|x^0\|^2,$$

and

$$\mathrm{Tr}\left[\left(\prod_{i=j+1}^{k} W_{\eta_i}\right)((R_* - R_L^2)(\mathcal{R}_j) + \sigma^2 L)\right] \leq \mathrm{Tr}(R_*(\mathcal{R}_j)) + \sigma^2 \mathrm{Tr}(L),$$

for $j = 1, \ldots, k$. Note that $\mathrm{Tr}(R_*(\mathcal{R}_j)) = \mathbb{E}\mathrm{Tr}(P\mathcal{R}_j P) \leq \mathbb{E}\mathrm{Tr}(\mathcal{R}_j) = \mathrm{Tr}(\mathcal{R}_j)$. By the inductive hypothesis,

$$\begin{aligned}\mathrm{Tr}(\mathcal{R}_{k+1}) &\leq \|x^0\|^2 + \eta_{2,m}\left(\frac{\|x^0\|^2 + \sigma^2\mathrm{Tr}(L)\eta_{2,m}}{1 - \eta_{2,m}} + \sigma^2 \mathrm{Tr}(L)\right) \\ &= \frac{\|x^0\|^2 + \sigma^2 \mathrm{Tr}(L)\eta_{2,m}}{1 - \eta_{2,m}}.\end{aligned}$$

This completes the proof. □

We will need the following technical lemma, of which the proof is postponed to Appendix A.

Lemma 4.3. *Let $s \geq 0$, $\tau \in [0,1]$, and $\eta_1, \dots, \eta_k \in [0,1]$. One has*

$$\left[\prod_{l=1}^{k}(1-\eta_l\tau)\right]\tau^s \leq \frac{2s^s+2}{1+\left(\sum_{l=1}^{k}\eta_l\right)^s}, \tag{42}$$

where we abuse the notation a little bit to define $0^0 := 1$ when $s = 0$.

Theorem 4.4 below estimates the expected squared norm of the error vector r_{k+1} for Algorithm (2), without specific design of the step-sizes. The rates of strong convergence in Theorem 2.6 with RK-FH setting is a corollary of Theorem 4.4.

Theorem 4.4. *Let $\{(\xi_i, y_i)\}_{i=1}^m$ be drawn from Model $\mathbb{M}(\mathbb{S}, \rho, x^0, m, \sigma^2)$ with $\sigma^2 \geq 0$, $1 \leq m \leq \infty$, and $x^0 \in L^{s_1}(H)$ for some $s_1 > 0$. Assume $\{\eta_j\}_{j=1}^m \subset (0,1)$ and $\eta_{2,m} := \sum_{j=1}^m \eta_j^2 < 1$ for Algorithm (2). Then for $1 \leq k \leq m$ (or $1 \leq k < \infty$ if $m = \infty$), one has*

$$\mathbb{E}\left[\|r_{k+1}\|^2\right] = \mathrm{Tr}(\mathcal{R}_{k+1}) \leq \frac{4(s_1^{s_1}+1)^2\|L^{-s_1}x^0\|^2}{1+\left(\sum_{j=1}^{k}\eta_j\right)^{2s_1}} + \frac{\|x^0\|^2+\sigma^2\mathrm{Tr}(L)}{1-\eta_{2,m}}\sum_{j=1}^{k}\eta_j^2. \tag{43}$$

Proof. We bound the trace of $\mathcal{R}_{k+1}$ according to the expansion (39). Since $x^0 \in L^{s_1}(H)$, we write $x^0 = \sum_{i\in\mathcal{I}} b_i\lambda_i^{s_1}\phi_i$, so $\sum_i b_i^2 = \|L^{-s_1}x^0\|^2$. By the definition of W_η, we have

$$\mathrm{Tr}\left[\left(\prod_{l=1}^{k}W_{\eta_l}\right)(\mathcal{R}_1)\right] = \left\|\left[\prod_{l=1}^{k}(1-\eta_l L)\right]x^0\right\|^2 = \sum_{i\in\mathcal{I}} b_i^2\lambda_i^{2s_1}\left[\prod_{l=1}^{k}(1-\eta_l\lambda_i)\right]^2.$$

Since $\lambda_i \in [0,1]$ for any i, we have by Lemma 4.3,

$$\begin{aligned}\mathrm{Tr}\left[\left(\prod_{l=1}^{k}W_{\eta_l}\right)(\mathcal{R}_1)\right] &\leq \|L^{-s_1}x^0\|^2\left[\max_{\tau\in[0,1]}\tau^{s_1}\prod_{l=1}^{k}(1-\eta_l\tau)\right]^2 \\ &\leq \frac{4(s_1^{s_1}+1)^2\|L^{-s_1}x^0\|^2}{1+\left(\sum_{l=1}^{k}\eta_l\right)^{2s_1}}.\end{aligned} \tag{44}$$

Now we estimate the trace of the second term on the right-hand side of (39). Recall Lemma 4.1. Since $R_L^2(\mathcal{R}_j) = L\mathcal{R}_jL$ is positive semi-definite, and $\mathrm{Tr}(R_*(\mathcal{R}_j)) = \mathbb{E}\mathrm{Tr}(P\mathcal{R}_jP) \leq \mathrm{Tr}(\mathcal{R}_j)$, one uses the bound (40) to obtain

$$\begin{aligned}&\mathrm{Tr}\left[\sum_{j=1}^{k}\eta_j^2\left(\prod_{l=j+1}^{k}W_{\eta_l}\right)((R_*-R_L^2)(\mathcal{R}_j)+\sigma^2L)\right] \\ &\leq \sum_{j=1}^{k}\eta_j^2\mathrm{Tr}\left[\left(\prod_{l=j+1}^{k}W_{\eta_l}\right)(R_*(\mathcal{R}_j)+\sigma^2L)\right] \leq \sum_{j=1}^{k}\eta_j^2(\mathrm{Tr}(\mathcal{R}_j)+\sigma^2\mathrm{Tr}(L)) \\ &\leq \left(\frac{\|x^0\|^2+\sigma^2\mathrm{Tr}(L)\eta_{2,m}}{1-\eta_{2,m}}+\sigma^2\mathrm{Tr}(L)\right)\sum_{j=1}^{k}\eta_j^2 = \frac{\|x^0\|^2+\sigma^2\mathrm{Tr}(L)}{1-\eta_{2,m}}\sum_{j=1}^{k}\eta_j^2.\end{aligned} \tag{45}$$

One completes the proof by combining (45) and (44). □

For proving Theorem 2.6, we need the following technical lemma, of which the proof is put to Appendix A for completeness.

Lemma 4.5. *For any $b > 0$ and $\theta \geq 0$,*

$$\int_0^b \frac{dx}{1 + x^\theta} \leq \begin{cases} \frac{1}{1-\theta} b^{1-\theta}, & \text{if } 0 \leq \theta < 1, \\ \log(b+1), & \text{if } \theta = 1, \\ \frac{\theta}{\theta - 1}, & \text{if } \theta > 1. \end{cases}$$

Proof of Theorem 2.6. When the step-sizes in Algorithm (2) are fixed to be a constant η, the expansion (39) takes the form (for $1 \leq k \leq m$)

$$\mathcal{R}_{k+1} = W_\eta^k(\mathcal{R}_1) + \eta^2 \sum_{j=1}^{k} W_\eta^{k-j} \left(R_*(\mathcal{R}_j) - R_L^2(\mathcal{R}_j) + \sigma^2 L \right).$$

First, we apply the estimate (44) to get

$$\mathrm{Tr}\left(W_\eta^m(\mathcal{R}_1) \right) \leq \frac{4(s_1^{s_1} + 1)^2 \|L^{-s_1} x^0\|^2}{1 + (m\eta)^{2s_1}}.$$

We note that $m\eta \geq 1$, which is because when $s_1 + s_* \geq 1$, $m\eta = m^{1 - \frac{2s_1 + s_*}{1 + 2s_1 + s_*}} = m^{\frac{1}{1 + 2s_1 + s_*}} \geq 1$, and when $s_1 + s_* < 1$, $m\eta = m^{1 - \frac{1 + s_1}{2 + s_1}} = m^{\frac{1}{2 + s_1}} \geq 1$.

Next, we rewrite the estimate (43) in Theorem 4.4, in the constant step-size-setting, as

$$\mathrm{Tr}(\mathcal{R}_{k+1}) \leq D_1 \left(\frac{1}{1 + (k\eta)^{2s_1}} + k\eta^2 \right), \quad 1 \leq k \leq m, \tag{46}$$

where $D_1 = \max\left\{ 4(s_1^{s_1} + 1)^2 \|L^{-s_1} x^0\|^2, \left(\|x^0\|^2 + \sigma^2 \mathrm{Tr}(L) \right) / \left(1 - 2^{\frac{2}{2 + s_1 + (s_1 + s_* - 1)_+} - 1} \right) \right\}$, with $(t)_+ := \max\{t, 0\}$ for any $t \in \mathbb{R}$, and we have used the following estimates for $\eta_{2,m}$ thanks to $m \geq 2$. When $s_1 + s_* \geq 1$,

$$\eta_{2,m} = \eta^2 m = m^{\frac{1 + s_* + 2s_1 - 2s_* - 4s_1}{1 + 2s_1 + s_*}} \leq 2^{-\frac{s_* + 2s_1 - 1}{1 + 2s_1 + s_*}} = 2^{\frac{2}{2 + s_1 + (s_1 + s_* - 1)_+} - 1} < 1,$$

and when $s_1 + s_* < 1$,

$$\eta_{2,m} = \eta^2 m = m^{\frac{2 + s_1 - 2 - 2s_1}{2 + s_1}} \leq 2^{\frac{-s_1}{2 + s_1}} = 2^{\frac{2}{2 + s_1 + (s_1 + s_* - 1)_+} - 1} < 1.$$

In particular, when $k = 0$, the estimate (46) still holds true since $\mathrm{Tr}(\mathcal{R}_1) = \|x^0\|^2 \leq \frac{\|x^0\|^2 + \sigma^2 \mathrm{Tr}(L)}{1 - \eta_{2,m}} \leq D_1$. We apply Lemma 4.5 and the fact that $\eta m \geq 1$ to obtain

$$\begin{aligned} &\mathrm{Tr}\left(\eta^2 \sum_{j=1}^{m} W_\eta^{m-j} R_*(\mathcal{R}_j) \right) \leq \eta^2 \sum_{j=1}^{m} \mathrm{Tr}(\mathcal{R}_j) \leq D_1 \eta^2 \sum_{j=0}^{m-1} \left(\frac{1}{1 + (j\eta)^{2s_1}} + j\eta^2 \right) \\ \leq & D_1 \eta^2 + D_1 \eta^2 \int_0^{m-1} \frac{dx}{1 + (\eta x)^{2s_1}} + D_1 \eta^4 \frac{m^2}{2} \leq \frac{3}{2} D_1 \eta^4 m^2 + D_1 \eta \int_0^{\eta(m-1)} \frac{dx}{1 + x^{2s_1}} \\ \leq & \frac{3}{2} D_1 \eta^4 m^2 + D_1 \begin{cases} \frac{1}{1 - 2s_1} \eta (\eta m)^{1 - 2s_1}, & \text{if } 0 < s_1 < 1/2, \\ \eta \log(\eta m - \eta + 1), & \text{if } s_1 = 1/2, \\ 2s_1 \eta / (2s_1 - 1), & \text{if } s_1 > 1/2. \end{cases} \end{aligned}$$

Then, similar to the argument used in (45), one has

$$\mathrm{Tr}\left(\eta^2 \sum_{j=1}^{m} W_\eta^{m-j} (-R_L^2)(\mathcal{R}_j) \right) \leq 0.$$

Finally, recall that $\mathrm{Tr}(L^{s_*}) = \sum_{i\in\mathcal{I}} \lambda_i^{s_*}$. By Lemma 4.3 (recalled that we use $0^0 := 1$),

$$\begin{aligned}
&\mathrm{Tr}\left(\eta^2 \sum_{j=1}^{m} W_\eta^{m-j}(\sigma^2 L)\right) = \sigma^2\eta^2 \sum_{j=1}^{m}\sum_{i\in\mathcal{I}} (1-\eta\lambda_i)^{2(m-j)}\lambda_i \\
\le&\sigma^2\eta^2 \sum_{j=1}^{m}\sum_{i\in\mathcal{I}} \left((1-\eta\lambda_i)^{2(m-j)}\lambda_i^{1-s_*}\right)\lambda_i^{s_*} \\
\le&\sigma^2\eta^2 \sum_{j=1}^{m}\left(\max_{\tau\in[0,1]}(1-\eta\tau)^{2(m-j)}\tau^{1-s_*}\right)\mathrm{Tr}(L^{s_*}) \\
\le&2((1-s_*)^{1-s_*}+1)\sigma^2\mathrm{Tr}(L^{s_*})\eta^2\sum_{j=1}^{m}\frac{1}{1+(2(m-j)\eta)^{1-s_*}}.
\end{aligned}$$

Since $0 < s_* \le 1$, Lemma 4.5 implies

$$\begin{aligned}
\eta\sum_{j=1}^{m}\frac{1}{1+(2\eta(m-j))^{1-s_*}} &= \eta\sum_{j=0}^{m-1}\frac{1}{1+(2\eta j)^{1-s_*}} \le \eta + \eta\int_0^{m-1}\frac{dx}{1+(2\eta x)^{1-s_*}} \\
&\le \eta + \frac{1}{2}\int_0^{2\eta m}\frac{dx}{1+x^{1-s_*}} \le \eta + \frac{(2\eta m)^{s_*}}{2s_*}.
\end{aligned}$$

We summarize the above analysis to get

$$\begin{aligned}
\mathrm{Tr}(\mathcal{R}_{m+1}) \le&\frac{4(s_1^{s_1}+1)^2\|L^{-s_1}x^0\|^2}{1+(m\eta)^{2s_1}} + \frac{3}{2}D_1\eta^4m^2 \\
&+ 2((1-s_*)^{1-s_*}+1)\sigma^2\mathrm{Tr}(L^{s_*})(\eta^2 + s_*^{-1}\eta^{1+s_*}m^{s_*}) \\
&+ D_1\begin{cases} \frac{1}{1-2s_1}\eta^{2-2s_1}m^{1-2s_1}, & \text{if } 0<s_1<1/2, \\ \left(\log 2 + s_*^{-1}\right)\eta^{1+s_*}m^{s_*}, & \text{if } s_1 = 1/2, \\ 2s_1\eta/(2s_1-1) & \text{if } s_1 > 1/2, \end{cases}
\end{aligned} \tag{47}$$

where for the case $s_1 = 1/2$, (recall that $\eta m \ge 1$) we integrate the inequality $t^{-1} \le t^{s_*-1}$ for $t \ge 1$ to obtain $\log(\eta m) \le (\eta m)^{s_*}/s_*$, and we have applied the inequality

$$\eta\log(\eta m - \eta + 1) \le \eta\log(2\eta m) \le \eta\log 2 + \eta(\eta m)^{s_*}/s_* \le \left(\log 2 + s_*^{-1}\right)\eta^{1+s_*}m^{s_*}.$$

Since $\eta m \ge 1$ and $\eta \le 1$, the estimate (47) is further tidied as

$$\begin{aligned}
\mathrm{Tr}(\mathcal{R}_{m+1}) \le&D_2(s_1,s_*)((m\eta)^{-2s_1} + \eta^4m^2 + \eta^{1+s_*}m^{s_*}) + \\
&D_2(s_1,s_*)\begin{cases} \eta^{2-2s_1}m^{1-2s_1}, & \text{if } 0<s_1<1/2, \\ 0, & \text{if } s_1 \ge 1/2, \end{cases}
\end{aligned}$$

where (note that $s_*^{-1} \ge 1$, $(1-s_*)^{1-s_*}+1 \le 2$, $\eta m \ge 1$, and $\eta^2 \le \eta \le \eta^{1+s_*}m^{s_*}$),

$$\begin{aligned}
D_2(s_1,s_*) =&4(s_1^{s_1}+1)^2\|L^{-s_1}x^0\|^2 + \frac{3}{2}D_1 + 8s_*^{-1}\sigma^2\mathrm{Tr}(L^{s_*}) + \\
&D_1\begin{cases} \frac{1}{1-2s_1}, & \text{if } 0<s_1<1/2, \\ s_*^{-1} + \log 2, & \text{if } s_1 = 1/2, \\ 2s_1/(2s_1-1), & \text{if } s_1 > 1/2. \end{cases} .
\end{aligned}$$

We derive the error bounds in two cases.

Case 1, $s_1 + s_* \geq 1$. In this case $\eta = m^{-(2s_1+s_*)/(1+2s_1+s_*)}$ implies $(m\eta)^{-2s_1} = \eta^{1+s_*} m^{s_*} = m^{-2s_1/(1+2s_1+s_*)}$, and

$$\eta^4 m^2 = m^{2-\frac{8s_1+4s_*}{1+2s_1+s_*}} = m^{-\frac{2(s_1+s_*-1)+2s_1}{1+2s_1+s_*}} \leq m^{-\frac{2s_1}{1+2s_1+s_*}}.$$

Moreover, when $0 < s_1 < 1/2$, $1 \leq s_1 + s_* \leq 2s_1 + s_*$ and $\eta m \geq 1$ implies $\eta^{2-2s_1} m^{1-2s_1} = \eta(\eta m)^{1-2s_1} \leq \eta^{1+s_*} m^{s_*}$. Therefore we have

$$\mathrm{Tr}(\mathcal{R}_{m+1}) \leq 4D_2(s_1, s_*) m^{-\frac{2s_1}{1+2s_1+s_*}}.$$

Case 2, $s_1 + s_* < 1$. In this case $\eta = m^{-(1+s_1)/(2+s_1)}$ implies $(m\eta)^{-2s_1} = \eta^4 m^2 = m^{-2s_1/(2+s_1)}$, and

$$\eta^{1+s_*} m^{s_*} = m^{-\frac{1+s_*+s_1+s_* s_1 - 2s_* - s_1 s_*}{2+s_1}} \leq m^{-\frac{2s_1}{2+s_1}}.$$

In particular, when $0 < s_1 < 1/2$,

$$\eta^{2-2s_1} m^{1-2s_1} = m^{-\frac{1+s_1-1+2s_1}{2+s_1}} \leq m^{-\frac{2s_1}{2+s_1}}.$$

Therefore we have

$$\mathrm{Tr}(\mathcal{R}_{m+1}) \leq 4D_2(s_1, s_*) m^{-\frac{2s_1}{2+s_1}}.$$

We complete the proof by setting the constant (recall that $(t)_+ := \max\{t, 0\}$ for any $t \in \mathbb{R}$)

$$\begin{aligned} C_2 =& 4D_2(s_1, s_*) = 16(s_1^{s_1}+1)^2 \|L^{-s_1} x^0\|^2 + 8 s_*^{-1} \left((1-s_*)^{1-s_*} + 1\right) \sigma^2 \mathrm{Tr}(L^{s_*}) + \\ & 4 \max\left\{ 4\left(s_1^{s_1}+1\right)^2 \|L^{-s_1} x^0\|^2, \left(\|x^0\|^2 + \sigma^2 \mathrm{Tr}(L)\right) / \left(1 - 2^{\frac{2}{2+s_1+(s_1+s_*-1)_+} - 1}\right)\right\} \\ & \times \begin{cases} \frac{3}{2} + \frac{1}{1-2s_1}, & \text{if } 0 < s_1 < 1/2, \\ \frac{3}{2} + s_*^{-1} + \log 2, & \text{if } s_1 = 1/2, \\ \frac{3}{2} + \frac{2s_1}{2s_1 - 1}, & \text{if } s_1 > 1/2. \end{cases} \end{aligned} \tag{48}$$

The proof of Theorem 2.6 is complete. □

Consider a sequence $\{(\xi_i, y_i)\}_{i=1}^m$ from Model $\mathbb{M}(\mathbb{S}, \rho, x^0, m, \sigma^2)$ with $1 \leq m \leq \infty$ and $\sigma^2 \geq 0$. If $m = \infty$, we also denote $m + 1 = \infty$. For the sake of simplicity, we define the sum of an empty set (e.g., a sum with lower index greater than its upper index) as zero. Recall that $r_k = x_k - x^0$. Theorem 4.6 below estimates the expected weak error for Algorithm (2), with general design of the step-sizes. The rates of weak convergence in Theorem 2.5 with RK-VS setting, and in Theorem 2.9 with RK-FH setting, are both corollaries of Theorem 4.6.

Theorem 4.6. *Assume $x^0 \in L^{s_1}(H)$ for some $s_1 > 0$. Let $x' \in L^{s_2}(H)$ be a measurement vector with $s_2 > 0$. Assume $\{\eta_j\}_{j=1}^m \subset (0,1)$ and $\eta_{2,m} := \sum_{j=1}^m \eta_j^2 < 1$ for Algorithm (2). Then for $1 \leq k \leq m$ (or $1 \leq k < \infty$ if $m = \infty$), one has*

$$\mathbb{E}\left[\langle x', r_{k+1}\rangle^2\right] \leq C_4 \left\{ \frac{1}{1 + \left(\sum_{l=1}^k \eta_l\right)^{2s_1+2s_2}} + \sum_{j=1}^k \frac{(1-\eta_{2,m})^{-1} \eta_j^2}{1 + \left(\sum_{l=j+1}^k \eta_l\right)^{2s_2+\frac{1}{2}}} \right\}, \tag{49}$$

where C_4 is a constant independent of $\{\eta_j\}$ or k, and it will be specified in (54) at the end of the proof.

In particular, if we replace x' by a random vector $\xi' \sim (\mathbb{S}, \rho)$, and assume $\mathrm{Tr}(L^{2-4s_0}) < \infty$ *for some $s_0 \in [1/4, 1/2)$, then for $1 \le k \le m$ (or $1 \le k < \infty$ if $m = \infty$),*

$$\mathbb{E}\left[\langle \xi', r_{k+1}\rangle^2\right] \le C_4' \left\{ \frac{1}{1 + \left(\sum_{l=1}^k \eta_l\right)^{2s_1+2s_0}} + \sum_{j=1}^k \frac{(1-\eta_{2,m})^{-1}\eta_j^2}{1 + \left(\sum_{l=j+1}^k \eta_l\right)^{2s_0+\frac{1}{2}}} \right\}, \tag{50}$$

where C_4' is a constant independent of $\{\eta_j\}$ or k, and it will be specified in (55) at the end of the proof.

Proof. Write $U = x' \otimes x'$. Since $\mathbb{E}\left[\langle x', r_{k+1}\rangle^2\right] = \langle U, \mathcal{R}_{k+1}\rangle_{\mathsf{HS}}$, we prove the estimates (49) and (50) with the help of the expansion (39).

First, since $x^0 \in L^{s_1}(H)$ and $x' \in L^{s_2}(H)$, $\|R_L^{-2s_1}(\mathcal{R}_1)\|_{\mathsf{HS}} = \|L^{-s_1}x^0\|^2$ and $\|R_L^{-2s_2}(U)\|_{\mathsf{HS}} = \|L^{-s_2}x'\|^2$. We have

$$\left\langle U, \left(\prod_{l=1}^k W_{\eta_l}\right)(\mathcal{R}_1)\right\rangle_{\mathsf{HS}} = \left\langle R_L^{-2s_2}(U), \left[R_L^{2s_2}\left(\prod_{l=1}^k W_{\eta_l}\right)R_L^{2s_1}\right] R_L^{-2s_1}(\mathcal{R}_1)\right\rangle_{\mathsf{HS}}.$$

Recall that W_η and R_L share the same set $\{\phi_i \otimes \phi_j\}_{i,j\in\mathcal{I}}$ of eigenvectors, and for each eigenvector $\phi_i \otimes \phi_j$, their eigenvalues are $(1-\eta\lambda_i)(1-\eta\lambda_j)$ and $\sqrt{\lambda_i\lambda_j}$ respectively. So these two operators commute, and one estimates the operator norm of $R_L^{s_2}\left(\prod_{l=1}^k W_{\eta_l}\right)R_L^{2s_1}$ on $\mathcal{J}_2(H)$, by Lemma 4.3 as

$$\begin{aligned}\left\|R_L^{2s_2}\left(\prod_{l=1}^k W_{\eta_l}\right)R_L^{2s_1}\right\|_{\mathcal{J}_2(H)\to\mathcal{J}_2(H)} &\le \left(\max_{\tau\in[0,1]} \tau^{s_1+s_2}\prod_{l=1}^k(1-\eta_l\tau)\right)^2 \\ &\le \frac{4((s_1+s_2)^{s_1+s_2}+1)^2}{1+\left(\sum_{l=1}^k \eta_l\right)^{2s_1+2s_2}}.\end{aligned} \tag{51}$$

Therefore

$$\left\langle U, \left(\prod_{l=1}^k W_{\eta_l}\right)(\mathcal{R}_1)\right\rangle_{\mathsf{HS}} \le \frac{(2((s_1+s_2)^{s_1+s_2}+1)\|L^{-s_1}x^0\|\|L^{-s_2}x'\|)^2}{1+\left(\sum_{l=1}^k \eta_l\right)^{2s_1+2s_2}}.$$

When x' is replaced by $\xi' \sim (\mathbb{S}, \rho)$, one has $\mathbb{E}\left[\langle \xi', r_{k+1}\rangle^2\right] = \langle L, \mathcal{R}_{k+1}\rangle_{\mathsf{HS}}$. Assume $\mathrm{Tr}(L^{2-4s_0}) < \infty$. Note that $R_L^{-2s_0}(L) = L^{1-2s_0}$. In the same way, we estimate the error $\langle L, \mathcal{R}_{k+1}\rangle_{\mathsf{HS}}$ according to the expansion (39).

$$\begin{aligned}\left\langle L, \left(\prod_{l=1}^k W_{\eta_l}\right)(\mathcal{R}_1)\right\rangle_{\mathsf{HS}} &= \left\langle L^{1-2s_0}, \left[R_L^{2s_0}\left(\prod_{l=1}^k W_{\eta_l}\right)R_L^{2s_1}\right] R_L^{-2s_1}\mathcal{R}_1\right\rangle_{\mathsf{HS}} \\ &\le \sqrt{\mathrm{Tr}(L^{2-4s_0})}\frac{(2((s_0+s_1)^{s_0+s_1}+1)\|L^{-s_1}x^0\|)^2}{1+\left(\sum_{l=1}^k \eta_l\right)^{2s_0+2s_1}}.\end{aligned}$$

Next,

$$\begin{aligned}&\left\langle U, \sum_{j=1}^k \eta_j^2\left(\prod_{l=j+1}^k W_{\eta_l}\right)(\sigma^2 L)\right\rangle_{\mathsf{HS}} \\ =&\sigma^2\sum_{j=1}^k \eta_j^2\left\langle R_L^{-2s_2}(U), \left[R_L^{2s_2}\left(\prod_{l=j+1}^k W_{\eta_l}\right)R_L^{1/2}\right](L^{1/2})\right\rangle_{\mathsf{HS}} \\ \le&\sigma^2\sum_{j=1}^k \eta_j^2\frac{\sqrt{\mathrm{Tr}(L)}(2((s_2+\frac{1}{4})^{s_2+\frac{1}{4}}+1)\|L^{-s_2}x'\|)^2}{1+(\sum_{l=j+1}^k \eta_l)^{2s_2+\frac{1}{2}}}.\end{aligned}$$

Similarly, when x' is replaced by the random vector ξ', we use the assumption $\mathrm{Tr}(L^{2-4s_0}) < \infty$, and Bound (51) (by replacing s_1 and s_2 both with s_0) to have

$$\begin{aligned}
&\left\langle L, \sum_{j=1}^{k} \eta_j^2 \left(\prod_{l=j+1}^{k} W_{\eta_l} \right) L \right\rangle_{\mathsf{HS}} = \sum_{j=1}^{k} \eta_j^2 \left\langle L^{1-2s_0}, R_L^{2s_0} \left(\prod_{l=j+1}^{k} W_{\eta_l} \right) R_L^{2s_0} L^{1-2s_0} \right\rangle_{\mathsf{HS}} \\
\le& \sum_{j=1}^{k} \eta_j^2 \left\langle L^{1-2s_0}, L^{1-2s_0} \right\rangle_{\mathsf{HS}} \frac{4((2s_0)^{2s_0}+1)^2}{1+\left(\sum_{l=j+1}^{k} \eta_l\right)^{4s_0}} \le \sum_{j=1}^{k} \eta_j^2 \frac{8\mathrm{Tr}(L^{2-4s_0})((2s_0)^{2s_0}+1)^2}{1+\left(\sum_{l=j+1}^{k} \eta_l\right)^{2s_0+\frac{1}{2}}},
\end{aligned}$$

where the last inequality is because $s_0 \ge 1/4$ and $\frac{1}{1+t^{\theta_1}} \le \frac{2}{1+t^{\theta_2}}$ for any $t \ge 0$ and $\theta_1 \ge \theta_2 > 0$.

Then, for $X = U$ or $X = L$,

$$\begin{aligned}
&\left\langle X, \sum_{j=1}^{k} \eta_j^2 \left(\prod_{l=j+1}^{k} W_{\eta_l} \right) (-R_L^2)\mathcal{R}_j \right\rangle_{\mathsf{HS}} \\
&= -\sum_{j=1}^{k} \eta_j^2 \mathbb{E} \left\langle r_j, R_L^2 \left(\prod_{l=j+1}^{k} W_{\eta_l} \right) X r_j \right\rangle \le 0.
\end{aligned}$$

Finally, recall that $R_* \preceq S_L$, for any $x \in \mathscr{J}_2(H)$, one applies Lemma B.1 to obtain $\|S_L^{-1/2} R_*^{1/2} x\|_{\mathsf{HS}} \le \|x\|_{\mathsf{HS}}$. A single computation shows

$$\|R_*^{1/2}\mathcal{R}_j\|_{\mathsf{HS}}^2 = \mathrm{Tr}(\mathcal{R}_j R_*(\mathcal{R}_j)) = \mathbb{E}\mathrm{Tr}(P\mathcal{R}_j P P \mathcal{R}_j P) \le \mathrm{Tr}(\mathcal{R}_j)^2.$$

We have

$$\left\langle U, \left(\prod_{l=j+1}^{k} W_{\eta_l} \right) R_*(\mathcal{R}_j) \right\rangle_{\mathsf{HS}} = \left\langle \Upsilon_3, \left[R_L^{2s_2} \left(\prod_{l=j+1}^{k} W_{\eta_l} \right) S_L^{1/2} \right] \Upsilon_4 \right\rangle_{\mathsf{HS}}, \tag{52}$$

where $\Upsilon_3 = R_L^{-2s_2}(U) =: \sum_{i,j\in\mathcal{I}} \Upsilon_3^{ij} \phi_i \otimes \phi_j$ with $\sum_{i,j} (\Upsilon_3^{ij})^2 = \|R_L^{-2s_2}(U)\|_{\mathsf{HS}}^2 = \|L^{-s_2}x'\|^4$, and $\Upsilon_4 = S_L^{-1/2} R_*^{1/2}(R_*^{1/2}(\mathcal{R}_j)) =: \sum_{i,j\in\mathcal{I}} \Upsilon_4^{ij} \phi_i \otimes \phi_j$ with $\sum_{i,j} (\Upsilon_4^{ij})^2 \le \|R_*^{1/2}(\mathcal{R}_j)\|_{\mathsf{HS}}^2 \le \mathrm{Tr}(\mathcal{R}_j)^2 \le \left(\frac{\|x^0\|^2+\sigma^2\mathrm{Tr}(L)\eta_{2,m}}{1-\eta_{2,m}}\right)^2$ (where the last inequality follows Lemma 4.2). Since operators R_L, W_η, and S_L share the same set of eigenvectors $\{\phi_i \otimes \phi_t\}_{i,t\in\mathcal{I}}$, the eigenvalue of $R_L^{2s_2} \left(\prod_{l=j+1}^{k} W_{\eta_l} \right) S_L^{1/2}$ corresponding to the eigenvector $\phi_i \otimes \phi_t$ is bounded by (recall that $4\lambda_i\lambda_t \le (\lambda_i + \lambda_t)^2$)

$$\begin{aligned}
&(\lambda_i\lambda_t)^{s_2} \left(\frac{\lambda_i+\lambda_t}{2} \right)^{1/2} \prod_{l=j+1}^{k} \left(1 - \eta_l(\lambda_i+\lambda_t) + \eta_l^2 \lambda_i\lambda_t \right) \\
\le& \left(\frac{\lambda_i+\lambda_t}{2} \right)^{2s_2+\frac{1}{2}} \prod_{l=j+1}^{k} \left[\left(1 - \eta_l \frac{\lambda_i+\lambda_t}{2} \right)^2 \right] \le \frac{2\left[(2s_2+\frac{1}{2})^{2s_2+\frac{1}{2}}+1\right]}{1+(\sum_{l=j+1}^{k} \eta_l)^{2s_2+\frac{1}{2}}}.
\end{aligned}$$

We further apply the spectral theorem to obtain

$$\left\langle U, \left(\prod_{l=j+1}^{k} W_{\eta_l} \right) R_*(\mathcal{R}_j) \right\rangle_{\mathsf{HS}} \le \|L^{-s_2}x'\|^2 \frac{\|x^0\|^2 + \sigma^2\mathrm{Tr}(L)\eta_{2,m}}{1-\eta_{2,m}} \frac{2\left[(2s_2+\frac{1}{2})^{2s_2+\frac{1}{2}}+1\right]}{1+(\sum_{l=j+1}^{k} \eta_l)^{2s_2+\frac{1}{2}}}. \tag{53}$$

When x' is replaced by a random vector ξ', we use a similar computation to give

$$\begin{aligned}\left\langle L, \left(\prod_{l=j+1}^{k} W_{\eta_l}\right) R_*(\mathcal{R}_j)\right\rangle_{\mathsf{HS}} &= \left\langle L^{1-2s_0}, \left[R_L^{2s_0}\left(\prod_{l=j+1}^{k} W_{\eta_l}\right) S_L^{1/2}\right]\Upsilon_4\right\rangle_{\mathsf{HS}} \\ &\le \sqrt{\mathrm{Tr}(L^{2-4s_0})}\frac{\|x^0\|^2+\sigma^2\mathrm{Tr}(L)\eta_{2,m}}{1-\eta_{2,m}}\frac{2\left[(2s_0+\frac{1}{2})^{2s_0+\frac{1}{2}}+1\right]}{1+(\sum_{l=j+1}^{k}\eta_l)^{2s_0+\frac{1}{2}}}.\end{aligned}$$

The proof is completed by combining the above four parts of analysis, and setting the constants

$$\begin{aligned}C_4 = \max\big\{&(2((s_1+s_2)^{s_1+s_2}+1)\|L^{-s_1}x^0\|\|L^{-s_2}x'\|)^2, \\ &\sigma^2\sqrt{\mathrm{Tr}(L)}(2((s_2+\frac{1}{4})^{s_2+\frac{1}{4}}+1)\|L^{-s_2}x'\|)^2, \\ &\|L^{-s_2}x'\|^2\left(\|x^0\|^2+\sigma^2\mathrm{Tr}(L)\right)\times 2\left[(2s_2+\frac{1}{2})^{2s_2+\frac{1}{2}}+1\right]\big\},\end{aligned} \tag{54}$$

and

$$\begin{aligned}C_4' = \max\big\{&(2((s_1+s_0)^{s_1+s_0}+1)\|L^{-s_1}x^0\|)^2\sqrt{\mathrm{Tr}(L^{2-4s_0})}, \\ &8\mathrm{Tr}(L^{2-4s_0})((2s_0)^{2s_0}+1)^2, \\ &\sqrt{\mathrm{Tr}(L^{2-4s_0})}\left(\|x^0\|^2+\sigma^2\mathrm{Tr}(L)\right)\times 2\left[(2s_0+\frac{1}{2})^{2s_0+\frac{1}{2}}+1\right]\big\}.\end{aligned} \tag{55}$$

□

Proof of Theorem 2.9. Consider Equation (49). Set $\eta_k \equiv m^{-\omega}$. Recall that from (19), when $0 < s_2 < 1/4$, $1-\omega = 2/(3+4s_1)$ and when $s_2 > 1/4$, $1-\omega = 1/(1+2s_1+2s_2)$. Recall $m \ge 2$, so $(\log m)/\log 2 \ge 1$. On the one hand,

$$\begin{aligned}\frac{1}{1+(\sum_{l=1}^{m}\eta_l)^{2s_1+2s_2}} &\le (m^{1-\omega})^{-2s_1-2s_2} \\ &\le \begin{cases} m^{-(4s_1+4s_2)/(3+4s_1)}, & \text{if } 0 < s_2 < 1/4, \\ m^{-(4s_1+1)/(4s_1+3)}(\log m)/\log 2, & \text{if } s_2 = 1/4, \\ m^{-(2s_1+2s_2)/(1+2s_1+2s_2)}, & \text{if } s_2 > 1/4. \end{cases}\end{aligned} \tag{56}$$

On the other hand,

$$\begin{aligned}\sum_{j=1}^{m}\frac{\eta_j^2}{1+\left(\sum_{l=j+1}^{m}\eta_l\right)^{2s_2+\frac{1}{2}}} &= m^{-2\omega}+\sum_{j=1}^{m-1}\frac{m^{-2\omega}}{1+((m-j)m^{-\omega})^{2s_2+\frac{1}{2}}} \\ \le m^{-2\omega}+m^{-2\omega}\int_0^{m-1}\frac{dx}{1+(m^{-\omega}x)^{2s_2+\frac{1}{2}}} &= m^{-2\omega}+m^{-\omega}\int_0^{m^{-\omega}(m-1)}\frac{dx}{1+x^{2s_2+\frac{1}{2}}}.\end{aligned} \tag{57}$$

When $0 < s_2 < 1/4$, $-2\omega = \frac{-2-8s_1}{3+4s_1} < -2\frac{4s_1+4s_2}{3+4s_1}$, we apply Lemma 4.5 to obtain

$$\begin{aligned}m^{-2\omega}+m^{-\omega}\int_0^{m^{-\omega}(m-1)}\frac{dx}{1+x^{2s_2+\frac{1}{2}}} &\le m^{-2\omega}+\frac{1}{\frac{1}{2}-2s_2}m^{-\omega+(1-\omega)(\frac{1}{2}-2s_2)} \\ \le \left(1+\frac{1}{\frac{1}{2}-2s_2}\right)m^{(1-4s_2-1-4s_1)/(3+4s_1)} &= \frac{3-4s_2}{1-4s_2}m^{-(4s_1+4s_2)/(3+4s_1)}.\end{aligned}$$

When $s_2 = 1/4$, $\omega = \frac{2s_1+2s_2}{1+2s_1+2s_2} = \frac{1+4s_1}{3+4s_1} < 1$. Recall that $m^{-\omega} \leq 1 \leq 2\log m$, so

$$\begin{aligned} m^{-2\omega} + m^{-\omega} \int_0^{m^{-\omega}(m-1)} \frac{dx}{1+x^{2s_2+\frac{1}{2}}} &\leq m^{-2\omega} + m^{-\omega}\log(m^{1-\omega}+1) \\ &\leq 2m^{-\omega}\log m + m^{-\omega}\log(2m^{1-\omega}) \\ &\leq 2m^{-\omega}\log m + m^{-\omega}\log 2 + (1-\omega)m^{-\omega}\log m \leq 4m^{-\omega}\log m. \end{aligned}$$

When $s_2 > 1/4$, $\omega = \frac{2s_1+2s_2}{1+2s_1+2s_2}$ and

$$m^{-2\omega} + m^{-\omega} \int_0^{m^{-\omega}(m-1)} \frac{dx}{1+x^{2s_2+\frac{1}{2}}} \leq \frac{8s_2}{4s_2-1} m^{-\omega}.$$

Finally, the assumptions $\omega > 1/2$ and $m \geq 2$ yield the bound $\eta_{2,m} = m\eta^2 = m^{1-2\omega} \leq 2^{1-2\omega} < 1$. We combine the above analysis for different s_2, (56), and (57) to obtain (20) with a constant C_3 given by

$$\begin{aligned} C_3 = \max\big\{&(2((s_1+s_2)^{s_1+s_2}+1)\|L^{-s_1}x^0\|\|L^{-s_2}x'\|)^2, \\ &\sigma^2\sqrt{\mathrm{Tr}(L)}(2((s_2+\frac{1}{4})^{s_2+\frac{1}{4}}+1)\|L^{-s_2}x'\|)^2, \\ &\|L^{-s_2}x'\|^2\left(\|x^0\|^2+\sigma^2\mathrm{Tr}(L)\right) \times 2\left[(2s_2+\frac{1}{2})^{2s_2+\frac{1}{2}}+1\right]\big\}\times \\ &\begin{cases} 1+\frac{3-4s_2}{(1-4s_2)(1-2^{1-2\omega})}, & \text{if } 0 < s_2 < 1/4, \\ \frac{1}{\log 2}+\frac{4}{1-2^{1-2\omega}}, & \text{if } s_2 = 1/4, \\ 1+\frac{8s_2}{(4s_2-1)(1-2^{1-2\omega})}, & \text{if } s_2 > 1/4. \end{cases} \end{aligned} \tag{58}$$

The estimate (21) is obtained by substituting s_2 with s_0 in the above analysis. Note that now we have $\eta_{2,m} = m\eta^2 \leq 2^{1-2\omega'} < 1$. The constant C_3' is defined by

$$\begin{aligned} C_3' = \max\big\{&(2((s_1+s_0)^{s_1+s_0}+1)\|L^{-s_1}x^0\|)^2\sqrt{\mathrm{Tr}(L^{2-4s_0})}, \\ &8\mathrm{Tr}(L^{2-4s_0})((2s_0)^{2s_0}+1)^2, \\ &\sqrt{\mathrm{Tr}(L^{2-4s_0})}\left(\|x^0\|^2+\sigma^2\mathrm{Tr}(L)\right)\times 2\left[(2s_0+\frac{1}{2})^{2s_0+\frac{1}{2}}+1\right]\big\}\times \\ &\begin{cases} \frac{1}{\log 2}+\frac{4}{1-2^{1-2\omega'}}, & \text{if } s_0 = 1/4, \\ 1+\frac{8s_0}{(4s_0-1)(1-2^{1-2\omega'})}, & \text{if } 1/4 < s_0 < 1/2. \end{cases} \end{aligned} \tag{59}$$

The proof of Theorem 2.9 is complete. □

Lemma 4.7. *For any $b \geq 2$, $1/2 < \theta < 1$, and $s \geq 0$,*

$$\int_1^b \frac{x^{-2\theta}dx}{1+(b^{1-\theta}-x^{1-\theta})^s} \leq C_5 b^{\max\{-\theta,-s(1-\theta)\}}. \tag{60}$$

where C_5 is a constant depending on s but independent of b, and will be specified in (66), (67), and (68), respectively, for different values of s, in the proof.

The technical proof of Lemma 4.7 is put to Appendix A.

Proof of Theorem 2.5. Note that by definition $1/2 < \omega < 1$, so

$$\eta_{2,\infty} = \sum_{j=1}^{\infty}\eta_j^2 \leq \frac{2\omega-1}{2\omega}\sum_{j=1}^{\infty} j^{-2\omega} < \frac{2\omega-1}{2\omega}\left(1+\int_1^{\infty} x^{-2\omega}dx\right) = 1,$$

where the condition in Theorem 4.6 is satisfied. To finish the proof, we only need to substitute the step-sizes into the inequalities (49) and (50) respectively.

First, for any $k \geq 1$, one has $\frac{k+1}{2} \geq 1$. So

$$\sum_{l=1}^{k} \eta_l \geq \eta_1 \int_1^{k+1} x^{-\omega} dx = \frac{\eta_1}{1-\omega}[(k+1)^{1-\omega} - 1] \geq \frac{\eta_1(1-(1/2)^{1-\omega})}{1-\omega}(k+1)^{1-\omega}.$$

Therefore,

$$\frac{1}{1+\left(\sum_{l=1}^{k}\eta_l\right)^{2s_1+2s_2}} \leq \left[\frac{1-\omega}{\eta_1(1-(1/2)^{1-\omega})}\right]^{2s_1+2s_2} (k+1)^{-(1-\omega)(2s_1+2s_2)}.$$

Note that by the definition $\omega = (2s_1+2s_2)/(1+2s_1+2s_2)$,

$$(k+1)^{-(1-\omega)(2s_1+2s_2)} = (k+1)^{-\omega} \tag{61}$$

Then, we bound the second part of the right-hand side of (49) below. When $k \geq j+1$,

$$\sum_{l=j+1}^{k} \eta_l \geq \eta_1 \int_{j+1}^{k+1} x^{-\omega} dx = \frac{\eta_1\left[(k+1)^{1-\omega} - (j+1)^{1-\omega}\right]}{1-\omega}.$$

Note that $j \geq 1$ implies $j \geq (j+2)/3$. So

$$\begin{aligned}
&\sum_{j=1}^{k} \frac{\eta_j^2}{1+\left(\sum_{l=j+1}^{k}\eta_l\right)^{2s_2+\frac{1}{2}}} \\
\leq &\eta_1^2 k^{-2\omega} + \eta_1^2 \sum_{j=1}^{k-1} \frac{(1/3)^{-2\omega}(j+2)^{-2\omega}}{1+\left(\frac{\eta_1}{1-\omega}\right)^{2s_2+\frac{1}{2}}[(k+1)^{1-\omega} - (j+1)^{1-\omega}]^{2s_2+\frac{1}{2}}} \\
\leq &\eta_1^2 k^{-2\omega} + \frac{9^\omega \eta_1^2}{\min\left\{1, \left(\frac{\eta_1}{1-\omega}\right)^{2s_2+\frac{1}{2}}\right\}} \sum_{j=1}^{k-1} \frac{(j+2)^{-2\omega}}{1+[(k+1)^{1-\omega} - (j+1)^{1-\omega}]^{2s_2+\frac{1}{2}}}.
\end{aligned}$$

Recall that $s_1+s_2 > 1/2$, so $\omega = \frac{2s_1+2s_2}{1+2s_1+2s_2} \in (1/2, 1)$. In the inequality below, for any $j = 1, \ldots, k-1$, we let $x \in [j+1, j+2]$ to have $(j+2)^{-2\omega} \leq x^{-2\omega}$ and $(j+1)^{1-\omega} \leq x^{1-\omega}$. By Lemma 4.7,

$$\begin{aligned}
&\sum_{j=1}^{k-1} \frac{(j+2)^{-2\omega}}{1+[(k+1)^{1-\omega} - (j+1)^{1-\omega}]^{2s_2+\frac{1}{2}}} \leq \int_2^{k+1} \frac{x^{-2\omega}dx}{1+[(k+1)^{1-\omega} - x^{1-\omega}]^{2s_2+\frac{1}{2}}} \\
\leq &C_5(k+1)^{\max\{-\omega, -(2s_2+\frac{1}{2})(1-\omega)\}}.
\end{aligned}$$

Since $k \geq 1$ implies $k \geq (k+1)/2$, the term $k^{-2\omega}$ is bounded by $k^{-2\omega} \leq k^{-\omega} \leq 2^\omega (k+1)^{-\omega}$. We apply Theorem 4.6 and summarize the above analysis to obtain

$$\mathbb{E}\left[\langle x', r_{k+1}\rangle^2\right] \leq C_1(k+1)^{\max\{-\omega, -(2s_2+\frac{1}{2})(1-\omega)\}}$$

where

$$C_1 = C_4\left(\frac{1-\omega}{\eta_1\left(1-(1/2)^{1-\omega}\right)}\right)^{2s_1+2s_2} + C_4\left(\frac{9^\omega \eta_1^2 C_5}{\min\left\{1, \left(\frac{\eta_1}{1-\omega}\right)^{2s_2+\frac{1}{2}}\right\}} + 2^\omega \eta_1^2\right)\left(1 - \frac{2\omega-1}{2\omega}\sum_{j=1}^{\infty} j^{-2\omega}\right)^{-1}. \tag{62}$$

Here for different scopes of s, C_5 is defined in (66), (67), and (68), respectively. The estimate (14) is proved in the same way as above, with s_2 replaced by s_0 and

$$C_1' = C_4 \left(\frac{1-\omega}{\eta_1 \left(1-(1/2)^{1-\omega}\right)} \right)^{2s_1+2s_0} + C_4 \left(\frac{9^{\omega}\eta_1^2 C_5}{\min\left\{1, \left(\frac{\eta_1}{1-\omega}\right)^{2s_0+\frac{1}{2}}\right\}} + 2^{\omega}\eta_1^2 \right) \left(1 - \frac{2\omega-1}{2\omega}\sum_{j=1}^{\infty} j^{-2\omega} \right)^{-1}. \tag{63}$$

The proof of Theorem 2.5 is complete. □

Acknowledgments

Xin Guo would like to thank Professor Gerlind Plonka-Hoch and Professor Kurt Jetter for their helpful discussions. We appreciate the constructive comments from the anonymous referees. The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project's No. PolyU 15334616], the NSFC/RGC Joint Research Scheme [RGC Project No. N-CityU102/20 and NSFC Project No. 12061160462], Germany/Hong Kong Joint Research Scheme [Project No. G-CityU101/20], the Laboratory for AI-Powered Financial Technologies, and Hong Kong Institute of Data Science. Part of the work of Xin Guo was done when he worked at The Hong Kong Polytechnic University. The corresponding author is Xin Guo.

Appendix

A Some Technical Proofs

In this section of appendix we provide some technical proofs.

Proof of the assertions in Example 2.4. It is evident that $L = \mathbb{E}P = \sum_{i=1}^{\infty} q_i e_i \otimes e_i$. So in this model, $\{(q_i, e_i)\}_{i=1}^{\infty}$ form the orthonormal eigensystem of L. Therefore, for any $s \geq 1$, $R_*(L^s) = \sum_{i=1}^{\infty} q_i (e_i \otimes e_i) L^s (e_i \otimes e_i) = \sum_{i=1}^{\infty} q_i^{1+s} e_i \otimes e_i = L^{1+s}$. Recall that $Q_\eta = I - 2\eta S_L + \eta^2 R_*$. In general, if f is a polynomial without constant term, then $R_* f(L) = L f(L)$, $2S_L f(L) = L f(L) + f(L) L = 2L f(L)$, and thus $Q_\eta(f(L)) = f(L) - 2\eta L f(L) + \eta^2 L f(L) = (I - (2\eta - \eta^2)L) f(L)$.

Since the step-sizes are fixed to be η, (37) implies

$$\begin{aligned} \mathcal{R}_{k+1} = Q_\eta(\mathcal{R}_k) + \sigma^2\eta^2 L = \cdots &= Q_\eta^k(\mathcal{R}_1) + \sigma^2\eta^2 \sum_{j=0}^{k-1} Q_\eta^j(L) \\ &= Q_\eta^k(\mathcal{R}_1) + \sigma^2\eta^2 \sum_{j=0}^{k-1} (I - (2\eta - \eta^2)L)^j L. \end{aligned}$$

Since $Q_\eta^k(\mathcal{R}_1)$ is positive semi-definite,

$$\begin{aligned} \lim_{k\to\infty} \mathbb{E}\left[\|r_{k+1}\|^2\right] = \lim_{k\to\infty} \mathrm{Tr}(\mathcal{R}_{k+1}) &\geq \sigma^2\eta^2 \sum_{j=0}^{\infty}\sum_{i=1}^{\infty} q_i (1 - (2\eta - \eta^2) q_i)^j \\ &= \sigma^2\eta^2 \sum_{i=1}^{\infty} \frac{q_i}{(2\eta - \eta^2) q_i} = \infty. \end{aligned}$$

Now we consider the error in weak sense. We have

$$\begin{aligned}\mathbb{E}\left[\langle r_{k+1}, x'\rangle^2\right] &= \left\langle x' \otimes x', Q_\eta^k(\mathcal{R}_{k+1})\right\rangle_{\mathsf{HS}} \\ &= \left\langle x' \otimes x', Q_\eta^k(\mathcal{R}_1)\right\rangle_{\mathsf{HS}} + \sigma^2\eta^2 \sum_{j=0}^{k-1} \left\langle x' \otimes x', (I-(2\eta-\eta^2)L)^j L\right\rangle_{\mathsf{HS}}.\end{aligned}$$

For any $i, j \in \mathcal{I}$, it is straightforward to see that

$$Q_\eta(e_i \otimes e_j) = e_i \otimes e_j - \eta(q_i + q_j)e_i \otimes e_j + \eta^2\sqrt{q_i q_j}\delta_{ij} e_i \otimes e_j,$$

where δ_{ij} is the Kronecker delta which takes value 1 when $i = j$ and 0 otherwise. Therefore $\{e_i \otimes e_j\}_{i,j=1}^\infty$ form a set of eigenvectors of Q_η, and thus Q_η and R_L commute. By Lemma 4.3,

$$\begin{aligned}&\left\|R_L^{2s_1+2s_2} Q_\eta^k\right\|_{\mathcal{J}_2(H)\to\mathcal{J}_2(H)} \le \max_{x,y\in[0,1]} \left(1 - \eta(x+y) + \eta^2\sqrt{xy}\right)^k (xy)^{s_1+s_2} \\ &\le \max_{x,y\in[0,1]} \left(1 - 2\eta\sqrt{xy} + \eta^2\sqrt{xy}\right)^k (xy)^{s_1+s_2} \le \frac{2(2s_1+2s_2)^{2s_1+2s_2}+2}{1+[k(2\eta-\eta^2)]^{2s_1+2s_2}} \le \frac{2(2s_1+2s_2)^{2s_1+2s_2}+2}{(k\eta)^{2s_1+2s_2}},\end{aligned}$$

therefore $\lim_{k\to\infty} \left\langle x' \otimes x', Q_\eta^k(\mathcal{R}_1)\right\rangle_{\mathsf{HS}} = 0$. On the other hand,

$$\begin{aligned}&\eta^2 \sum_{j=1}^{k-1} \left\langle x' \otimes x', (I-(2\eta-\eta^2)L)^j L\right\rangle_{\mathsf{HS}} = \eta^2 \sum_{j=0}^{k-1} \left\langle L^{-s_2}x', (I-(2\eta-\eta^2)L)^j L^{1+2s_2}(L^{-s_2}x')\right\rangle \\ &\le \eta^2\|L^{-s_2}x'\|^2 \sum_{j=0}^{\infty} \frac{2(1+2s_2)^{1+2s_2}+2}{1+(j\eta)^{1+2s_2}} = O(\eta)\end{aligned}$$

uniformly for $k \ge 1$ as $\eta \to 0$, where the last step is obtained by applying Lemma 4.5, and noting that $s_2 > 0$. We have proved that

$$\limsup_{k\to\infty} \mathbb{E}\left[\left\langle x', x_{k+1} - x^0\right\rangle^2\right] = O(\eta), \quad \text{as } \eta \to 0^+.$$

□

Proof of Lemma 4.3. When $s = 0$, the left-hand side (LHS) of (42) is no greater than 1, and the right-hand side (RHS) of (42) equals 2, so (42) holds true. Below we assume $s > 0$.

When $\tau = 0$, (42) is trivial. When $\tau\eta_l = 1$ for some $l = 1, \dots, k$, (42) is trivial. Below we assume $\tau > 0$ and $\tau\eta_l < 1$ for all l.

When $\sum_{l=1}^k \eta_l = 0$, LHS of (42) is no greater than 1 and RHS of (42) is $2 + 2s^s > 1$, so (42) holds true. Below we assume $\sum_{l=1}^k \eta_l > 0$.

Thanks to the inequality $\log(1-c) \le -c$ for $c < 1$, we have

$$\left[\prod_{l=1}^k (1-\eta_l\tau)\right]\tau^s = \tau^s \exp\left\{\sum_{l=1}^k \log(1-\eta_l\tau)\right\} \le \tau^s \exp\left\{-\tau\sum_{l=1}^k \eta_l\right\}.$$

Some simple calculus shows that since $\sum_{l=1}^k \eta_l > 0$,

$$\tau^s \exp\left\{-\tau\sum_{l=1}^k \eta_l\right\} \le \left(\frac{s}{e\sum_{l=1}^k \eta_l}\right)^s, \tag{64}$$

which is valid on $\tau \in [0,\infty)$, and the maximum is achieved at $\tau = s/\sum_{l=1}^k \eta_l$. Since $\tau \in (0,1]$, while (64) is still valid, we have further that

$$\left[\prod_{l=1}^k (1-\eta_l\tau)\right]\tau^s \le 1. \tag{65}$$

We combine (64) and (65) together with the inequality $\min(a,b) \le 2ab/(a+b)$ for $a,b>0$, and the inequality $\frac{s}{s+t} \le \frac{s+1}{1+t}$ for $s,t>0$ to obtain

$$\left[\prod_{l=1}^{k}(1-\eta_l\tau)\right]\tau^s \le \frac{2s^s}{s^s+(e\sum_{l=1}^{k}\eta_l)^s} \le \frac{2s^s+2}{1+\left(\sum_{l=1}^{k}\eta_l\right)^s}.$$

The proof is complete. □

Proof of Lemma 4.5. When $0 \le \theta < 1$,

$$\int_0^b \frac{dx}{1+x^\theta} \le \int_0^b \frac{dx}{(1+x)^\theta} = \frac{(b+1)^{1-\theta}-1}{1-\theta} \le \frac{b^{1-\theta}+1^{1-\theta}-1}{1-\theta} = \frac{b^{1-\theta}}{1-\theta}.$$

When $\theta = 1$, $\int_0^b \frac{dx}{1+x^\theta} = \log(b+1)$. When $\theta > 1$,

$$\int_0^b \frac{dx}{1+x^\theta} \le \int_0^1 \frac{dx}{1} + \int_1^\infty \frac{dx}{x^\theta} = 1 + \frac{1}{\theta-1} = \frac{\theta}{\theta-1}.$$

□

Proof of Lemma 4.7. We divide the integral interval to two parts, $[1, b/2]$ and $[b/2, b]$. First, since $-2\theta < -1$,

$$\begin{aligned}\int_1^{b/2} \frac{x^{-2\theta}dx}{1+(b^{1-\theta}-x^{1-\theta})^s} &\le \frac{1}{1+(b^{1-\theta}-(b/2)^{1-\theta})^s}\int_1^{b/2} x^{-2\theta}dx \\ &\le \frac{b^{-s(1-\theta)}}{(1-(1/2)^{1-\theta})^s}\cdot\frac{1-(b/2)^{1-2\theta}}{2\theta-1} \le \frac{b^{-s(1-\theta)}}{(1-(1/2)^{1-\theta})^s(2\theta-1)}.\end{aligned}$$

Change a variable $y = b^{1-\theta} - x^{1-\theta}$ to give $dy = -(1-\theta)x^{-\theta}dx$. We apply Lemma 4.5 to have

$$\begin{aligned}\int_{b/2}^b \frac{x^{-2\theta}dx}{1+(b^{1-\theta}-x^{1-\theta})^s} &= \int_0^{b^{1-\theta}-(b/2)^{1-\theta}} \frac{x^{-2\theta+\theta}dy}{(1+y^s)(1-\theta)} \le \frac{(b/2)^{-\theta}}{1-\theta}\int_0^{b^{1-\theta}-(b/2)^{1-\theta}} \frac{dy}{1+y^s} \\ &\le \frac{(b/2)^{-\theta}}{1-\theta}\begin{cases} \frac{1}{1-s}\left[1-(1/2)^{1-\theta}\right]^{1-s} b^{(1-\theta)(1-s)}, & \text{if } 0 \le s < 1, \\ \log\left[(1-(1/2)^{1-\theta})b^{1-\theta}+1\right], & \text{if } s = 1, \\ \frac{s}{s-1}, & \text{if } s > 1. \end{cases}\end{aligned}$$

When $0 \le s < 1$, since $-\theta + (1-\theta)(1-s) = 1-2\theta-s(1-\theta) < -s(1-\theta)$, (60) is proved with

$$C_5 = \frac{1}{(1-(1/2)^{1-\theta})^s(2\theta-1)} + \frac{2^\theta(1-(1/2)^{1-\theta})^{1-s}}{(1-\theta)(1-s)}. \tag{66}$$

When $s = 1$, recall $b \ge 2$. Simple calculus shows that $b^{1-2\theta}\log b \le \frac{1}{e(2\theta-1)}$ with the equality attained at $b = \exp\left(\frac{1}{2\theta-1}\right) > e > 2$. So

$$\log\left[(1-(1/2)^{1-\theta})b^{1-\theta}+1\right] \le \log(2b^{1-\theta}) \le \log 2 + (1-\theta)\log b \le (2-\theta)\log b \le \frac{2-\theta}{e(2\theta-1)}b^{2\theta-1}.$$

Therefore, (60) is proved with

$$C_5 = \frac{1}{(1-(1/2)^{1-\theta})(2\theta-1)} + \frac{2^\theta(2-\theta)}{e(1-\theta)(2\theta-1)}. \tag{67}$$

At the end, when $s > 1$, (60) is obtained with

$$C_5 = \frac{1}{(1-(1/2)^{1-\theta})^s(2\theta-1)} + \frac{2^\theta s}{(1-\theta)(s-1)}. \tag{68}$$

The proof is complete. □

B A Technical Lemma

It is well known that for any two positive semi-definite matrices M_1 and M_2, if $M_1 \preceq M_2$, that is, if $M_2 - M_1$ is positive semi-definite, then $\mathrm{Im}(M_1) \subset \mathrm{Im}(M_2)$, that is, the image of M_1 is a subset of that of M_2. In general, one has the following lemma, which should be well known, but we find it difficult to locate the lemma from the literature, so we include its proof for the sake of completeness.

Lemma B.1. *Let H be a real separable Hilbert space. Let F and G be two bounded linear positive semi-definite operators on H such that $F \preceq G$. Then for any $x \in H$, there exists some $w \in H$ such that $G^{1/2}w = F^{1/2}x$ and $\|w\| \leq \|x\|$.*

As a rough outline, Lemma B.1 is proved by making sense to the vector "$G^{-1/2}F^{1/2}x$", and using it as w. This is done by continuously extending the operator $F^{1/2}G^{-1/2}$ from $G^{1/2}(H)$ to its closure, and then taking the adjoint.

Proof of Lemma B.1. Since G is a bounded positive semi-definite operator, there exists a unique bounded positive semi-definite operator $G^{1/2}$, such that $(G^{1/2})^2 = G$. See, for example, [4, Theorem 4]. We define $F^{1/2}$ in the same way for F.

Write $\mathrm{Im}(G^{1/2}) := G^{1/2}(H)$. Recall that $\overline{\mathrm{Im}(G^{1/2})}$ is the orthogonal complement of $\mathrm{Null}(G^{1/2})$ (the null space of $G^{1/2}$) in H. For any $\xi \in \mathrm{Im}(G^{1/2})$, write $G^{-1/2}\xi$ the corresponding preimage vector in $\overline{\mathrm{Im}(G^{1/2})}$ (the existence is evident). The vector $G^{-1/2}\xi$ is uniquely defined, otherwise the difference between any two such vectors falls into $\mathrm{Null}(G^{1/2}) \cap \overline{\mathrm{Im}(G^{1/2})} = \{0\}$, making a contradiction.

The map $G^{-1/2} : \mathrm{Im}(G^{1/2}) \to \overline{\mathrm{Im}(G^{1/2})}$ is a linear operator. In fact, for any $\xi, \eta \in \mathrm{Im}(G^{1/2})$ and any $b \in \mathbb{R}$, $G^{-1/2}\xi + bG^{-1/2}\eta \in \overline{\mathrm{Im}(G^{1/2})}$, and $G^{1/2}(G^{-1/2}\xi + bG^{-1/2}\eta) = \xi + b\eta$, therefore $G^{-1/2}(\xi + b\eta) = G^{-1/2}\xi + bG^{-1/2}\eta$.

Consider $F^{1/2}G^{-1/2} : \mathrm{Im}(G^{1/2}) \to H$. For any $\xi \in \mathrm{Im}(G^{1/2})$, one writes $\eta = G^{-1/2}\xi$ to have

$$\left\|F^{1/2}G^{-1/2}\xi\right\|^2 = \langle \eta, F\eta \rangle \leq \langle \eta, G\eta \rangle = \|\xi\|^2 .$$

So $F^{1/2}G^{-1/2}$ is a bounded linear operator on $\mathrm{Im}(G^{1/2})$. Write A the continuous extension of $F^{1/2}G^{-1/2}$ onto $\overline{\mathrm{Im}(G^{1/2})}$. We further extend A to H by setting $Ax = 0$ for any $x \in \mathrm{Null}(G^{1/2})$. Write A^T the adjoint operator of A. We have $\|A^T\|_{H \to H} = \|A\|_{H \to H} \leq 1$.

For any $x \in H$, we claim that $w = A^T x$ is a vector that satisfies the lemma. In fact, first, $\|w\| = \|A^T x\| \leq \|x\|$. Second, for any $\xi \in \overline{\mathrm{Im}(G^{1/2})}$,

$$\left\langle \xi, G^{1/2}A^T x \right\rangle = \left\langle AG^{1/2}\xi, x \right\rangle = \left\langle F^{1/2}\xi, x \right\rangle = \left\langle \xi, F^{1/2}x \right\rangle .$$

When $\xi \in \mathrm{Null}(G^{1/2})$, $0 \leq \|F^{1/2}\xi\|^2 \leq \|G^{1/2}\xi\|^2 = 0$, so one still has $\left\langle \xi, G^{1/2}A^T x \right\rangle = \left\langle \xi, F^{1/2}x \right\rangle$. Therefore $G^{1/2}w = G^{1/2}A^T x = F^{1/2}x$. This completes the proof. □

References

[1] Zeyuan Allen-Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, pages 1110–1119, 2016.

[2] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

[3] Heinz H. Bauschke and Jonathan M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Rev.*, 38(3):367–426, 1996.

[4] S. J. Bernau. The square root of a positive self-adjoint operator. *J. Austral. Math. Soc.*, 8:17–36, 1968.

[5] Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2576–2586. Curran Associates, Inc., 2020.

[6] Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Found. Comput. Math.*, 18(4):971–1013, 2018.

[7] Yair Censor. Row-action methods for huge and sparse systems and their applications. *SIAM Rev.*, 23(4):444–466, 1981.

[8] Nicolò Cesa-Bianchi, Philip M. Long, and Manfred K. Warmuth. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Trans. Neural Networks*, 7(3):604–619, 1996.

[9] Xuemei Chen and Alexander M. Powell. Almost sure convergence of the Kaczmarz algorithm with random measurements. *J. Fourier Anal. Appl.*, 18(6):1195–1214, 2012.

[10] Xuemei Chen and Alexander M. Powell. Randomized subspace actions and fusion frames. *Constr. Approx.*, 43(1):103–134, 2016.

[11] P. L. Combettes. Hilbertian convex feasibility problem: convergence of projection methods. *Appl. Math. Optim.*, 35(3):311–330, 1997.

[12] Frank Deutsch. *Best Approximation in Inner Product Spaces*, volume 7 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC.* Springer-Verlag, New York, 2001.

[13] Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large stepsizes. *Ann. Statist.*, 44(4):1363–1399, 2016.

[14] Nelson Dunford and Jacob T. Schwartz. *Linear Operators. Part II.* Wiley Classics Library. John Wiley & Sons, Inc., New York, 1988.

[15] René Escalante and Marcos Raydan. *Alternating Projection Methods*, volume 8 of *Fundamentals of Algorithms.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011.

[16] Qin Fang, Min Xu, and Yiming Ying. Faster convergence of a randomized coordinate descent method for linearly constrained optimization problems. *Anal. Appl.*, 16(5):741–755, 2018.

[17] Aurél Galántai. *Projectors and Projection Methods*, volume 6 of *Advances in Mathematics (Dordrecht).* Kluwer Academic Publishers, Boston, MA, 2004.

[18] Robert M. Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM J. Matrix Anal. Appl.*, 36(4):1660–1690, 2015.

[19] Michael Griebel and Peter Oswald. Schwarz iterative methods: infinite space splittings. *Constr. Approx.*, 44(1):121–139, 2016.

[20] Michael Griebel and Peter Oswald. Stochastic subspace correction in Hilbert space. *Constr. Approx.*, 48(3):501–521, 2018.

[21] Xin Guo. Learning gradients via an early stopping gradient descent method. *J. Approx. Theory*, 162(11):1919–1944, 2010.

[22] Xin Guo, Jun Fan, and Ding-Xuan Zhou. Sparsity and error analysis of empirical feature-based regularization schemes. *J. Mach. Learn. Res.*, 17:Paper No. 89, 34, 2016.

[23] Xin Guo, Lexin Li, and Qiang Wu. Modeling interactive components by coordinate kernel polynomial models. *Mathematical Foundations of Computing*, 3(4):263–277, 2020.

[24] Zheng-Chu Guo and Lei Shi. Fast and strong convergence of online learning algorithms. *Adv. Comput. Math.*, 45(5-6):2745–2770, 2019.

[25] Israel Halperin. The product of projection operators. *Acta Sci. Math. (Szeged)*, 23:96–99, 1962.

[26] Gabor T. Herman. *Fundamentals of Computerized Tomography*. Advances in Pattern Recognition. Springer, Dordrecht, second edition, 2009. Image reconstruction from projections.

[27] Stefan Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres*, 35:355–357, 1937.

[28] John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. *J. Mach. Learn. Res.*, 10:777–801, 2009.

[29] Junhong Lin and Lorenzo Rosasco. Optimal learning for multi-pass stochastic gradient methods. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4556–4564. Curran Associates, Inc., 2016.

[30] Junhong Lin and Ding-Xuan Zhou. Learning theory of randomized Kaczmarz algorithm. *J. Mach. Learn. Res.*, 16:3341–3365, 2015.

[31] Ji Liu and Stephen J. Wright. An accelerated randomized Kaczmarz algorithm. *Math. Comp.*, 85(297):153–178, 2016.

[32] Karl Löwner. Über monotone Matrixfunktionen. *Math. Z.*, 38(1):177–216, 1934.

[33] Deanna Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT*, 50(2):395–403, 2010.

[34] Yu. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

[35] Anupan Netyanun and Donald C. Solmon. Iterated products of projections in Hilbert space. *Amer. Math. Monthly*, 113(7):644–648, 2006.

[36] Peter Oswald and Weiqi Zhou. Convergence analysis for Kaczmarz-type methods in a Hilbert space framework. *Linear Algebra Appl.*, 478:131–161, 2015.

[37] Gert K. Pedersen. Some operator monotone functions. *Proc. Amer. Math. Soc.*, 36:309–310, 1972.

[38] Frank Schöpfer and Dirk A. Lorenz. Linear convergence of the randomized sparse Kaczmarz method. *Math. Program.*, 173(1-2, Ser. A):509–536, 2019.

[39] Wei Shen, Zhenhuan Yang, Yiming Ying, and Xiaoming Yuan. Stability and optimization error of stochastic gradient descent for pairwise learning. *Anal. Appl.*, 18(5):887–927, 2020.

[40] Wen-Jun Shen, Yu Ting Wei, Xin Guo, Stephen Smale, Hau-San Wong, and Shuai Cheng Li. Mhc binding prediction with kernelrlspan and its variations. *Journal of Immunological Methods*, 406:10–20, 2014.

[41] Wen-Jun Shen, Hau-San Wong, Quan-Wu Xiao, Xin Guo, and Stephen Smale. Introduction to the peptide binding problem of computational immunology: new results. *Found. Comput. Math.*, 14(5):951–984, 2014.

[42] Steve Smale and Yuan Yao. Online learning algorithms. *Found. Comput. Math.*, 6(2):145–170, 2006.

[43] Thomas Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.*, 15(2):262–278, 2009.

[44] Aditya Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of SGD for least-squares in the interpolation regime, 2021.

[45] Cheng Wang and Ting Hu. Online minimum error entropy algorithm with unbounded sampling. *Anal. Appl.*, 17(2):293–322, 2019.

[46] Norbert Wiener. On the factorization of matrices. *Comment. Math. Helv.*, 29:97–111, 1955.

[47] Yuan Yao. On complexity issues of online learning algorithms. *IEEE Trans. Inform. Theory*, 56(12):6470–6481, 2010.

[48] Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Found. Comput. Math.*, 8(5):561–596, 2008.

[49] Yiming Ying and Ding-Xuan Zhou. Online regularized classification algorithms. *IEEE Trans. Inform. Theory*, 52(11):4775–4788, 2006.

[50] Kôsaku Yosida. *Functional Analysis*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the sixth (1980) edition.

[51] Ding-Xuan Zhou. Deep distributed convolutional neural networks: universality. *Anal. Appl.*, 16(6):895–919, 2018.

[52] Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Appl. Comput. Harmon. Anal.*, 48(2):787–794, 2020.