Compressed Online Learning of Conditional Mean Embedding

author names withheld

Editor: Under Review for COLT 2024

Abstract

The conditional mean embedding (CME) encodes Markovian stochastic kernels through their actions on probability distributions embedded within the reproducing kernel Hilbert spaces (RKHS). The CME plays a key role in several well-known machine learning tasks such as reinforcement learning, analysis of dynamical systems, etc. We present an algorithm to learn the CME *incrementally* from data via an operator-valued stochastic gradient descent. As is well-known, function learning in RKHS suffers from scalability challenges from large data. We utilize a *compression* mechanism to counter the scalability challenge. The core contribution of this paper is a finite-sample performance guarantee on the *last iterate* of the online compressed operator learning algorithm with fast-mixing Markovian samples, when the target CME may not be contained in the hypothesis space. We illustrate the efficacy of our algorithm by applying it to the analysis of an example dynamical system.

Keywords: Online learning, Reproducing kernel Hilbert space, Conditional mean embedding Dynamical systems

1. Introduction

Stochastic kernels can be studied through their actions on reproducing kernel Hilbert space (RKHS) through the conditional mean embedding (CME). Computing expectations of distributions typically involves high-dimensional integrations. The CME reduces this problem to lightweight dimension-free inner product calculations. Such a non-parametric computational framework avoids subscriptions to specific parametric descriptions of functions and probability distributions, and has found applications in reinforcement learning Grunewalder et al. (2012); Lever et al. (2016), analysis of nonlinear dynamical systems Klus et al. (2020); Kostic et al. (2022); Hou et al. (2021, 2023b), and causal inference problems Mitrovic et al. (2018); Muandet et al. (2021), among others.

The CME μ has been defined in two ways in the literature. The first approach in Song et al. (2009) utilizes a composition of covariance operators and requires that the RKHS be closed under the action of the corresponding stochastic kernel. The second approach in Grünewälder et al. (2012); Park and Muandet (2020) is measure-theoretic and defines the CME as $\mu(x) = \mathbb{E}_{Y|x}[\kappa_Y(\cdot, Y)|X = x]$, where the right-hand side is understood as an $\mathcal{H}_Y$-valued Bochner-integral for random variables X, Y, where $\mathcal{H}_Y$ is an RKHS on the domain of Y with the reproducing kernel κ_Y. This definition allows the CME to be viewed as the solution to a regression problem in vector-valued RKHS and does not require the closure assumption needed for the first approach. In this paper, we leverage the regression viewpoint from Grünewälder et al. (2012); Li et al. (2022) to provide *last iterate performance guarantees on compressed online learning of CME with Markovian data*. Our online learning with Markovian sampling is particularly useful to analyze unknown time-homogenous Markov processes, for which the CME encodes the transition dynamics Klus et al. (2020); Kostic et al. (2022); Hou et al. (2023b).

1.1. Our Contributions

- We devise an online algorithm that learns a compressed representation of the CME. This stands in sharp contrast to prior works in Song et al. (2009); Grünewälder et al. (2012); Talwai et al. (2022); Li et al. (2022); Hou et al. (2023b) that estimate compressed/uncompressed versions of the same from a fixed batch of IID samples. Processing streaming data is particularly useful in sequential decision-making settings, e.g., with changing environments such as in Hazan and Seshadhri (2009); Dahlin et al. (2023) or in nonlinear adaptive filtering in Goodwin and Sin (2014).

- Thanks to the representer theorem Schölkopf et al. (2001); Carmeli et al. (2010), the solution of a regularized regression problem in RKHS can be represented as linear expansions in terms of kernel functions centered at samples. Such representations, however, become burdensome with growth in the size of the input dataset as in Kivinen et al. (2004); Rahimi and Recht (2007); Rudi et al. (2015); Koppel et al. (2019); Hou et al. (2023a). To enable scaling to large data sets, we generalize a variant of the compression scheme in Koppel et al. (2019) from the real-valued case to vector-valued CME estimation, which combats the growth of the complexity of the learned representation.

- As opposed to the operator-theoretic viewpoint in Grünewälder et al. (2012) that requires the target CME to be a part of a suitably-defined vector-valued RKHS, we tackle the so-called *hard learning* setting from Li et al. (2022) with batch processing of data where that assumption is invalid, but merge it with the online sparsification technique of Koppel et al. (2019).

- We adopt the Lypanuov-based argument described in Srikant and Ying (2019); Chen et al. (2022) to study last-iterate convergence guarantees with Markovian sampling. Compared to these papers, we innovate in two ways: (1) we extend their work from Euclidean spaces to the space of Hilbert-Schmidt operators. (2) We handle a compounding bias that arises from compression of the operator representation in each iteration by carefully controlling the step-sizes.

2. RKHS Preliminaries

2.1. Real-valued RKHS

We start by formally defining a real-valued RKHS (see Berlinet and Thomas-Agnan (2011)). A separable Hilbert space on $\mathbb{X}$ with its inner product $\left(\mathcal{H}_X, \langle\cdot,\cdot\rangle_{\mathcal{H}_X}\right)$ of functions $f : \mathbb{X} \to \mathbb{R}$ is an RKHS, if the evaluation functional defined by $\delta_x f = f(x)$ is bounded (continuous) for all $x \in \mathbb{X}$. The Riesz representation theorem implies that for all $f \in \mathcal{H}_X$, there exists an element $\phi(x) \in \mathcal{H}_X$ such that $\delta_x f = \langle f, \phi(x)\rangle_{\mathcal{H}_X}$. Define $\kappa_X : \mathbb{X}\times\mathbb{X} \to \mathbb{R}$ by $\kappa_X(x, x') := \langle \phi\left(x\right), \phi\left(x'\right)\rangle$. Then, κ_X is a positive definite kernel that satisfies $\kappa_X(\cdot, x) \in \mathcal{H}_X$, and $\langle f, \kappa_X(\cdot, x)\rangle_{\mathcal{H}_X} = f(x)$, $\forall x \in \mathbb{X}$, $\forall h \in \mathcal{H}_X$. Such κ_X is called a reproducing kernel and $\phi(x) := \kappa_X(\cdot, x)$ is a feature map. Throughout this paper, we assume that all RKHS in question are separable with bounded measurable kernels, and holds if κ is a continuous kernel on an Euclidean space.

Consider two separable measurable spaces $(\mathbb{X}, \mathcal{B}_X)$ and $(\mathbb{Y}, \mathcal{B}_Y)$ with Borel sigma-field $\mathcal{B}_X$ and $\mathcal{B}_Y$, respectively. Let ρ be a probability measure on $\mathbb{X} \times \mathbb{Y}$ with its marginal on X denoted by ρ_X. Denote $\mathcal{L}_2(\rho_X, \mathbb{R}) := \mathcal{L}_2(\rho_X)$ as the vector space of real-valued square-integrable functions with respect to ρ_X. Equip $\mathcal{L}_2(\rho_X)$ with the norm $\|\cdot\|_\rho$ such that $\|f\|_\rho := \left(\int_{\mathbb{X}} |f\left(x\right)|^2 \,\mathrm{d}\rho_X\right)^{1/2}$ for any $f \in \mathcal{L}_2(\rho_X)$. For any $f \in \mathcal{L}_2(\rho_X)$, its ρ_X-equivalent class comprises all functions $g \in \mathcal{L}_2(\rho_X)$ that

$\rho_X(\{f \neq g\}) = 0$. Let $L_2(\rho_X) := \mathcal{L}_2(\rho_X)_{/\sim}$ be the corresponding quotient space equipped with the norm $\|[f]_\sim\|_{L_2(\rho_X)} = \|f\|_\rho$ for any $f \in \mathcal{L}_2(\rho_X)$. In the sequel, we drop the sub-index $\sim$ for any $[f]_\sim \in L_2(\rho_X)$ and simply denote it by $[f]$. To investigate the relationship between RKHS $\mathcal{H}_X$ and $L_2(\rho_X)$, consider the inclusion map $I_\kappa : \mathcal{H}_X \to L_2(\rho_X)$ which maps a function $h \in \mathcal{H}_X$ to its ρ_X-equivalent class $[h]$. Our algorithm and analysis require the following assumptions.

Assumption 1 *(a)* $\sup_{x\in\mathbb{X}} \sqrt{\kappa_X(x,x)} \leq \sqrt{K} < \infty$, $\sup_{y\in\mathbb{Y}} \sqrt{\kappa_Y(y,y)} \leq \sqrt{K} < \infty$. *(b)* $I_\kappa : \mathcal{H}_X \to L_2(\rho_X)$ *is continuous.*

Part (a) implies $\sup_{x\in\mathbb{X}} \kappa_X(x,x) \leq K$. Part (b) implies that I_κ is an embedding, denoted $\mathcal{H}_X \hookrightarrow L_2(\rho_X)$. Define $[\mathcal{H}_X] := \{[f] \in L_2(\rho_X) : f \in \mathcal{H}_X\}$. Parts (a), (b) imply that I_κ is compact.

2.2. Vector-valued RKHS and Tensor Product Hilbert Space

Let $\mathcal{H}_Y$ be a real-valued Hilbert space and $\mathcal{L}(\mathcal{H}_Y)$ be the Banach space of bounded operators from $\mathcal{H}_Y$ to itself. A $\mathcal{H}_Y$-valued Hilbert space $\left(\mathcal{H}_V, \langle\cdot,\cdot\rangle_{\mathcal{H}_V}\right)$ of functions $v : \mathbb{X} \to \mathcal{H}_Y$ is an $\mathcal{H}_Y$-valued RKHS if for each $x \in \mathbb{X}$, $y \in \mathcal{H}_Y$, the linear functional $v \mapsto \langle y, v(x)\rangle_{\mathcal{H}_Y}$ is bounded. $\mathcal{H}_V$ admits an operator-valued reproducing kernel of positive type $\Gamma : \mathbb{X} \times \mathbb{X} \to \mathcal{L}(\mathcal{H}_Y)$ which satisfies $\langle v(x), y\rangle_{\mathcal{H}_Y} = \langle v, \Gamma(\cdot,x)y\rangle_{\mathcal{H}_V}$ and $\langle y, \Gamma(x,x')y'\rangle_{\mathcal{H}_Y} = \langle \Gamma(\cdot,x)y, \Gamma(\cdot,x')y'\rangle_{\mathcal{H}_V}$ for all $x, x' \in \mathbb{X}$, $y, y \in \mathcal{H}_Y$ and $v \in \mathcal{H}_V$. Throughout this paper, we restrict our attention to the vector-valued RKHS associated with the operator-valued kernel $\kappa_X(x,x')\,\mathrm{Id}_Y$ where Id_Y is the identity map on $\mathcal{H}_Y$ and denote it by $\mathcal{H}_V$.

Consider two separable real-valued Hilbert spaces $\mathcal{H}_X, \mathcal{H}_Y$ on separable measurable spaces $\mathbb{X}$ and $\mathbb{Y}$. A bounded linear operator A is a Hilbert-Schmidt (HS) operator from $\mathcal{H}_X$ to $\mathcal{H}_Y$ if $\sum_{i\in\mathbb{N}} \|Ae_i\|^2_{\mathcal{H}_Y} < \infty$ with $\{e_i\}_{i\in\mathbb{N}}$ an orthonormal basis (ONB) of $\mathcal{H}_X$. The quantity $\|A\|_{\mathrm{HS}} = \left(\sum_{i\in\mathbb{N}} \|Ae_i\|^2_{\mathcal{H}_Y}\right)^{1/2}$ is the Hilbert-Schmidt norm of A and is independent of the choice of ONB. We denote $\mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y)$ as the Hilbert space of HS operators from $\mathcal{H}_X$ to $\mathcal{H}_Y$, endowed with the norm $\|\cdot\|_{\mathrm{HS}}$. See Appendix A for a detailed introduction to HS operators. For $f_1 \in \mathcal{H}_X$ and $f_2 \in \mathcal{H}_Y$, the tensor product $f_1 \otimes f_2$ is defined as a rank-one operator from $\mathcal{H}_Y$ to $\mathcal{H}_X$,

$$(f_1 \otimes f_2)\, g \mapsto \langle g, f_2\rangle_{\mathcal{H}_Y} f_1, \quad \forall g \in \mathcal{H}_Y. \tag{1}$$

This rank-one operator is Hilbert-Schmidt. Given a second operator $f_1' \otimes f_2'$ for $f_1' \in \mathcal{H}_X$, $f_2' \in \mathcal{H}_Y$, their inner product is $\langle f_1 \otimes f_2, f_1' \otimes f_2'\rangle_{\mathrm{HS}} = \langle f_1, f_1'\rangle_{\mathcal{H}_X} \langle f_2, f_2'\rangle_{\mathcal{H}_Y}$. Denote by $\mathcal{H}_X \otimes \mathcal{H}_Y$, the tensor product of two Hilbert spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$ which is the completion of the algebraic tensor product with respect to the norm induced by the aforementioned inner product. Moreover, $\mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y)$ is isometrically isomorphic to $\mathcal{H}_Y \otimes \mathcal{H}_X$, per (Park and Muandet, 2020, Lemma C.1).

Let $L_2(\rho_X, \mathcal{H}_Y)$ be the $\mathcal{H}_Y$-valued Bochner square-integrable functions $y : x \mapsto y(x)$ with values in $\mathcal{H}_Y$ such that $\|y\|_{L_2(\rho_X,\mathcal{H}_Y)} = \left(\int_{\mathbb{X}} \|y(x)\|^2_{\mathcal{H}_Y}\, \mathrm{d}\rho_X\right)^{1/2} < \infty$. Consider a vector-valued RKHS $\mathcal{H}_V$ associated with the operator-valued kernel $\kappa_X(x,x')\,\mathrm{Id}_Y$, where κ_X is a real-valued kernel function on $\mathbb{X}$ and $\mathrm{Id}_Y \in \mathcal{L}(\mathcal{H}_Y)$ is the identity operator on $\mathcal{H}_Y$. Under Assumption 1, we have the lemma below whose proof is deferred to Appendix A.1.

Lemma 1 *Let Assumption 1 hold. Then,* $\mathcal{H}_V$ *and* $\mathcal{H}_Y \otimes \mathcal{H}_X$ *are isomorphic, i.e.,* $\mathcal{H}_V \cong \mathcal{H}_Y \otimes \mathcal{H}_X$. *Moreover,* $\mathcal{H}_Y \otimes [\mathcal{H}_X] \subseteq \mathcal{H}_Y \otimes L_2(\rho_X) \cong L_2(\rho_X, \mathcal{H}_Y)$.

Denote the isomorphism $\iota : L_2(\rho_X, \mathcal{H}_Y) \to \mathcal{H}_Y \otimes L_2(\rho_X)$ and $\iota_\kappa : \mathcal{H}_Y \otimes \mathcal{H}_X \to \mathcal{H}_V$. As we shall see in Section 3, we leverage the three pairs of isomorphism, $\mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y) \cong \mathcal{H}_Y \otimes \mathcal{H}_X$, $L_2(\rho_X, \mathcal{H}) \cong \mathcal{H} \otimes L_2(\rho_X)$ and $\mathcal{H}_V \cong \mathcal{H}_Y \otimes \mathcal{H}_X$, to study the CME learning problem within the space of Hilbert-Schmidt operators.

2.3. Embedding of Probability Distributions

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a σ-algebra $\mathcal{F}$ and a probability measure $\mathbb{P}$. Let $X : \Omega \to \mathbb{X}$ be a random variable with distribution $\mathbb{P}_X$. Let Assumption 1 hold. The *kernel mean embedding* (KME) of $\mathbb{P}_X$ in $\mathcal{H}_X$ is the Bochner integral $\mathsf{KME}_X := \mathbb{E}_X[\kappa_X(X, \cdot)]$, where $\mathbb{E}_X$ is the expectation with respect to $\mathbb{P}_X$. Suppose that $\mathbb{P}(X, Y)$ denotes a joint distribution over $\mathbb{X} \times \mathbb{Y}$, then $\mathbb{P}(X, Y)$ can be embedded into $\mathcal{H}_X \otimes \mathcal{H}_Y$, per Berlinet and Thomas-Agnan (2011), as

$$C_{XY} := \mathbb{E}_{XY}[\phi_X(X) \otimes \phi_Y(Y)], \tag{2}$$

where $\mathbb{E}_{XY}$ is the expectation with respect to $\mathbb{P}(X, Y)$. We call C_{XY} (uncentered) cross-covariance operator. Likewise, the (uncentered) covariance operator is defined as $C_{XX} = \mathbb{E}_X[\phi_X(X) \otimes \phi_X(X)]$, which can be viewed as the embedding of the marginal distribution $\mathbb{P}_X$ in $\mathcal{H}_X \otimes \mathcal{H}_X$. Let $\mathbb{P}_{Y|x}$ denote the conditional distribution of Y, given $X = x \in \mathbb{X}$. The $\mathcal{H}_Y$-valued *conditional mean embedding* (CME) $\mu : \mathbb{X} \to \mathcal{H}_Y$ is defined as

$$\mu(\cdot) := \mathbb{E}_{Y|x}[\kappa_Y(\cdot, Y)|X = \cdot] \in L_2(\rho_X, \mathcal{H}_Y), \tag{3}$$

according to Park and Muandet (2020). In addition, for all $f_Y \in \mathcal{H}_Y$ and $x \in \mathbb{X}$, we have $\mathbb{E}_{Y|x}[f_Y(Y)|X = x] = \langle f_Y, \mu(x) \rangle_{\mathcal{H}_Y}$. CME defined in Song et al. (2009) requires C_{XX} to be injective and $\mathbb{E}_{Y|x}[f(Y)|X = x] \in \mathcal{H}_X$ for all $f \in \mathcal{H}_Y$ and $x \in \mathbb{X}$. Under such conditions, μ admits the closed-form $\mu(x) = C_{YX} C_{XX}^{\dagger} \phi_X(x)$. Since our interest lies in the "hard learning" setting, we bypass such restrictive assumptions (see Klebanov et al. (2020); Park and Muandet (2020)).

3. Compressed Online Learning of the CMEs

3.1. Learning the CMEs via Nonlinear Least Squares Regression

Recall that ρ is a joint distribution over $\mathbb{X} \times \mathbb{Y}$ and ρ_X its marginal on $\mathbb{X}$. Park and Muandet (2020) considers an equivalent definition of μ in (3) as the minimizer of a least squares regression problem in the space of $\mathcal{H}_Y$-valued functions in $L_2(\rho_X, \mathcal{H}_Y)$ as

$$\mu_\star := \operatorname*{argmin}_{g \in L_2(\rho_X, \mathcal{H}_Y)} \int_{\mathbb{X} \times \mathbb{Y}} \|g(x) - \phi_Y(y)\|_{\mathcal{H}_Y}^2 \, \mathrm{d}\rho(x, y). \tag{4}$$

In addition, $\mu_\star$ is almost surely unique per (Park and Muandet, 2020, Theorem 4.2). By the isomorphism in Lemma 1, for every $\mu_\star$, there exists a unique Hilbert-Schmidt operator $U_\star$ mapping from $L_2(\rho_X)$ to $\mathcal{H}_Y$ given by $U_\star = \iota^{-1}(\mu_\star)$. In the sequel, we call $U_\star$ the CME operator. For well-posedness and over-fitting prevention, consider now its regularized variant over $\mathcal{H}_V$ as,

$$\mu_\lambda := \operatorname*{argmin}_{g \in \mathcal{H}_V} \frac{1}{2} \int_{\mathbb{X} \times \mathbb{Y}} \|g(x) - \phi_Y(y)\|_{\mathcal{H}_Y}^2 \, \mathrm{d}\rho(x, y) + \frac{\lambda}{2} \|g\|_{\mathcal{H}_V}^2 \tag{5}$$

with $\lambda > 0$ as the regularization parameter. Again, with $\mu_\lambda \in \mathcal{H}_V$, associate a unique HS-operator $U_\lambda \in \mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y)$ such that

$$\mu_\lambda(x) = \iota_\kappa\left(U_\lambda\right)(x) = U_\lambda \phi_X(x), \tag{6}$$

where ι_κ is the isometric isomorphism defined in Lemma 1. We call this HS-operator U_λ as the regularized CME operator. The isomorphism also suggests that U_λ minimizes the regularized risk $R_\lambda : \mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y) \to \mathbb{R}$,

$$R_\lambda(U) := \frac{1}{2}\mathbb{E}\left[\|\phi_Y(y) - U\phi_X(x)\|^2_{\mathcal{H}_Y}\right] + \frac{\lambda}{2}\|U\|^2_{\mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y)}, \tag{7}$$

over $\mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y)$. We aim to solve (7) incrementally and then recover μ via (6). Key to our incremental learning paradigm is the gradient $\nabla_U R_\lambda(U)$ for $U \in \mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y)$ that we characterize in the next result; the proof is in Appendix B.1.

Lemma 2 $\nabla R_\lambda(U) = UC_{XX} - C_{YX} + \lambda U \in HS(\mathcal{H}_X, \mathcal{H}_Y)$ *for any* $U \in HS(\mathcal{H}_X, \mathcal{H}_Y)$.

Setting $\nabla R_\lambda(U)$ equal to 0 yields $-\lambda U = UC_{XX} - C_{YX}$. It admits the solution $U_\lambda = C_{YX}(C_{XX} + \lambda\mathrm{Id})^{-1}$ which is the definition of the regularized CME in Song et al. (2009), provided it belongs to $\mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y)$. We now study an online algorithm that solves (7) incrementally.

3.2. Online Learning of the CME Operator

Let $\mathcal{D}_t := \{(x_i, y_i)\}_{i=1}^t$ be a collection of t samples where $(x_i, y_i) \in \mathbb{X} \times \mathbb{Y}$ for $i = 1, \ldots, t$. The empirical estimation of U_λ is given by $U_{\lambda,\mathrm{emp}} = C_{YX,\mathrm{emp}}\left(C_{XX,\mathrm{emp}} + \lambda\mathrm{Id}\right)^{-1}$, per Hou et al. (2023a), where $C_{YX,\mathrm{emp}}$, $C_{XX,\mathrm{emp}}$ are the empirical estimates of the covariance operators,

$$C_{XX,\mathrm{emp}} = \frac{1}{t}\sum_{i=1}^{t} \phi_X\left(x_i\right) \otimes \phi_X\left(x_i\right), \quad C_{YX,\mathrm{emp}} = \frac{1}{t}\sum_{i=1}^{t} \phi_Y\left(y_i\right) \otimes \phi_X\left(x_i\right). \tag{8}$$

We next present our main algorithm that solves (7) incrementally using stochastic approximation. Let $\mathbb{T} = \mathbb{N}$ represent time. Let $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{T}}$ be a filtration where $\mathcal{F}_t$ is the σ-field generated by the history of data up to time t. Given a sample pair $(x_t, y_t) \in \mathbb{X} \times \mathbb{Y}$ for $t \in \mathbb{T}$, stochastic approximations based estimations of (cross)-covariance operators are given by

$$\widetilde{C}_{XX}(t) = \phi_X\left(x_t\right) \otimes \phi_X\left(x_t\right), \quad \widetilde{C}_{YX}(t) = \phi_Y\left(y_t\right) \otimes \phi_X\left(x_t\right). \tag{9}$$

Thus, an operator-valued stochastic variant of the gradient given in Lemma (2) is

$$\widetilde{\nabla}_t R_\lambda(x_t, y_t; U) = U\widetilde{C}_{XX}(t) - \widetilde{C}_{YX}(t) + \lambda U \in \mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y), \tag{10}$$

for all $U \in \mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y)$ and $t \in \mathbb{T}$. Assuming $\widetilde{U}_0 = 0$, we select a step-size sequence $\{\eta_t\}_{t\in\mathbb{T}} \subset \mathbb{R}$, and consider the $\mathcal{F}_t$-adapted process $\left\{\widetilde{U}_t\right\}_{t\in\mathbb{T}}$ taking values in $\mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y)$,

$$\widetilde{U}_{t+1} = \widetilde{U}_t - \eta_t \widetilde{\nabla}_t R_\lambda(x_t, y_t; \widetilde{U}_t) = (1 - \lambda\eta_t)\widetilde{U}_t - \eta_t\left(\widetilde{U}_t\widetilde{C}_{XX}(t) - \widetilde{C}_{YX}(t)\right), \quad t \in \mathbb{T}. \tag{11}$$

In what follows, we refer to (11) as the *base* operator-valued stochastic gradient descent (SGD) algorithm. Since $\mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y) \cong \mathcal{H}_Y \otimes \mathcal{H}_X$, we next show that (11) can be written as expansions in terms of elements in $\mathcal{H}_Y \otimes \mathcal{H}_X$. The proof is presented in Appendix D.

Lemma 3 *Let $\{\widetilde{U}_t\}_{t\in\mathbb{T}}$ be the sequence generated by (11). Let $\widetilde{\Phi}_{X,t} := [\phi_X(x_1), \ldots, \phi_X(x_t)]$ and $\widetilde{\Psi}_{Y,t} := [\phi_Y(y_1), \ldots, \phi_Y(y_t)]$. Then $\{\widetilde{U}_t\}_{t\in\mathbb{T}}$ admits the representation,*

$$\widetilde{U}_{t+1} = \sum_{i=1}^{t}\sum_{j=1}^{t} \widetilde{W}_t^{ij} \left(\kappa_Y(y_i,\cdot) \otimes \kappa_X(x_j,\cdot)\right) = \widetilde{\Psi}_{Y,t}\widetilde{W}_t\widetilde{\Phi}_{X,t}^\top, \quad \forall t \in \mathbb{T}, \tag{12}$$

$$\widetilde{W}_t^{ij} = (Id - \lambda\eta_t)\widetilde{W}_{t-1}^{ij},\ \widetilde{W}_t^{it} = -\eta_t \sum_{j=1}^{t-1} \widetilde{W}_{t-1}^{ij}\kappa_X(x_j, x_t),\ \widetilde{W}_t^{tt} = \eta_t,\ t \in \mathbb{T}\setminus\{0\}, \widetilde{W}_0 = 0. \tag{13}$$

The above result states that the iterates $\{\widetilde{U}_t\}_{t\in\mathbb{T}}$ generated by the base operator-valued SGD (11) can be described by a linear combination of kernel functions centered at samples seen thus far. Therefore, the implementation of (11) can be decomposed into two parts–appending the new sample to the current dictionary $\widetilde{\mathcal{D}}_t \leftarrow \widetilde{\mathcal{D}}_{t-1} \cup \{(x_t, y_t)\}$ and updating coefficients according to (13). The complexity of such an algorithm grows with the size of $\widetilde{\mathcal{D}}_t$ that compounds even faster than for function learning. We aim to control that growth of $\widetilde{\mathcal{D}}_t$ by judiciously admitting a new sample only when the new sample brings sufficiently "new" information.

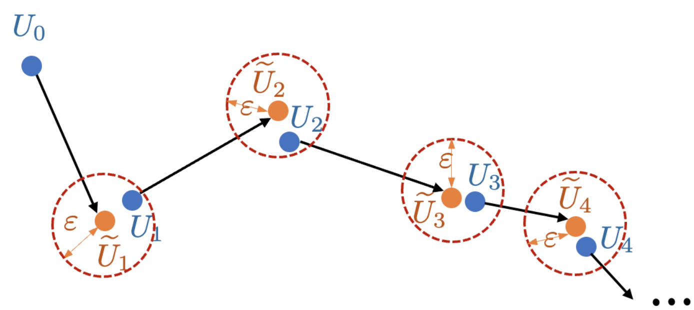


Figure 1: An illustration of our compressed operator-valued stochastic gradient descent: $\{U_t\}_{t\in\mathbb{T}}$ (blue) are the iterates generated by Algorithm 1, and $\left\{\widetilde{U}_t\right\}_{t\in\mathbb{T}}$ (orange) are the auxiliary variables computed based on $\widetilde{\mathcal{D}}_t$. Condition (16) ensures that at each step $t \in \mathbb{T}$, the compressed estimate U_t lies within the ε-ball around $\widetilde{U}_t$.

Denote the corresponding learning sequence by $\{U_t\}_{t\in\mathbb{T}}$. Let $U_0 = 0$, $\mathcal{D}_0 = \emptyset$. After receiving (x_1, y_1), define $\mathcal{D}_1 \leftarrow [(x_1, y_1)]$ and update $U_1 = \widetilde{U}_1 = \eta_1\kappa_Y(y_1,\cdot) \otimes \kappa_X(x_1,\cdot)$. At time $t-1$ for $t \geq 2$, suppose $\mathcal{D}_{t-1}$ is the dictionary which is a subset of all samples encountered up to time $t-1$. Let $\mathcal{I}_{t-1} = \{1, \ldots, |\mathcal{D}_{t-1}|\}$ be the set of indices of $\mathcal{D}_{t-1}$. After receiving a new sample pair (x_t, y_t), we decide whether to add it to the current dictionary $\mathcal{D}_{t-1}$ or discard it based on its contribution to steer the iterates toward the desired direction. More precisely, if we admit the new data into the dictionary, i.e., $\widetilde{\mathcal{D}}_t \leftarrow [\mathcal{D}_{t-1}, (x_t, y_t)]$, then we utilize (11) to update

$$\widetilde{U}_{t+1} = U_t - \widetilde{\nabla}_t R_\lambda\left(x_t, y_t; U_t\right). \tag{14}$$

$\widetilde{W}_t$ is updated according to (13), based on $\widetilde{\mathcal{D}}_t$.

We now test whether $\widetilde{U}_{t+1}$ can be well approximated within the ε-accuracy level by a combination of kernel functions centered at elements in the old dictionary $\mathcal{D}_{t-1}$. To this end, we consider the projection of $\widetilde{U}_{t+1}$ onto the closed subspace span $\{\kappa_Y(y_i,\cdot)\otimes\kappa_X(x_j,\cdot) : i,j\in\mathcal{I}_{t-1}\}$, i.e.,

$$\widehat{U}_{t+1} := \Pi_{\mathcal{D}_{t-1}}\left[\widetilde{U}_{t+1}\right]. \tag{15}$$

We now distinguish between two cases. In the first case, the estimation error due to compression is within a pre-selected compression budget ε for all $t\in\mathbb{T}$, i.e.,

$$\left\|\widehat{U}_{t+1}-\widetilde{U}_{t+1}\right\|_{\mathrm{HS}}\leq\varepsilon. \tag{16}$$

In other words, $\widetilde{U}_{t+1}$ can be ε-well approximated based on $\mathcal{D}_{t-1}\subseteq\widetilde{\mathcal{D}}_t$. Therefore, we discard the new sample (x_t,y_t) and maintain the same dictionary as before, i.e., $\mathcal{D}_t\leftarrow\mathcal{D}_{t-1}$, $\mathcal{I}_t\leftarrow\mathcal{I}_{t-1}$. We then update the coefficients by incorporating the effect of (x_t,y_t) as

$$W_t=\operatorname*{argmin}_{Z\in\mathbb{R}^{|\mathcal{I}_t|\times|\mathcal{I}_t|}}\left\|\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{I}_t}Z^{ij}\kappa_Y(y_i,\cdot)\otimes\kappa_X(x_j,\cdot)-\sum_{i=1}^{t}\sum_{j=1}^{t}\widetilde{W}_t^{ij}\kappa_Y(y_i,\cdot)\otimes\kappa_X(x_j,\cdot)\right\|_{\mathrm{HS}}^2. \tag{17}$$

In the second case, where condition (16) is violated, we append the new sample (x_t,y_t) to the dictionary, i.e., $\mathcal{D}_t\leftarrow\mathcal{D}_{t-1}\cup(x_t,y_t)$. The coefficient matrix is $W_t\leftarrow\widetilde{W}_t$ according to (13). In both cases, the estimate at time $t+1$ can be computed based on $\mathcal{D}_t$ and W_t as

$$U_{t+1}=\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{I}_t}W_t^{ij}\kappa_Y(y_i,\cdot)\otimes\kappa_X(x_j,\cdot). \tag{18}$$

In summary, our approach attains a compressed representation of U_{t+1} by construct, and the complexity of description only depends on the cardinality of $\mathcal{D}_t$ at each $t\in\mathbb{T}$. The implementation of such an algorithm only involves finite-dimensional kernel matrices; see Appendix G for details. The procedure is summarized in Figure 1 and Algorithm 1 in Appendix E. While our algorithm is inspired by that in Koppel et al. (2019), we generalize the framework therein to operator-valued RKHS which is in that we provide last-iterate finite-time guarantees (compared to asymptotic results) with Markovian sampling (compared to i.i.d. samples), and only prune the dictionary by considering the admittance of a new sample (instead of optimizing over the whole dictionary every iteration). As our empirical results will confirm in Section 5, the compressed dictionary satisfies $|\mathcal{D}_t|\ll t$, thus highlighting the efficacy of our compression mechanism.

4. Convergence Analysis of Compressed Online Learning

In this section, we study the convergence behavior of the error $[\mu_t]-\mu_\star$. To this end, we aim to bound the right-hand side of the following relation,

$$\|[\mu_t]-\mu_\star\|_\gamma^2\leq 2\,\|[\mu_t-\mu_\lambda]\|_\gamma^2+2\,\|[\mu_\lambda]-\mu_\star\|_\gamma^2\,. \tag{19}$$

where $\|\cdot\|_\gamma$ is defined via *the power spaces* that we present next. Such spaces are necessary for the analysis due to the fact that $\mu_\star$ may not lie in $\mathcal{H}_V\cong\mathrm{HS}(\mathcal{H}_X,\mathcal{H}_Y)$. The first term corresponds to the sampling error and depends on the stochastic sample path. The second term is deterministic and captures the bias due to regularization and how far $\mu_\star$ lies from $\mathcal{H}_V$.

4.1. The Power Spaces

Recall that we do not assume $\mu_\star \in \mathcal{H}_V$. Instead, following Li et al. (2022), we rather assume $\mu_\star$ being an element of an intermediate vector-valued space $[H_V]^\beta$ that lies between $L_2(\rho_X, \mathcal{H}_Y)$ and $\mathcal{H}_V$ for some $\beta \in (0, 2]$, where β is a measure of regularity of the space $[H_V]^\beta$. To characterize $[H_V]^\beta$, we start by introducing the real-valued power spaces in Steinwart and Scovel (2012) and then present its extension to the vector-valued case given by Li et al. (2022).

Define the integral operator $L_\kappa : L_2(\rho_X) \to L_2(\rho_X)$ associated with a reproducing kernel κ as

$$L_\kappa [f] := \int_{\mathbb{X}} \kappa(\cdot, x) g(x) \mathrm{d}\rho_X(x) \quad \forall g \in [f] \tag{20}$$

for any $[f] \in L_2(\rho_X)$. Under Assumption 1, L_κ is continuous, self-adjoint, positive trace-class, and compact per (Steinwart and Scovel, 2012, Lemmas 2.2 and 2.3). The spectral theorem for self-adjoint compact operators (Steinwart and Christmann, 2008, Theorem A.5.13) indicates that there exists an at most countable index set $\mathbb{I}$ that either $\mathbb{I} = \{1, 2, \cdots, n\}$ or $\mathbb{I} = \mathbb{N}$, a non-increasing, summable sequence $(\sigma_i)_{i\in\mathbb{I}} \in (0, \infty)$ converging to 0 and a family $(e_i)_{i\in\mathbb{I}} \subset \mathcal{H}_X$ such that $([e_i])_{i\in\mathbb{I}} \subset L_2(\rho_X)$ is an orthonormal system (ONS) of $L_2(\rho_X)$, and

$$L_\kappa[f] = \sum_{i\in\mathbb{I}} \sigma_i \langle [f], [e_i] \rangle_\rho [e_i], \quad [f] \in L_2(\rho_X). \tag{21}$$

In addition, $(\sigma_i)_{i\in\mathbb{I}}$ is the family of non-zero eigenvalues of L_κ and $([e_i])_{i\in\mathbb{I}}$ consists of the corresponding eigenvectors of L_κ. For some fixed $\beta \geq 0$, Steinwart and Scovel (2012) define the β-power space $[H_X]^\beta$ using eigensystems of the integral operator L_κ as

$$[H_X]^\beta := \left\{ \sum_{i\in\mathbb{I}} a_i \sigma_i^{\beta/2} [e_i] : (a_i) \in l_2(\mathbb{I}) \right\} \subseteq L_2(\rho_X), \tag{22}$$

equipped with norm $\left\| \sum_{i\in\mathbb{I}} a_i \sigma_i^{\beta/2} [e_i] \right\|_{[H]^\beta} := \|(a_i)_{i\in\mathbb{I}}\|_{l_2(\mathbb{I})}$. To simplify notation, we use $\|\cdot\|_\beta$ instead of $\|\cdot\|_{[H]^\beta}$. For $(a_i)_{i\in\mathbb{I}} \in l_2(\mathbb{I})$, $[H_X]^\beta$ is a separable Hilbert space with ONB $\left(\sigma_i^{\beta/2} [e_i]\right)_{i\in\mathbb{I}}$. In addition, for every $\alpha \in (0, \beta)$, we have the following chain of embeddings $[H_X]^\beta \hookrightarrow [H_X]^\alpha \hookrightarrow [H_X]^0 \subseteq L_2(\rho_X)$ per Steinwart and Scovel (2012).

Analogous reasoning as before allows us to embed the $\mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y)$ into $\mathrm{HS}(L_2(\rho_X), \mathcal{H}_Y)$, and hence, we can define an intermediate space consisting of vector-valued functions per (Li et al., 2022, Definition 3) as below.

$$[H_V]^\beta := \iota\left(\mathrm{HS}\left([H_X]^\beta, \mathcal{H}_Y\right)\right) = \left\{\iota(U) : U \in \mathrm{HS}\left([H_X]^\beta, \mathcal{H}_Y\right)\right\}, \tag{23}$$

equipped with the norm $\|v\|_\beta := \|U\|_{\mathrm{HS}([H_X]^\beta, \mathcal{H}_Y)}$, where ι is the isomorphism in Lemma 1. In the following, we assume $\mu_\star \in [H_V]^\beta$ for some $\beta \in (0, 2]$. When $\beta \geq 1$, $\mu_\star \in \mathcal{H}_V$, while for $\beta \in (0, 1)$, $\mu_\star \notin \mathcal{H}_V$. We call the latter the hard learning setting in that the target operator does not lie in the hypothesis space, and hence, a bias in the estimate even without compression is unavoidable.

4.2. Bounding $\|[\mu_t - \mu_\lambda]\|_\gamma^2$

Recall that we aim to bound the right-hand side of (19). We bound the first term via Lemma 9 in Appendix C.1 to infer

$$\|[\mu_t - \mu_\lambda]\|_\gamma^2 \leq \lambda^{-(\gamma+1)} (K+\lambda)^2 \|U_t - U_\lambda\|_{\text{HS}}^2 . \tag{24}$$

Next, we study the convergence of $\{U_t\}$ to U_λ in HS-norm. Figure 1 illustrates that compression in each iteration induces an extra error at each iterate. To this end, define an $\{\mathcal{F}_t\}_{t\in\mathbb{T}}$-adapted sequence $\{E_t\}_{t\in\mathbb{T}}$ where $E_t := U_{t+1} - \widetilde{U}_{t+1}$ encodes the error due to compression to write the iterates of our algorithm as

$$U_{t+1} = U_t + \eta_t \left(-\widetilde{\nabla} R_\lambda(x_t, y_t; U_t) + \frac{E_t}{\eta_t} \right), \quad U_0 = 0. \tag{25}$$

Here, $\|E_t\|_{\text{HS}} \leq \varepsilon$ from (16). We make the following assumption.

Assumption 2 *(a) Consider a constant step-size sequence $\{\eta_t\}_{t\in\mathbb{T}}$ with $\eta_t = \eta$ for some $0 \leq \eta \leq \min(1, 1/\lambda)$ for all $t \in \mathbb{T}$, and (b) $\varepsilon \leq B_{cmp}\eta^2$ for some $B_{cmp} > 0$.*

We delineate precise requirements on the Markovian data generation process in Assumption 3, presenting which needs the following definition.

Definition 4 *(Exponentially ergodic and mixing time of a Markov process.) Let $\{M_t\}_{t\in\mathbb{T}}$ be a Markov process on a filtered probability space $(\mathbb{M}, \{\mathcal{F}_t\}_{t\in\mathbb{T}}, \mathbb{P})$ where M_t is $\mathcal{F}_t$-adapted. Let $P_t(\cdot|x)$ be a version of the conditional distribution of M_t given $M_0 = x \in \mathbb{M}$, and ρ_X be the unique stationary distribution of the Markov process over $\mathbb{M}$. Then, $\{M_t\}_{t\in\mathbb{T}}$ is exponentially ergodic if there exists some finite $\overline{M} > 0$ and $c \in (0,1)$ such that $\sup_{x\in\mathbb{M}} \mathbb{E}\,\|P_t(\cdot|x) - \rho_X(\cdot)\|_{TV} \leq \overline{M}c^t$, $t \in \mathbb{T}$, where $\|\cdot\|_{TV}$ is the total variation distance. Furthermore, for $\delta > 0$, the mixing time τ_δ of $\{M_t\}_{t\in\mathbb{T}}$ with precision δ is*

$$\tau_\delta := \mathit{min}\left\{ t > 0 : \sup_{x\in\mathbb{M}} \mathbb{E}\,\|P_t(\cdot|x) - \rho_X(\cdot)\|_{TV} \leq \delta \right\}. \tag{26}$$

Assumption 3 *$\{(X_t, Y_t)\}_{t\in\mathbb{T}}$ is exponentially ergodic with a unique stationary distribution ρ. In addition, $P_t(\cdot|\mathcal{F}_0)$ and ρ are absolutely continuous with respect to some underlying measure ν on $\mathbb{X}\times\mathbb{Y}$ for all $t \in \mathbb{T}$.*

Under Assumption 3, the process has sufficiently mixed after τ_δ steps. By (26), this mixing time τ_δ satisfies $\overline{M}c^{\tau_\delta} \leq \delta$ and $\overline{M}c^{\tau_\delta - 1} \geq \delta$, which implies

$$\tau_\delta \leq \frac{\log(\overline{M}/c) + \log(1/\delta)}{\log(1/c)} \leq \frac{\log(\overline{M}/c) + \log(\overline{M}/c)\log(1/\delta)}{\log(1/c)} = B_{\text{mix}}\left(\log\frac{1}{\delta} + 1\right) \tag{27}$$

for $B_{\text{mix}} = \frac{\log(\overline{M}/c)}{\log(1/c)}$. Unlike the IID case, under Markovian sampling, the gradient steps are biased. The following result bounds this bias. Its proof is deferred to Appendix F.1.

Lemma 5 *Let Assumptions 1 and 3 hold. For any $\delta > 0$ and $t \geq \tau_\delta$, we have*

$$\left\| \mathbb{E}\left[\widetilde{\nabla} R_\lambda(x_t, y_t, U) \,|\mathcal{F}_0 \right] - \nabla R_\lambda(U) \right\|_{HS} \leq 2B_\kappa\delta\left(\|U\|_{HS} + 1\right). \tag{28}$$

We adopt a Lyapunov-style analysis to study the compressed operator-valued stochastic approximation in Algorithm 1. The argument closely resembles the (informal) analysis of the continuous-time dynamics $\dot{U}(t) = -\nabla R_\lambda (U(t))$ for $U \in \mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y)$ for which one can show that $dW(U(t))/dt \leq -2\lambda W(U(t))$ for $W(U) = \|U(t) - U_\lambda\|^2_{\mathrm{HS}}$, and then view (25) as its discrete, biased, and stochastic counterpart. In the sequel, we denote

$$\Xi_\lambda := \|U_\lambda\|_{\mathrm{HS}} + 1, \quad B_\kappa := K + \lambda, \quad B = B_\kappa + B_{\mathrm{cmp}}, \quad \check{B} = 98B^2 + 32B.$$

The following result provides the one-time-step drift of the above $W(U_t)$ in expectation; see Appendix F.2 for a proof.

Lemma 6 *Let Assumptions 1,2, and 3 hold. Suppose* $\eta\tau_\eta \leq \min\left\{1/4B, \lambda/\check{B}\right\}$*. Then, for all* $t \geq \tau_\eta$*, we have*

$$\mathbb{E}\left[\|U_{t+1} - U_\lambda\|^2_{HS}\right] \leq (1 - \lambda\eta)\, \mathbb{E}\left[\|U_t - U_\lambda\|^2_{HS}\right] + \check{B}\tau_\eta\eta^2\Xi_\lambda^2 + 4\varepsilon K/\lambda. \tag{29}$$

By (27), $\eta\tau_\eta \to 0$ as $\eta \to 0$, implying that η can be chosen to satisfy the conditions required for the above result. In the analysis of classic SGD in Euclidean spaces, e.g., in Borkar (2009), such one-step inequalities are commonplace. In such analysis, one typically requires the constant terms in the upper bound to scale as $\mathcal{O}(\eta^2)$. We achieve the same with $\varepsilon = \mathcal{O}(\eta^2)$, per Assumption 2(b).

4.3. The Final Result

The one-step inequality established in Lemma 6 allows us to bound $\|U_t - U_\lambda\|_{\mathrm{HS}}$ that in turn bounds the first term on the right-hand side of (19). The second term on the right-hand side of (19) equals the bias in trying to approximate an operator that does not lie in the hypothesis space. By (Li et al., 2022, Lemma 1), we know

$$\|\mu_\lambda - \mu_\star\|^2_\gamma \leq \lambda^{\beta-\gamma} \|\mu\|^2_\beta\,. \tag{30}$$

These two bounds ultimately yield the following main result of this paper; its proof is in Appendix F.3. Define

$$\Theta_1 := 2\lambda^{-(\gamma+1)} (K+\lambda)^2\, \Xi_\lambda^2, \quad \Theta_2 := 2\lambda^{-(\gamma+2)} (K+\lambda)^2 \left(\check{B}\tau_\eta\Xi_\lambda^2 + 4B_{\mathrm{cmp}}K/\lambda\right), \tag{31}$$

where Θ_2 captures the effect of compression through B_{cmp} defined in Assumption 2.

Theorem 7 *(Last iterate convergence with constant step-size) Suppose* $\mu_\star \in [H_V]^\beta$ *for some* $0 < \beta \leq 2$ *and Assumptions 1, 2, and 3 hold. If* $\eta\tau_\eta \leq \min\left\{1/4B, \lambda/\check{B}\right\}$*, then for* $t \geq \tau_\eta$*,*

$$\mathbb{E}\left[\|[\mu_t] - \mu_\star\|^2_\gamma\right] \leq \Theta_1 (1 - \lambda\eta)^{t-\tau_\eta} + \Theta_2\eta + 2\lambda^{\beta-\gamma} \|\mu_\star\|^2_\beta\,, \tag{32}$$

for some $\gamma \in [0, 1]$ *with* $\gamma < \beta$*.*

The above result says that after an initial transient period, while the Markov chain settles down to its stationary distribution, the error decays exponentially fast in the mean square sense and the iterates converge to a ball centered at $\mu_\star$, with a radius depending on the step-size η, the compression budget ε, regularization parameter λ and the hardness of the learning problem encoded in β. The settling time τ_η grows with smaller η, but $\eta\tau_\eta$ can be made arbitrarily small, owing to the assumption that the sampling distribution guided by a Markov process is exponentially ergodic. Notice that we only allow small enough compression budgets in our analysis that depend on the step-size choice, according to Assumption 2(b). In effect, we do not allow the biases introduced due to compression to derail the progress of the online learning algorithm. Furthermore, the convergence is established in the γ-norm that defines norm in $[H_V]^\gamma$, where recall that for the conditions on γ laid out in Theorem 7, we have $[H_V]^\beta \subseteq [H_V]^\gamma \subseteq L_2(\rho_X, \mathcal{H}_Y)$ for $0 \leq \gamma \leq \beta$. In the special case where $\gamma = 0$, we obtain the mean square convergence in the $L(\rho_X, \mathcal{H}_Y)$-norm.

5. Application to Learning Dynamical Systems

Let $\{X_t\}_{t\in\mathbb{T}}$ be a $\mathbb{R}^n$-valued time-homogeneous Markov process defined via the transition kernel density p as $\mathbb{P}\{X_{t+1} \in \mathbb{A} | X_t = x\} = \int_{\mathbb{A}} p(y|x)\mathrm{d}y$ for measurable $\mathbb{A} \subseteq \mathbb{R}^n$. Let $f \in L^1(\mathbb{R}^n)$ be a probability density over $\mathbb{R}^n$ and $g \in L^\infty(\mathbb{R}^n)$ a scalar function of $\mathbb{R}^n$. Then, the Perron–Frobenius (PF) operator $\mathcal{P} : L^1(\mathbb{R}^n) \to L^1(\mathbb{R}^n)$ and the Koopman operator $\mathcal{K} : L^\infty(\mathbb{R}^n) \to L^\infty(\mathbb{R}^n)$ act on f and g, respectively, as

$$(\mathcal{P}f)(y) = \int p(y|x)f(x)\mathrm{d}x, \quad (\mathcal{K}g)(x) = \int p(y|x)g(y)\mathrm{d}y. \tag{33}$$

These transfer operators are infinite-dimensional but linear. When interacting with an RKHS $\mathcal{H}$, they are related to CME as follows. Let X^+ be the system state at the next time-step starting from X, the Koopman operator satisfies

$$\langle \mathcal{K}g, \phi(x)\rangle = \mathbb{E}\left[g(y)|x_t = x\right] = \left\langle g, \mu_{X^+|X}(x)\right\rangle \overset{(a)}{=} \left\langle g, U_{X^+|X}\phi(x)\right\rangle_{\mathcal{H}}, \quad g \in \mathcal{H}, \tag{34}$$

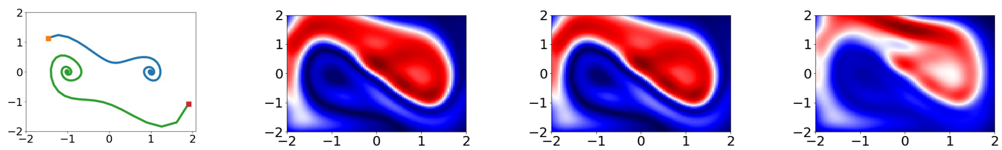
where (a) follows from $U_{X^+|X} = \iota_\kappa^{-1}\left(\mu_{X^+|X}\right)$ as in Lemma 1. We call $U_{X^+|X} \in \mathrm{HS}(\mathcal{H})$ the CME operator. On the other hand, we exploit the definition of KME as in Section 2 to obtain

$$\mathrm{KME}_{X^+} = \mathbb{E}_{X^+}\left[\phi\left(X^+\right)\right] \overset{(a)}{=} \mathbb{E}_X\left[\mathcal{U}_{X^+|X}\phi(X)\right] \overset{(b)}{=} U_{X^+|X}\mathbb{E}_X\left[\phi(X)\right] = U_{X^+|X}\mathrm{KME}_X. \tag{35}$$

In the above derivation, (a) follows from the law of total expectation and the definitions of μ, U in (3), while (b) follows from the linearity $\mathcal{U}_{X^+|X}$. The above result suggests we can identify $\mathcal{P}$ which acts on $\mathcal{H}$ as the CME operator, i.e., $\mathcal{P} = U_{X^+|X}$ and $\mathcal{P} : \mathrm{KME}_X \mapsto \mathrm{KME}_{X^+}$ propagates the embedded distribution of states through the system dynamics. Furthermore, $\mathcal{K}$ then becomes the adjoint of $U_{X^+|X}$, i.e., $\mathcal{K} = U^*_{X^+|X}$. We remark that the prevalent way of relating transfer operators to CME in Klus et al. (2020); Hou et al. (2023b) goes through the covariance operator route in Song et al. (2009) that requires $\mathcal{H}$ to be closed under the action of the system dynamics that we do not require.

Let $d_t = |\mathcal{D}_t|$. Define matrices $\Phi_{X,t} = [\phi_X(x_1), \ldots, \phi_X(x_{d_t})]$, $\Psi_{Y,t} = [\phi_Y(y_1), \ldots, \phi_Y(y_{d_t})]$, and $G_{YX,t} = \Psi_{Y,t}^\top \Phi_{X,t}$. By Lemma 3, the iterates $\{U_t\}_{t\in\mathbb{T}}$ generated by Algorithm 1 can be expressed as $U_t = \Psi_{Y,t}W_t\Phi_{X,t}^\top$ for all $t \in \mathbb{T}$. Therefore, the compressed Koopman operator becomes $\widehat{\mathcal{K}}_t = \Phi_{X,t}W_t^\top\Psi_{Y,t}^\top$. From Klus et al. (2020, Proposition 3.1), the eigenfunction φ_λ of $\mathcal{K}_t$ associated with eigenvalue λ can then be computed as $\varphi_\lambda(x) = (\Phi_{X,t}\boldsymbol{v})(x) = \sum_{i\in\mathcal{I}_t} v_i\kappa_X(x_i, x)$, where $\boldsymbol{v} \in \mathbb{R}^{d_t}$ is a right eigenvector of a finite-dimensional matrix $W_t^\top G_{YX,t}$ with the same eigenvalue.

Leading eigenfunctions of the Koopman operator can characterize regions of attraction (e.g., see Hou et al. (2023b)). Consider the unforced Duffing oscillator, described by $\ddot{z} = -\delta\dot{z} - z\left(\beta + \alpha z^2\right)$, with $\delta = 0.5$, $\beta = -1$, and $\alpha = 1$, where $z \in \mathbb{R}$ and $\dot{z} \in \mathbb{R}$ are the scalar position and velocity, respectively. Let $x = (z, \dot{z})$. As Figure 2(*a*) reveals, the Duffing dynamics exhibits two regions of attraction, corresponding to equilibrium points $x = (-1, 0)$ and $x = (1, 0)$. Although Assumption 3 does not hold in this case, our online algorithm serves as an efficient computational technique to analyze dynamical system properties.



(*a*) (*b*) $\varepsilon = 0$ (*c*) $\varepsilon = 2\eta^3$ (*d*) $\varepsilon = 1.5\eta^2$

Figure 2: (a) Two trajectories of the Duffing oscillator that converge to two different equilibrium points. (b)-(d) Leading eigenfunction of $\mathcal{K}$ with eigenvalue 1 at $t = 3550$ under various compression budget with (b)$\varepsilon = 0$,$|\mathcal{D}_t| = 3550$, (c) $\varepsilon = 2\eta^3$,$|\mathcal{D}_t| = 300$, and (d) $\varepsilon = 1.5\eta^2$,$|\mathcal{D}_t| = 190$.

To approximate $\mathcal{K}$, the streaming data consists of samples from 355 trajectories over region $[-2, 2] \times [-2, 2]$. We used a sampling interval of 0.25s and did not require the sampled trajectories to be fully mixed. We utilized a Gaussian kernel $\kappa(x_1, x_2) = \exp(-\|x_1 - x_2\|_2^2/(2 \times 0.3^2))$ and implemented Algorithm 1 with a constant step-size $\eta = 0.2$ and processed 3550 steaming sample pairs. Figure 2(*b*)-2(*d*) portrays heat maps of the leading eigenfunctions of $\mathcal{K}$ after 3550 iterations with various values of compression budget ε. Upon increasing ε, the dictionary becomes more compressed with fewer elements. As shown in Figure 2(*c*), the resulting eigenfunctions accurately reveal the distinct regions of attractions, even with merely 10% of total data points. However, the characterization becomes less sound with higher ε as the algorithm discards too many points.

6. Conclusions

In this paper, we have presented an algorithm that learns a compressed version of the CME incrementally. Then, we studied the convergence behavior of the last literate under Markovian sampling. We have applied this framework to analyzing unknown nonlinear dynamical systems. Numerical examples confirmed the effectiveness of the proposed mechanism.

From the standpoint of pure theoretical analysis, we want to extend our last iterate convergence in the mean square sense to almost sure guarantees and those with high probability, and investigate asymptotic behavior. We also want to study if the multiple variants of scalar-valued stochastic gradient descent can be extended to the operator setting. In addition, we will investigate extending a variety of compression schemes for function learning in RKHS to the operator setting, e.g., windowed retention of old points Kivinen et al. (2004), approximate linear dependence Engel et al. (2004), random Fourier features Rahimi and Recht (2007), Nyström method Rudi et al. (2015), and coherence-based sparsification Richard et al. (2008). In terms of conceptual directions for future work, perhaps our main interest lies in the generalization of the online compressed CME learning framework to the question of reinforcement learning through approaches mirroring those in Grunewalder et al. (2012). Another direction of interest is the online fusion of compressed CME estimates, where multiple agents maintains and communicates separate compressed estimates over a network.

References

Jean-Pierre Aubin. *Applied functional analysis*. John Wiley & Sons, 2011.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

Zaiwei Chen, Sheng Zhang, Thinh T Doan, John-Paul Clarke, and Siva Theja Maguluri. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, 146:110623, 2022.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. *Advances in neural information processing systems*, 29, 2016.

Nathan Dahlin, Subhonmesh Bose, and Venugopal V Veeravalli. Controlling a markov decision process with an abrupt change in the transition kernel. In *2023 American Control Conference (ACC)*, pages 3401–3408. IEEE, 2023.

Yaakov Engel, Shie Mannor, and Ron Meir. The kernel recursive least-squares algorithm. *IEEE Transactions on signal processing*, 52(8):2275–2285, 2004.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *The Journal of Machine Learning Research*, 21(1):8464–8501, 2020.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan): 73–99, 2004.

Graham C Goodwin and Kwai Sang Sin. *Adaptive filtering prediction and control*. Courier Corporation, 2014.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional mean embeddings as regressors-supplementary. *International Conference on Machine Learning*, 2012.

Steffen Grunewalder, Guy Lever, Luca Baldassarre, Massi Pontil, and Arthur Gretton. Modelling transition dynamics in MDPs with RKHS embeddings. *International Conference on Machine Learning*, 2012.

Elad Hazan and Comandur Seshadhri. Efficient learning algorithms for changing environments. In *Proceedings of the 26th annual international conference on machine learning*, pages 393–400, 2009.

Boya Hou, Subhonmesh Bose, and Umesh Vaidya. Sparse learning of kernel transfer operators. In *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pages 130–134. IEEE, 2021.

Boya Hou, Sina Sanjari, Nathan Dahlin, and Subhonmesh Bose. Compressed decentralized learning of conditional mean embedding operators in reproducing kernel hilbert spaces. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023a.

Boya Hou, Sina Sanjari, Nathan Dahlin, Subhonmesh Bose, and Umesh Vaidya. Sparse learning of dynamical systems in RKHS: An operator-theoretic approach. In *International Conference on Machine Learning*, pages 13325–13352. PMLR, 2023b.

Jyrki Kivinen, Alexander J Smola, and Robert C Williamson. Online learning with kernels. *IEEE transactions on signal processing*, 52(8):2165–2176, 2004.

Ilja Klebanov, Ingmar Schuster, and Timothy John Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020.

Stefan Klus, Ingmar Schuster, and Krikamol Muandet. Eigendecompositions of transfer operators in reproducing kernel hilbert spaces. *Journal of Nonlinear Science*, 30(1):283–315, 2020.

Alec Koppel, Garrett Warnell, Ethan Stump, and Alejandro Ribeiro. Parsimonious online learning with kernels via sparse projections in function space. *The Journal of Machine Learning Research*, 20(1):83–126, 2019.

Vladimir Kostic, Pietro Novelli, Andreas Maurer, Carlo Ciliberto, Lorenzo Rosasco, and Massimiliano Pontil. Learning dynamical systems via koopman operator regression in reproducing kernel hilbert spaces. *Advances in Neural Information Processing Systems*, 35:4017–4031, 2022.

Guy Lever, John Shawe-Taylor, Ronnie Stafford, and Csaba Szepesvári. Compressed conditional mean embeddings for model-based reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning. *Advances in Neural Information Processing Systems*, 35: 4433–4445, 2022.

David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.

Jovana Mitrovic, Dino Sejdinovic, and Yee Whye Teh. Causal inference via kernel deviance measures. *Advances in neural information processing systems*, 31, 2018.

Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, and Sanparith Marukatat. Counterfactual mean embeddings. *The Journal of Machine Learning Research*, 22(1):7322–7392, 2021.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in Neural Information Processing Systems*, 33:21247–21259, 2020.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Cédric Richard, José Carlos M Bermudez, and Paul Honeine. Online prediction of time series data with kernels. *IEEE Transactions on Signal Processing*, 57(3):1058–1067, 2008.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 28, 2015.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35:363–417, 2012.

Kunio Takezawa. *Introduction to nonparametric regression*. John Wiley & Sons, 2005.

Prem Talwai, Ali Shameli, and David Simchi-Levi. Sobolev norm learning rates for conditional mean embeddings. In *International conference on artificial intelligence and statistics*, pages 10422–10447. PMLR, 2022.

Max A Woodbury. *Inverting modified matrices*. Department of Statistics, Princeton University, 1950.

Appendix A provides a brief introduction to the tensor product Hilbert spaces and Hilbert-Schmidt operators. Appendix B contains the supporting materials regarding operator-valued gradients introduced in Section 3. Appendix C includes results that will be useful in handling the hard learning scenario. Appendix G presents a pseudo-code of our main algorithm. This is then followed by Appendix F that provides detailed proof of theorems regarding the convergence results. In the rest of the appendix, we present a detailed description of implementing the algorithm.

Appendix A. A Primer on Tensor Product Hilbert Space and Hilbert-Schmidt Operators

This appendix recalls some facts regarding tensor product Hilbert spaces and Hilbert-Schmidt operators. We refer the reader to (Aubin, 2011, Chapter 12) for details.

Consider two separable real-valued Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ defined on separable measurable spaces $\mathbb{X}$ and $\mathbb{Y}$, respectively. Let $\{e_i\}_{i\in\mathbb{N}}$ be an orthonormal basis (ONB) of $\mathcal{H}_1$. A bounded linear operator $A : \mathcal{H}_1 \to \mathcal{H}_2$ is a Hilbert-Schmidt (HS) operator if $\sum_{i\in\mathbb{N}} \|Ae_i\|^2_{\mathcal{H}_2} < \infty$. The quantity $\|A\|_{\text{HS}} = \left(\sum_{i\in\mathbb{N}} \|Ae_i\|^2_{\mathcal{H}_2}\right)^{1/2}$ is the Hilbert-Schmidt norm of A and is independent of the choice of the ONB. For two HS operators A and B from $\mathcal{H}_1$ to $\mathcal{H}_2$, their Hilbert–Schmidt inner product can be defined as

$$\langle A, B\rangle_{\text{HS}(\mathcal{H}_1,\mathcal{H}_2)} = \text{Tr}(A^*B) = \sum_{i\in\mathbb{N}} \langle Ae_i, Be_i\rangle_{\mathcal{H}_2}. \tag{36}$$

In addition, for a Hilbert-Schmidt operator A and a bounded linear operator B, we have

$$\|A\|_{\text{HS}} = \text{Tr}(A^*A)^{1/2}, \tag{37}$$

$$\|A\|_{\text{HS}} = \|A^*\|_{\text{HS}}, \tag{38}$$

$$\|A\|_{\text{op}} \le \|A\|_{\text{HS}}, \tag{39}$$

$$\|BA\|_{\text{HS}} \le \|B\|_{\text{op}} \|A\|_{\text{HS}}, \quad \|AB\|_{\text{HS}} \le \|A\|_{\text{HS}} \|B\|_{\text{op}}, \tag{40}$$

where A^* is the adjoint of A and $\|A\|_{\text{op}}$ is the operator norm of A. Let $f \in \mathcal{H}_1, g \in \mathcal{H}_2$, the tensor product $f \otimes g : \mathcal{H}_2 \to \mathcal{H}_1$ can be viewed as the linear rank-one operator defined by $(f \otimes g)h = \langle h, g\rangle_{\mathcal{H}_2} f$ for all $h \in \mathcal{H}_2$. Thus, for any bounded linear operator A from $\mathcal{H}_1$ to itself, we have

$$A\left((f \otimes g)\, h\right) = A\left(\langle h, g\rangle_{\mathcal{H}_2} f\right) = \langle h, g\rangle_{\mathcal{H}_2} (Af) = \left((Af) \otimes g\right) h, \quad f \in \mathcal{H}_1, h \in \mathcal{H}_2. \tag{41}$$

That is,

$$A\,(f \otimes g) = (Af) \otimes g. \tag{42}$$

Furthermore, if $\{e_i\}_{i\in\mathbb{N}}$ is an orthonormal systems (ONS) of $\mathcal{H}_1$ and $\{e'_j\}_{j\in\mathbb{N}}$ is an ONS of $\mathcal{H}_2$, then $\{e_i \otimes e'_j\}_{i,j\in\mathbb{N}}$ is an ONS of $\mathcal{H}_1 \otimes \mathcal{H}_2$.

Now consider $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$ and A is an HS operator mapping from $\mathcal{H}_2$ to $\mathcal{H}_1$. Let $(e_i)_{i\in\mathbb{N}}$ be an orthonormal basis of $\mathcal{H}_2$. Then we have the Fourier series expansion of $g \in \mathcal{H}_2$ as

$$g = \sum_{i\in\mathbb{N}} \langle g, e_i\rangle_{\mathcal{H}_2} e_i. \tag{43}$$

Therefore, using (36), we have

$$
\begin{aligned}
\langle f \otimes g, A\rangle_{\mathrm{HS}} &= \sum_{i\in\mathbb{N}} \langle (f\otimes g)\, e_i, Ae_i\rangle_{\mathcal{H}_1} \\
&= \sum_{i\in\mathbb{N}} \left\langle \langle g, e_i\rangle_{\mathcal{H}_2} f, Ae_i \right\rangle_{\mathcal{H}_1} \\
&= \sum_{i\in\mathbb{N}} \langle g, e_i\rangle_{\mathcal{H}_2} \langle f, Ae_i\rangle_{\mathcal{H}_1} \\
&= \sum_{i\in\mathbb{N}} \langle g, e_i\rangle_{\mathcal{H}_2} \langle A^* f, e_i\rangle_{\mathcal{H}_2} \\
&= \left\langle \left\{ \langle g, e_i\rangle_{\mathcal{H}_2} \right\}_{i\in\mathbb{N}}, \left\{ \langle A^* f, e_i\rangle_{\mathcal{H}_2} \right\}_{i\in\mathbb{N}} \right\rangle_{l_2(\mathbb{N})}.
\end{aligned}
\tag{44}
$$

Since a separable Hilbert space is isomorphic to $l_2(\mathbb{N})$ (Aubin, 2011, Theorem 1.7.2), let $T(g) = \left\{ \langle g, e_i\rangle_{\mathcal{H}_2} \right\}_{i\in\mathbb{N}}$ denote such an isomorphism $T : \mathcal{H}_2 \mapsto l_2(\mathbb{N})$ then we have

$$
\langle f \otimes g, A\rangle_{\mathrm{HS}} = \langle T(g), T(A^* f)\rangle_{l_2(\mathbb{N})} = \langle g, A^* f\rangle_{\mathcal{H}_2} = \langle f, Ag\rangle_{\mathcal{H}_1}. \tag{45}
$$

A.1. Proof of Lemma 1

For the first claim, (Ciliberto et al., 2016, Lemma 15) and (Li et al., 2022, Theorem 1) show that there exists an isometric isomorphism $\iota_\kappa : \mathcal{H}_Y \otimes \mathcal{H}_X \to \mathcal{H}_V$. Moreover, Li et al. (2022) further establishes that for each $v \in \mathcal{H}_V$, there exists a unique $V \in \mathcal{H}_Y \otimes \mathcal{H}_X$ given by $V = \iota_\kappa^{-1}(v)$ such that $\|v\|_{\mathcal{H}_V} = \|V\|_{\mathrm{HS}}$ and the *operator reproducing property* holds, i.e.,

$$
v(x) = V\phi_X(x) \in \mathcal{H}_Y, \quad \forall x \in \mathbb{X}. \tag{46}
$$

For the second claim, $\mathcal{H}_Y \otimes [\mathcal{H}_X] \subseteq \mathcal{H}_Y \otimes L_2(\rho_X)$. In addition, according to (Aubin, 2011, Theorem 12.6.1), there exists an isomorphism $\iota : L_2(\rho_X, \mathcal{H}_Y) \to \mathcal{H}_Y \otimes L_2(\rho_X)$ [1].

Appendix B. Operator-valued Gradient and Its Properties

B.1. Proof of Lemma 2

Consider $R_\lambda : \mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y) \to \mathbb{R}$ in (7). For any $h \in \mathbb{R}$ and $U, U' \in \mathrm{HS}(\mathcal{H}_X, \mathcal{H}_Y)$, we have

$$
R_\lambda(U + hU') = R_\lambda(U) + h\, \langle \nabla R_\lambda(U), U'\rangle_{\mathrm{HS}} + \mathcal{O}(h^2). \tag{47}
$$

1. The original statement of (Aubin, 2011, Theorem 12.6.1) only claims isometry. Yet their proof shows that there exists a linear mapping from $L_2(\rho_X, \mathcal{H})$ to $\mathcal{H} \otimes L_2(\rho_X)$ that is isometry and surjective.

Hence,

$$
\begin{aligned}
\langle \nabla R_\lambda(U), U' \rangle_{\text{HS}} &= \lim_{h\to 0} \frac{R_\lambda(U+hU') - R_\lambda(U)}{h} \\
&= \underbrace{\lim_{h\to 0} \frac{1}{2h} \mathbb{E}\left[\| \phi_Y(y) - U\phi_X(x) - hU'\phi_X(x) \|^2_{\mathcal{H}_Y} - \| \phi_Y(y) - U\phi_X(x) \|^2_{\mathcal{H}_Y} \right]}_{T_1} \\
&+ \underbrace{\lim_{h\to 0} \frac{\lambda \| U + hU' \|^2_{\text{HS}} - \lambda \| U \|^2_{\text{HS}}}{2h}}_{T_2}.
\end{aligned}
\tag{48}
$$

To compute T_1 , first notice that

$$
\begin{aligned}
&\frac{1}{2h} \mathbb{E}\left[\| \phi_Y(y) - U\phi_X(x) - hU'\phi_X(x) \|^2_{\mathcal{H}_Y} - \| \phi_Y(y) - U\phi_X(x) \|^2_{\mathcal{H}_Y} \right] \\
=&\mathbb{E}\left[\frac{-2h \langle \phi_Y(y) - U\phi_X(x), U'\phi_X(x) \rangle_{\mathcal{H}_Y} + \| hU'\phi_X(x) \|^2_{\mathcal{H}_Y}}{2h} \right] \\
=&\mathbb{E}\left[- \langle \phi_Y(y) - U\phi_X(x), U'\phi_X(x) \rangle_{\mathcal{H}_Y} + \frac{h}{2} \| U'\phi_X(x) \|^2_{\mathcal{H}_Y} \right].
\end{aligned}
\tag{49}
$$

By Assumption 1, the kernel function is bounded. Since U, U' are Hilbert-Schmidt from $\mathcal{H}_X$ to $\mathcal{H}_Y$, we can apply the dominated convergence theorem to obtain

$$
\begin{aligned}
T_1 = &\lim_{h\to 0} \mathbb{E}\left[- \langle \phi_Y(y) - U\phi_X(x), U'\phi_X(x) \rangle_{\mathcal{H}_Y} + \frac{h}{2} \| U'\phi_X(x) \|^2_{\mathcal{H}_Y} \right] \\
=&\mathbb{E}\left[- \langle \phi_Y(y) - U\phi_X(x), U'\phi_X(x) \rangle_{\mathcal{H}_Y} \right].
\end{aligned}
\tag{50}
$$

From (45), the last line above can be written as

$$
\begin{aligned}
T_1 &= - \mathbb{E}\left[\langle (\phi_Y(y) - U\phi_X(x)) \otimes (\phi_X(x)), U' \rangle_{\text{HS}} \right] \\
&= - \langle \mathbb{E}\left[(\phi_Y(y) - U\phi_X(x)) \otimes (\phi_X(x)) \right], U' \rangle_{\text{HS}},
\end{aligned}
\tag{51}
$$

Likewise for T_2, we have

$$
T_2 = \frac{\lambda}{2} \lim_{h\to 0} \frac{\| U + hU' \|^2_{\text{HS}} - \| U \|^2_{\text{HS}}}{h} = \lambda \langle U, U' \rangle_{\text{HS}}. \tag{52}
$$

Putting together, (48) becomes

$$
\langle \nabla R_\lambda(U), U' \rangle_{\text{HS}} = \langle -\mathbb{E}\left[(\phi_Y(y) - U\phi_X(x)) \otimes (\phi_X(x)) \right] + \lambda U, U' \rangle_{\text{HS}}. \tag{53}
$$

Hence, the operator gradient of $R_\lambda(U)$ is given by

$$
\begin{aligned}
\nabla R_\lambda(U) = &- \mathbb{E}\left[(\phi_Y(y) - U\phi_X(x)) \otimes (\phi_X(x)) \right] + \lambda U \\
=&UC_{XX} - C_{YX} + \lambda U.
\end{aligned}
\tag{54}
$$

In addition, since C_{XX},C_{YX} is Hilbert Schmidt when the kernel function is uniformly bounded Fukumizu et al. (2004), we get that $\nabla R_\lambda(U) \in \text{HS}(\mathcal{H}_X, \mathcal{H}_Y)$, which completes the proof.

B.2. Properties of Operator-valued Gradients

Consider $(x, y) \in \mathbb{X} \times \mathbb{Y}$ and define

$$\widetilde{C}_{XX}(x) = \phi_X(x) \otimes \phi_X(x),$$
$$\widetilde{C}_{YX}(y, x) = \phi_Y(y) \otimes \phi_X(x)$$

Under Assumption 1(i), we have

$$\begin{aligned}\left\|\widetilde{C}_{XX}(x)\right\|_{\text{HS}}^2 &= \langle \phi_X(x) \otimes \phi_X(x), \phi_X(x) \otimes \phi_X(x) \rangle_{\text{HS}} = \kappa_X(x, x)\kappa_X(x, x) \leq K^2 \\ \left\|\widetilde{C}_{YX}(y, x)\right\|_{\text{HS}}^2 &= \langle \phi_Y(y) \otimes \phi_X(x), \phi_Y(y) \otimes \phi_X(x) \rangle_{\text{HS}} = \kappa_X(x, x)\kappa_Y(y, y) \leq K^2.\end{aligned} \tag{55}$$

Let $B_\kappa = K + \lambda$. Then,

$$\begin{aligned}&\max_{x \in \mathbb{X}} \left\|\widetilde{C}_{XX}(x) + \lambda \text{Id}\right\|_{\text{op}} \leq \max_{x \in \mathbb{X}} \left(\left\|\widetilde{C}_{XX}(x)\right\|_{\text{op}} + \lambda \|\text{Id}\|_{\text{op}}\right) \leq \max_{x \in \mathbb{X}} \left(\left\|\widetilde{C}_{XX}(x)\right\|_{\text{HS}}\right) + \lambda \leq B_\kappa \\ &\max_{x \in \mathbb{X}, y \in \mathbb{Y}} \left\|\widetilde{C}_{YX}(y, x)\right\|_{\text{HS}} \leq K \leq B_\kappa.\end{aligned} \tag{56}$$

We have the following properties regarding $\nabla R_\lambda(U)$,$\widetilde{\nabla} R_\lambda(x, y, U)$ which are needed for the convergence analysis.

Lemma 8 *(Properties of gradients) Under Assumptions 1 and 2, $\nabla R_\lambda(U)$ and its stochastic approximation $\widetilde{\nabla} R_\lambda(x, y, U)$ satisfy the following.*

(a) (Strong convexity) For $U \in HS(\mathcal{H}_X, \mathcal{H}_Y)$,

$$\langle -\nabla R_\lambda(U), U - U_\lambda \rangle_{HS} \leq -\lambda \|U - U_\lambda\|_{HS}^2,$$

where $U_\lambda := \operatorname{argmin}_{U \in HS(\mathcal{H}_X, \mathcal{H}_Y)} R_\lambda(U)$.

(b) (Lipschitz gradient) $\nabla R_\lambda(U)$ and $\widetilde{\nabla} R_\lambda(x, y, U)$ are Lipschitz continuous with respect to U for all $x \in \mathbb{X}, y \in \mathbb{Y}$, i.e.,

$$\begin{aligned}\|\nabla R_\lambda(U_1) - \nabla R_\lambda(U_2)\|_{HS} &\leq B_\kappa \|U_1 - U_2\|_{HS} \\ \left\|\widetilde{\nabla} R_\lambda(x, y, U_1) - \widetilde{\nabla} R_\lambda(x, y, U_2)\right\|_{HS} &\leq B_\kappa \|U_1 - U_2\|_{HS}\end{aligned}$$

for all $U_1, U_2 \in HS(\mathcal{H}_X, \mathcal{H}_Y)$ and $x \in \mathbb{X}, y \in \mathbb{Y}$.

(c) (Affine scaling) For $U \in HS(\mathcal{H}_X, \mathcal{H}_Y)$,

$$\begin{aligned}\|\nabla R_\lambda(U)\|_{HS} &\leq B_\kappa (\|U\|_{HS} + 1) \\ \left\|\widetilde{\nabla} R_\lambda(x, y, U)\right\|_{HS} &\leq B_\kappa (\|U\|_{HS} + 1)\end{aligned}$$

for all $x \in \mathbb{X}, y \in \mathbb{Y}$. As a result, $\nabla R_\lambda(U)$ and $\widetilde{\nabla} R_\lambda(x, y, U)$ are Bochner-integrable.

Proof Denote $g(U) := R_\lambda(U) - \frac{\lambda}{2}\left\|U\right\|_{\text{HS}}^2$. Then for $U_1, U_2 \in \text{HS}(\mathcal{H}_X, \mathcal{H}_Y)$ and $\alpha \in (0,1]$,

$$\begin{aligned} g\left(\alpha U_1 + (1-\alpha)\,U_2\right) =& \frac{1}{2}\mathbb{E}\left[\left\|\phi_Y(y) - \left(\alpha U_1 + (1-\alpha)\,U_2\right)\phi_X(x)\right\|_{\mathcal{H}_Y}^2\right] \\ =& \frac{1}{2}\mathbb{E}\left[\left\|\left(\alpha + (1-\alpha)\right)\phi_Y(y) - \left(\alpha U_1 + (1-\alpha)\,U_2\right)\phi_X(x)\right\|_{\mathcal{H}_Y}^2\right] \\ =& \frac{1}{2}\mathbb{E}\left[\left\|\alpha\left(\phi_Y(y) - U_1\phi_X(x)\right) + (1-\alpha)\left(\phi_Y(y) - U_2\phi_X(x)\right)\right\|_{\mathcal{H}_Y}^2\right]. \end{aligned} \tag{57}$$

Let $T_1 := \phi_Y(y) - U_1\phi_X(x)$ and $T_2 := \phi_Y(y) - U_2\phi_X(x)$. Then, we have

$$\begin{aligned} g\left(\alpha U_1 + (1-\alpha)\,U_2\right) &= \frac{1}{2}\mathbb{E}\left\|\alpha T_1 + (1-\alpha)\,T_2\right\|_{\mathcal{H}_Y}^2 \\ &= \frac{1}{2}\mathbb{E}\left[\alpha^2\left\|T_1\right\|_{\mathcal{H}_Y}^2 + (1-\alpha)^2\left\|T_2\right\|_{\mathcal{H}_Y}^2 + 2\alpha\,(1-\alpha)\left\langle T_1, T_2\right\rangle_{\mathcal{H}_Y}\right]. \end{aligned} \tag{58}$$

Hence, for $\alpha \in (0,1]$, we have

$$\begin{aligned} & \alpha g\left(U_1\right) + (1-\alpha)\,g\left(U_2\right) - g\left(\alpha U_1 + (1-\alpha)\,U_2\right) \\ =& \frac{\alpha}{2}\mathbb{E}\left\|T_1\right\|_{\mathcal{H}_Y}^2 + \frac{1-\alpha}{2}\mathbb{E}\left\|T_2\right\|_{\mathcal{H}_Y}^2 - \frac{1}{2}\mathbb{E}\left\|\alpha T_1 + (1-\alpha)\,T_2\right\|_{\mathcal{H}_Y}^2 \\ =& \frac{1}{2}\mathbb{E}\left[\alpha\,(1-\alpha)\left\|T_1\right\|_{\mathcal{H}_Y}^2 + \alpha\,(1-\alpha)\left\|T_2\right\|_{\mathcal{H}_Y}^2 - 2\alpha\,(1-\alpha)\left\langle T_1, T_2\right\rangle_{\mathcal{H}_Y}\right] \\ =& \frac{\alpha\,(1-\alpha)}{2}\mathbb{E}\left[\left\|T_1\right\|_{\mathcal{H}_Y}^2 + \left\|T_2\right\|_{\mathcal{H}_Y}^2 - 2\left\langle T_1, T_2\right\rangle_{\mathcal{H}_Y}\right] \\ =& \frac{\alpha\,(1-\alpha)}{2}\mathbb{E}\left[\left\|T_1 - T_2\right\|_{\mathcal{H}_Y}^2\right] \\ \geq& 0. \end{aligned} \tag{59}$$

That is, $g : \text{HS}(\mathcal{H}_X, \mathcal{H}_Y) \to \mathbb{R}$ satisfies

$$g\left(\alpha U_1 + (1-\alpha)\,U_2\right) \leq \alpha g\left(U_1\right) + (1-\alpha)\,g\left(U_2\right), \tag{60}$$

and thus is a convex functional in the sense of (Luenberger, 1997, p. 190). Rearranging terms in the above equation, we have

$$g\left(U_1\right) - g\left(U_2\right) \geq \frac{g\left(U_2 + \alpha\left(U_1 - U_2\right)\right) - g\left(U_2\right)}{\alpha}, \quad \alpha \in (0,1]. \tag{61}$$

Taking $\alpha \to 0$, we have

$$\begin{aligned} g\left(U_1\right) - g\left(U_2\right) &\geq \lim_{\alpha \to 0}\frac{g\left(U_2 + \alpha\left(U_1 - U_2\right)\right) - g\left(U_2\right)}{\alpha} \\ &= \left\langle \nabla g\left(U_2\right), U_1 - U_2\right\rangle_{\text{HS}}, \end{aligned} \tag{62}$$

where the last line holds due to (48) by replacing $R_\lambda(\cdot)$ with $g(\cdot)$. Now, substituting $g(\cdot)$ with $R_\lambda(\cdot) - \frac{\lambda}{2}\left\|\cdot\right\|_{\text{HS}}^2$, we have for $U_1, U_2 \in \text{HS}(\mathcal{H}_X, \mathcal{H}_Y)$,

$$\left[R_\lambda\left(U_1\right) - \frac{\lambda}{2}\left\|U_1\right\|_{\text{HS}}^2\right] - \left[R_\lambda\left(U_2\right) - \frac{\lambda}{2}\left\|U_2\right\|_{\text{HS}}^2\right] \geq \left\langle \nabla R_\lambda\left(U_2\right) - \lambda U_2, U_1 - U_2\right\rangle_{\text{HS}}. \tag{63}$$

Rearranging terms gives

$$\begin{aligned} R_\lambda(U_1) - R_\lambda(U_2) &\geq \langle \nabla R_\lambda(U_2), U_1 - U_2\rangle_{\text{HS}} - \lambda \langle U_2, U_1 - U_2\rangle_{\text{HS}} + \frac{\lambda}{2}\|U_1\|_{\text{HS}}^2 - \frac{\lambda}{2}\|U_2\|_{\text{HS}}^2 \\ &= \langle \nabla R_\lambda(U_2), U_1 - U_2\rangle_{\text{HS}} - \lambda \langle U_2, U_1\rangle_{\text{HS}} + \frac{\lambda}{2}\|U_1\|_{\text{HS}}^2 + \frac{\lambda}{2}\|U_2\|_{\text{HS}}^2 \\ &= \langle \nabla R_\lambda(U_2), U_1 - U_2\rangle_{\text{HS}} + \frac{\lambda}{2}\|U_1 - U_2\|_{\text{HS}}^2 . \end{aligned} \tag{64}$$

In other words, R_λ is λ-strongly convex. From (64), we have for $U_1, U_2 \in \text{HS}(\mathcal{H}_X, \mathcal{H}_Y)$,

$$\begin{aligned} R_\lambda(U_1) - R_\lambda(U_2) \geq \langle \nabla R_\lambda(U_2), U_1 - U_2\rangle_{\text{HS}} + \frac{\lambda}{2}\|U_1 - U_2\|_{\text{HS}}^2 , \\ R_\lambda(U_2) - R_\lambda(U_1) \geq \langle \nabla R_\lambda(U_1), U_2 - U_1\rangle_{\text{HS}} + \frac{\lambda}{2}\|U_1 - U_2\|_{\text{HS}}^2 . \end{aligned} \tag{65}$$

Adding these two we have

$$0 \geq \langle \nabla R_\lambda(U_2) - \nabla R_\lambda(U_1), U_1 - U_2\rangle_{\text{HS}} + \lambda \|U_1 - U_2\|_{\text{HS}}^2 . \tag{66}$$

Let $U_1 = U, U_2 = U_\lambda$. Since $\nabla R_\lambda(U_\lambda) = 0$, we have

$$0 \geq \langle -\nabla R_\lambda(U_1), U_1 - U_\lambda\rangle_{\text{HS}} + \lambda \|U_1 - U_2\|_{\text{HS}}^2 . \tag{67}$$

Rearranging terms concludes the proof of claim (a).

To prove the second and third claims, we have

$$\begin{aligned} &\left\|\widetilde{\nabla} R_\lambda(x, y, U_1) - \widetilde{\nabla} R_\lambda(x, y, U_2)\right\|_{\text{HS}} \\ =& \left\|\left[U_1 \widetilde{C}_{XX}(x) - \widetilde{C}_{YX}(y, x) + \lambda U_1\right] - \left[U_2 \widetilde{C}_{XX}(x) - \widetilde{C}_{YX}(y, x) + \lambda U_2\right]\right\|_{\text{HS}} \\ =& \left\|(U_1 - U_2)\left(\widetilde{C}_{XX}(x) + \lambda \text{Id}\right)\right\|_{\text{HS}} . \end{aligned} \tag{68}$$

Using the fact that $\|AB\|_{\text{HS}} \leq \|A\|_{\text{HS}} \|B\|_{\text{op}}$ for Hilbert-Schmidt operator A and bounded linear operator B, we further have

$$\begin{aligned} \left\|\widetilde{\nabla} R_\lambda(x, y, U_1) - \widetilde{\nabla} R_\lambda(x, y, U_2)\right\|_{\text{HS}} &\leq \|U_1 - U_2\|_{\text{HS}} \left\|\widetilde{C}_{XX}(x) + \lambda \text{Id}\right\|_{\text{op}} \\ &\leq B_\kappa \|U_1 - U_2\|_{\text{HS}} . \end{aligned} \tag{69}$$

Together with triangle inequality w.r.t HS-norm, we then have

$$\begin{aligned} \left\|\widetilde{\nabla} R_\lambda(x, y, U)\right\|_{\text{HS}} &\leq \left\|\widetilde{\nabla} R_\lambda(x, y, U) - \widetilde{\nabla} R_\lambda(x, y, 0)\right\|_{\text{HS}} + \left\|\widetilde{\nabla} R_\lambda(x, y, 0)\right\|_{\text{HS}} \\ &\leq B_\kappa \|U\|_{\text{HS}} + \left\|\widetilde{C}_{YX}(y, x)\right\|_{\text{HS}} \\ &\leq B_\kappa (\|U\|_{\text{HS}} + 1) . \end{aligned} \tag{70}$$

Since U is Hilbert-Schmdit, the above inequality also implies that $\int_{\mathbb{X}\times\mathbb{Y}} \left\|\widetilde{\nabla} R_\lambda(x,y,U)\right\|_{\text{HS}} \mathrm{d}\rho(x,y) \leq B_\kappa \left(\|U\|_{\text{HS}}+1\right) < \infty$, i.e., $\widetilde{\nabla} R_\lambda(x,y,U)$ is Bochner-integrable. Using Jensen's inequality, we have

$$\begin{aligned}\|\nabla R_\lambda(U)\|_{\text{HS}} &= \left\|\mathbb{E}\left[\widetilde{\nabla} R_\lambda(x,y,U)\right]\right\|_{\text{HS}} \\ &\leq \mathbb{E}\left[\left\|\widetilde{\nabla} R_\lambda(x,y,U)\right\|_{\text{HS}}\right] \\ &\leq B_\kappa\left(\|U\|_{\text{HS}}+1\right).\end{aligned} \tag{71}$$

Similarly, using (69), we get

$$\begin{aligned}\|\nabla R_\lambda(U_1)-\nabla R_\lambda(U_2)\|_{\text{HS}} =& \left\|\mathbb{E}\left[\widetilde{\nabla} R_\lambda(x,y,U_1)-\widetilde{\nabla} R_\lambda(x,y,U_2)\right]\right\|_{\text{HS}} \\ \leq& \mathbb{E}\left[\left\|\widetilde{\nabla} R_\lambda(x,y,U_1)-\widetilde{\nabla} R_\lambda(x,y,U_2)\right\|_{\text{HS}}\right] \\ \leq& B_\kappa\|U_1-U_2\|_{\text{HS}}.\end{aligned} \tag{72}$$

■

Appendix C. Technical Lemmas Related to Learning in the Power Spaces

By spectral theorem, the integral operator L_κ defined in (20) enjoys the spectral representation (21) which is convergent in $L_2(\rho_X)$. Since L_κ is a strictly positive operator, then for any $r \in [0,\infty)$, the fractional power $L_\kappa^r : L_2(\rho_X) \to L_2(\rho_X)$ is defined by

$$L_\kappa^r[f] := \sum_{i\in\mathbb{I}} \sigma_i^r \langle [f],[e_i]\rangle_\rho [e_i], \quad [f] \in L_2(\rho_X). \tag{73}$$

Moreover, define the adjoint of I_κ by $I_\kappa{}^* : L_2(\rho_X) \to \mathcal{H}$. From the adjoint relation, we have for any $f \in L_2(\rho_X)$ and $h \in \mathcal{H}$

$$\langle I_\kappa{}^*[f], h\rangle_{\mathcal{H}} = \langle [f], I_\kappa h\rangle_\rho = \int_{\mathbb{X}} g(x)\, h(x)\mathrm{d}\rho_X(x) \quad \forall g \in [f]. \tag{74}$$

Taking $h = \kappa(\cdot,x) =: \kappa_x$ for $x \in X$ yields that

$$I_\kappa{}^*[f] = \int_{\mathbb{X}} \kappa(\cdot,x)g(x)\mathrm{d}\rho_X \quad \forall g \in [f]. \tag{75}$$

In addition, since $(\kappa_x \otimes \kappa_x)\, h = \langle h, \kappa_x\rangle_{\mathcal{H}} \kappa_x$, for all $h \in \mathcal{H}$, we have

$$C_\kappa h = (I_\kappa{}^* I_\kappa)h = I_\kappa{}^*(I_\kappa h) = \int_{\mathbb{X}} \kappa_x h(x)\mathrm{d}\rho_X(x) = \int_{\mathbb{X}} \kappa_x \langle h,\kappa_x\rangle_{\mathcal{H}} \mathrm{d}\rho_X(x) = \int_{\mathbb{X}} (\kappa_x\otimes\kappa_x)\, h\mathrm{d}\rho_X(x). \tag{76}$$

That is, the self-adjoint integral operator $C_\kappa := I_\kappa{}^* I_\kappa$ is the covariance operator C_{XX} defined in Section 2.3 with respect to the probability measure ρ_X.

Recall that $\left(\sigma_i^{1/2} e_i\right)_{i\in\mathbb{I}}$ is an ONB of $(\ker I_\kappa)^\perp$, $([e_i])_{i\in\mathbb{I}}$ an ONB of $\overline{\text{range} I_\kappa}$, and we have the spectral representation of C_κ with respect to the ONS $(\sqrt{\sigma_i} e_i)_{i\in\mathbb{I}}$ in $\mathcal{H}$.

$$C_\kappa = \sum_{i\in\mathbb{I}} \sigma_i \left\langle \cdot, \sqrt{\sigma_i} e_i \right\rangle_{\mathcal{H}} \sqrt{\sigma_i} e_i, \quad \mathcal{H} = \ker C_\kappa \oplus \overline{\text{span}\,(e_i, i \in \mathbb{I})}. \tag{77}$$

In addition, $I_\kappa{}^*$ also admits the spectral decomposition

$$I_\kappa{}^* = \sum_{i\in\mathbb{I}} \sigma_i \left\langle \cdot, [e_i] \right\rangle_\rho e_i, \tag{78}$$

per (Steinwart and Scovel, 2012, Theorem 2.11).

Recall that $\left(\sigma_i^{1/2} e_i\right)_{i\in\mathbb{I}}$ is an ONB of $(\ker I_\kappa)^\perp$, $([e_i])_{i\in\mathbb{I}}$ an ONB of $\overline{\text{range} I_\kappa}$. Since every ONS can be extended to an ONB, let $(\widetilde{e}_i)_{i\in\mathbb{J}}$ be an ONB of $\ker I_\kappa$ with $\mathbb{J} \cap \mathbb{N} = \varnothing$ such that $\left(\sigma_i^{1/2} e_i\right)_{i\in\mathbb{I}} \cup (\widetilde{e}_i)_{i\in\mathbb{J}}$ is an ONB of $\mathcal{H}_X$. We then have the spectral representation for $0 \le \gamma \le 1$ and $a > 0$,

$$C_\kappa{}^{\frac{1-\gamma}{2}} = \sum_{i\in\mathbb{I}} \sigma_i^{\frac{1-\gamma}{2}} \left\langle \cdot, \sigma_i^{1/2} e_i \right\rangle_{\mathcal{H}} \sigma_i^{1/2} e_i, \tag{79}$$

$$(C_\kappa + \lambda \text{Id})^{-a} = \sum_{i\in\mathbb{I}} (\sigma_i + \lambda)^{-a} \left\langle \sigma_i^{1/2} e_i, \cdot \right\rangle_{\mathcal{H}} \sigma_i^{1/2} e_i + \lambda^{-a} \sum_{j\in\mathbb{J}} \left\langle \widetilde{e}_j, \cdot \right\rangle_{\mathcal{H}} \widetilde{e}_j, \tag{80}$$

per Fischer and Steinwart (2020). Based on the above spectral representations, we derive the bound bound below which will be useful later. By the definition of operator norm, we have

$$\left\| C_\kappa{}^{\frac{1-\gamma}{2}} (C_\kappa + \lambda \text{Id})^{-1/2} \right\|_{\text{op}}^2 = \sup_{\|f\|_{\mathcal{H}}=1} \left\| C_\kappa{}^{\frac{1-\gamma}{2}} (C_\kappa + \lambda \text{Id})^{-1/2} f \right\|_{\mathcal{H}}^2, \tag{81}$$

We next expand $C_\kappa{}^{\frac{1-\gamma}{2}} (C_\kappa + \lambda \text{Id})^{-1/2}$ based on (79), (80), and we have

$$\begin{aligned}
&\left\| C_\kappa{}^{\frac{1-\gamma}{2}} (C_\kappa + \lambda \text{Id})^{-1/2} f \right\|_{\mathcal{H}}^2 \\
&= \left\| \sum_{i\in\mathbb{I}} \sigma_i^{\frac{1-\gamma}{2}} \left\langle \sum_{i'\in\mathbb{I}} (\sigma_{i'} + \lambda)^{-1/2} \left\langle \sigma_{i'}^{1/2} e_{i'}, f \right\rangle \sigma_{i'}^{1/2} e_{i'} + \lambda^{-1/2} \sum_{j\in\mathbb{J}} \left\langle \widetilde{e}_j, f \right\rangle_{\mathcal{H}} \widetilde{e}_j \,,\, \sigma_i^{1/2} e_i \right\rangle_{\mathcal{H}} \sigma_i^{1/2} e_i \right\|_{\mathcal{H}}^2 \\
&\overset{(a)}{=} \left\| \sum_{i\in\mathbb{I}} \sigma_i^{\frac{1-\gamma}{2}} (\sigma_i + \lambda)^{-1/2} \left\langle \sigma_i^{1/2} e_i, f \right\rangle_{\mathcal{H}} \sigma_i^{1/2} e_i \right\|_{\mathcal{H}}^2 \\
&= \sum_{i\in\mathbb{I}} \sum_{i'\in\mathbb{I}} \left\langle \sigma_i^{\frac{1-\gamma}{2}} (\sigma_i + \lambda)^{-1/2} \left\langle \sigma_i^{1/2} e_i, f \right\rangle_{\mathcal{H}} \sigma_i^{1/2} e_i \,,\, \sigma_{i'}^{\frac{1-\gamma}{2}} (\sigma_{i'} + \lambda)^{-1/2} \left\langle \sigma_{i'}^{1/2} e_{i'}, f \right\rangle_{\mathcal{H}} \sigma_{i'}^{1/2} e_{i'} \right\rangle_{\mathcal{H}} \\
&= \sum_{i\in\mathbb{I}} \sum_{i'\in\mathbb{I}} \sigma_i^{\frac{1-\gamma}{2}} (\sigma_i + \lambda)^{-1/2} \sigma_{i'}^{\frac{1-\gamma}{2}} (\sigma_{i'} + \lambda)^{-1/2} \left\langle \sigma_i^{1/2} e_i, f \right\rangle_{\mathcal{H}} \left\langle \sigma_{i'}^{1/2} e_{i'}, f \right\rangle_{\mathcal{H}} \left\langle \sigma_i^{1/2} e_i, \sigma_{i'}^{1/2} e_{i'} \right\rangle_{\mathcal{H}} \\
&= \sum_{i\in\mathbb{I}} \frac{\sigma_i^{1-\gamma}}{\sigma_i + \lambda} \left| \left\langle \sigma_i^{1/2} e_i, f \right\rangle_{\mathcal{H}} \right|^2.
\end{aligned} \tag{82}$$

In deriving the above expression, (a) holds since $\left(\sigma_i^{1/2} e_i\right)_{i \in \mathbb{I}}$ is an ONB of $(\ker I_\kappa)^\perp$ and $(\widetilde{e}_i)_{i \in \mathbb{J}}$ is an ONB of $\ker I_\kappa$. Therefore, we have

$$\begin{aligned}
\left\| C_\kappa^{\frac{1-\gamma}{2}} (C_\kappa + \lambda \mathrm{Id})^{-1/2} f \right\|_{\mathcal{H}}^2 &= \sup_{\|f\|_{\mathcal{H}}=1} \sum_{i \in \mathbb{I}} \frac{\sigma_i^{1-\gamma}}{\sigma_i + \lambda} \left| \left\langle \sigma_i^{1/2} e_i, f \right\rangle_{\mathcal{H}} \right|^2 \\
&\leq \sup_{\|f\|_{\mathcal{H}}=1} \left(\sup_{i \in \mathbb{I}} \frac{\sigma_i^{1-\gamma}}{\sigma_i + \lambda} \right) \sum_{i \in \mathbb{I}} \left| \left\langle \sigma_i^{1/2} e_i, f \right\rangle_{\mathcal{H}} \right|^2 \\
&\overset{(a)}{=} \sup_{\|f\|_{\mathcal{H}}=1} \left(\sup_{i \in \mathbb{I}} \frac{\sigma_i^{1-\gamma}}{\sigma_i + \lambda} \right) \|f\|_{\mathcal{H}}^2 \\
&= \sup_{i \in \mathbb{I}} \frac{\sigma_i^{1-\gamma}}{\sigma_i + \lambda} \\
&\leq \lambda^{-\gamma},
\end{aligned} \tag{83}$$

where (a) follows from the Parseval's identity.

C.1. Bounding the γ-norm

We next introduce the following lemma that upper bound the γ-norm using the HS-norm.

Lemma 9 *For $u \in \mathcal{H}_V$, let $U = \iota_\kappa^{-1}(u) \in HS(\mathcal{H}_X, \mathcal{H}_Y)$, where ι_κ is the linear isomorphism in Lemma 1. For any $\gamma \in [0,1]$ and $U \in HS(\mathcal{H}_X, \mathcal{H}_Y)$, we have*

$$\|[u]\|_\gamma^2 \leq \lambda^{-(\gamma+1)} (K + \lambda)^2 \|U\|_{HS}^2 . \tag{84}$$

Proof By (Li et al., 2022, Lemma 2), we have

$$\|[u]\|_\gamma \leq \left\| U C_{XX}^{\frac{1-\gamma}{2}} \right\|_{\mathrm{HS}} . \tag{85}$$

First, notice that for a self-adjoint invertible operator A, $A^{-1/2} \times A \times A^{-1/2} = \mathrm{Id}$. Thus we have

$$U C_{XX}^{\frac{1-\gamma}{2}} = U (C_{XX} + \lambda \mathrm{Id})^{-1/2} (C_{XX} + \lambda \mathrm{Id})^{1/2} (C_{XX} + \lambda \mathrm{Id})^{1/2} (C_{XX} + \lambda \mathrm{Id})^{-1/2} C_{XX}^{\frac{1-\gamma}{2}} . \tag{86}$$

Repeatedly applying (40) twice, we get

$$\begin{aligned}
\left\| U C_{XX}^{\frac{1-\gamma}{2}} \right\|_{\mathrm{HS}}^2 &= \left\| U (C_{XX} + \lambda \mathrm{Id})^{-1/2} (C_{XX} + \lambda \mathrm{Id}) (C_{XX} + \lambda \mathrm{Id})^{-1/2} C_{XX}^{\frac{1-\gamma}{2}} \right\|_{\mathrm{HS}}^2 \\
&\leq \left\| U (C_{XX} + \lambda \mathrm{Id})^{-1/2} (C_{XX} + \lambda \mathrm{Id}) \right\|_{\mathrm{HS}}^2 \left\| (C_{XX} + \lambda \mathrm{Id})^{-1/2} C_{XX}^{\frac{1-\gamma}{2}} \right\|_{\mathrm{op}}^2 \\
&\leq \left\| U (C_{XX} + \lambda \mathrm{Id})^{-1/2} \right\|_{\mathrm{HS}}^2 \times \|C_{XX} + \lambda \mathrm{Id}\|_{\mathrm{op}}^2 \times \left\| (C_{XX} + \lambda \mathrm{Id})^{-1/2} C_{XX}^{\frac{1-\gamma}{2}} \right\|_{\mathrm{op}}^2 .
\end{aligned} \tag{87}$$

We can upper bound the second term in in (87) as follows:

$$\|C_{XX}+\lambda\mathrm{Id}\|_{\mathrm{op}}^2 \le \left(\|C_{XX}\|_{\mathrm{op}}+\lambda\right)^2 \le \left(\|C_{XX}\|_{\mathrm{HS}}+\lambda\right)^2 \le (K+\lambda)^2 . \tag{88}$$

For the last term, by the self-adjointness of C_{XX}, we have

$$\left\|\left(C_{XX}+\lambda\mathrm{Id}\right)^{-1/2} C_{XX}^{\frac{1-\gamma}{2}}\right\|_{\mathrm{op}}^2 = \left\|C_{XX}^{\frac{1-\gamma}{2}}\left(C_{XX}+\lambda\mathrm{Id}\right)^{-1/2}\right\|_{\mathrm{op}}^2,$$

where the term on the right hand side is bounded in (83).

We next bound the first term in (87) via spectrum decomposition. Let $\{d_l\}_{l\in\mathbb{I}_Y}$ be a basis of $\mathcal{H}_Y$. Recall that $\left(\sigma_i^{1/2}e_i\right)_{i\in\mathbb{I}}\cup\left(\widetilde{e}_i\right)_{i\in\mathbb{J}}$ is an ONB of $\mathcal{H}_X$ and denote

$$f_i=\begin{cases}\sigma_i^{1/2}e_i & i\in\mathbb{I}\\ \widetilde{e}_i & i\in\mathbb{J}\end{cases}$$

Then $\{d_l\otimes f_i\}_{i\in\mathbb{I}\cup\mathbb{J},l\in\mathbb{I}_Y}$ is an ONB of $\mathrm{HS}\left(\mathcal{H}_X,\mathcal{H}_Y\right)$. Therefore, for $U\in\mathrm{HS}\left(\mathcal{H}_X,\mathcal{H}_Y\right)$, we have

$$U=\sum_{i\in\mathbb{I}}\sum_{l\in\mathbb{I}_Y}a_{il}d_l\otimes\sigma_i^{1/2}e_i+\sum_{j\in\mathbb{J}}\sum_{l\in\mathbb{I}_Y}a_{jl}d_l\otimes\widetilde{e}_j,\quad a_{il}=\langle U,d_l\otimes f_i\rangle_{\mathrm{HS}} . \tag{89}$$

From the above representation, we have

$$\begin{aligned}
&\left\|U\left(C_{XX}+\lambda\mathrm{Id}\right)^{-1/2}\right\|_{\mathrm{HS}}^2\\
=&\Bigg\|\left(\sum_{i\in\mathbb{I}}\sum_{l\in\mathbb{I}_Y}a_{il}d_l\otimes\sigma_i^{1/2}e_i+\sum_{j\in\mathbb{J}}\sum_{l\in\mathbb{I}_Y}a_{jl}d_l\otimes\widetilde{e}_j\right)\\
&\left(\sum_{k\in\mathbb{I}}\left(\sigma_k+\lambda\right)^{-1/2}\left\langle\cdot,\sigma_k^{1/2}e_k\right\rangle_{\mathrm{HS}}\sigma_k^{1/2}e_k+\lambda^{-1/2}\sum_{k'\in\mathbb{J}}\left\langle\cdot,\widetilde{e}_{k'}\right\rangle_{\mathrm{HS}}\widetilde{e}_{k'}\right)\Bigg\|_{\mathrm{HS}}^2\\
=&\left\|\sum_{i\in\mathbb{I}}\sum_{l\in\mathbb{I}_Y}a_{ij}\left(\sigma_i+\lambda\right)^{-1/2}\left\langle\sigma_i^{1/2}e_i,\sigma_i^{1/2}e_i\right\rangle_{\mathcal{H}_X}d_j\otimes\sigma_i^{1/2}e_i+\sum_{j\in\mathbb{J}}\sum_{l\in\mathbb{I}_Y}a_{jl}\lambda^{-1/2}\left\langle\widetilde{e}_j,\widetilde{e}_j\right\rangle_{\mathrm{HS}}d_l\otimes\widetilde{e}_j\right\|_{\mathrm{HS}}^2,
\end{aligned} \tag{90}$$

the last line follows from (42). Since $\left(\sigma_i^{1/2} e_i\right)_{i\in\mathbb{I}} \cup (\widetilde{e}_j)_{j\in\mathbb{J}}$ is an ONB of $\mathcal{H}_X$, we can simply the last line as

$$\begin{aligned}
\left\|U\left(C_{XX}+\lambda \mathrm{Id}\right)^{-1/2}\right\|_{\mathrm{HS}}^2 =& \left\|\sum_{i\in\mathbb{I}}\sum_{l\in\mathbb{I}_Y} \frac{a_{ij}}{(\sigma_i+\lambda)^{1/2}} d_j \otimes \sigma_i^{1/2} e_i + \sum_{j\in\mathbb{J}}\sum_{l\in\mathbb{I}_Y} \frac{a_{jl}}{\lambda^{1/2}} d_l \otimes \widetilde{e}_j\right\|_{\mathrm{HS}}^2 \\
\overset{(a)}{=}& \sum_{i\in\mathbb{I}}\sum_{l\in\mathbb{I}_Y}\left(\frac{a_{ij}}{(\sigma_i+\lambda)^{1/2}}\right)^2 + \sum_{j\in\mathbb{J}}\sum_{l\in\mathbb{I}_Y}\left(\frac{a_{jl}}{\lambda^{1/2}}\right)^2 \\
\leq& \sup_{i,j} \frac{1}{(\sigma_i+\lambda)^{1/2}} \sum_{i\in\mathbb{I}}\sum_{l\in\mathbb{I}_Y} a_{ij}^2 + \frac{1}{\lambda}\sum_{j\in\mathbb{J}}\sum_{l\in\mathbb{I}_Y} a_{jl}^2 \\
\leq& \lambda^{-1}\sum_{i\in\mathbb{I}}\sum_{l\in\mathbb{I}_Y} a_{ij}^2 + \frac{1}{\lambda}\sum_{j\in\mathbb{J}}\sum_{l\in\mathbb{I}_Y} a_{jl}^2 \\
\leq& \lambda^{-1}\sum_{i\in\mathbb{I}\cup\mathbb{J}}\sum_{l\in\mathbb{I}_Y} a_{ij}^2 \\
\overset{(b)}{=}& \|U\|_{\mathrm{HS}}^2/\lambda,
\end{aligned} \tag{91}$$

(a) and (b) follows from Parserval's identity w.r.t the ONB $\{d_l \otimes f_i\}_{i\in\mathbb{I}\cup\mathbb{J}, l\in\mathbb{I}_Y}$ of $\mathrm{HS}\left(\mathcal{H}_X, \mathcal{H}_Y\right)$. Combining the three bounds concludes the proof. ■

Appendix D. Proof of Lemma 3

We prove this lemma by induction. Let $\widetilde{U}_0 = 0$. After we receive (x_1, y_1), we update the estimate as

$$\widetilde{U}_1 = \eta_1 \widetilde{C}_{YX}(1) = \eta_1 \kappa_Y(y_1, \cdot) \otimes \kappa_X(x_1, \cdot). \tag{92}$$

Assume at t-th iteration, $\widetilde{U}_t = \sum_{i=1,j=1}^{t-1} \widetilde{W}_{t-1}^{ij} \kappa_Y(y_i, \cdot) \otimes \kappa_X(x_j, \cdot)$. Then, we have

$$\begin{aligned}
\widetilde{U}_t \widetilde{C}_{XX}(t) =& \Big[\sum_{i=1}^{t-1}\sum_{j=1}^{t-1} \widetilde{W}_{t-1}^{ij} \kappa_Y(y_i, \cdot) \otimes \kappa_X(x_j, \cdot)\Big] \Big[\kappa_X(x_t, \cdot) \otimes \kappa_X(x_t, \cdot)\Big] \\
\overset{(a)}{=}& \sum_{i=1,j=1}^{t-1} \widetilde{W}_{t-1}^{ij} \Big[\left(\kappa_Y(y_i, \cdot) \otimes \kappa_X(x_j, \cdot)\right) \kappa_X(x_t, \cdot)\Big] \otimes \kappa_X(x_t, \cdot) \\
\overset{(b)}{=}& \sum_{i=1,j=1}^{t-1} \widetilde{W}_{t-1}^{ij} \Big(\left\langle \kappa_X(x_t, \cdot), \kappa_X(x_j, \cdot)\right\rangle_{\mathcal{H}} \kappa_Y(y_i, \cdot)\Big) \otimes \kappa_X(x_t, \cdot) \\
=& \sum_{i=1,j=1}^{t-1} \widetilde{W}_{t-1}^{ij} \kappa_X(x_j, x_t) \; \left[\kappa_Y(y_i, \cdot) \otimes \kappa_X(x_t, \cdot)\right],
\end{aligned} \tag{93}$$

where (a) follows from (42) and (b) follows from the definition of tensor products.

Substituting the above equation into (11) for $t+1$ gives

$$
\begin{aligned}
\widetilde{U}_{t+1} =&(1-\lambda\eta_t)\widetilde{U}_t - \eta_t\left(\widetilde{U}_t\widetilde{C}_{XX}(t) - \widetilde{C}_{YX}(t)\right) \\
=&(1-\lambda\eta_t)\sum_{i=1,j=1}^{t-1}\widetilde{W}_{t-1}^{ij}\kappa_Y(y_i,\cdot)\otimes\kappa_X(x_j,\cdot) - \eta_t\sum_{i=1,j=1}^{t-1}\widetilde{W}_{t-1}^{ij}\kappa_X(x_j,x_t)\kappa_Y(y_i,\cdot)\otimes\kappa_X(x_t,\cdot) \\
&+\eta_t\kappa(y_t,\cdot)\otimes\kappa_X(x_t,\cdot) \\
=&\sum_{i=1,j=1}^{t}\widetilde{W}_t^{ij}\kappa_Y(y_i,\cdot)\otimes\kappa_X(x_j,\cdot) \\
=&\widetilde{\Psi}_{Y,t}\widetilde{W}_t\widetilde{\Phi}_{X,t}^\top,
\end{aligned} \tag{94}
$$

where the (i,j)-th element of $\widetilde{W}_t$ is given by (13).

Appendix E. Algorithm

The compressed online learning algorithm presented in Section 3.2 is summarized below. A more detailed implementation based on finite-dimensional Gram matrices is deferred to Appendix G.

Algorithm 1: Compressed Online Learning of the CME

input : Sample pairs $\{(x_t,y_t)\}_{t\in\mathbb{T}}$; Kernel κ; Step-sizes $\{\eta_t\}_{t\in\mathbb{T}}$; Compression budget $\{\varepsilon_t\}_{t\in\mathbb{T}}$

Initialize $U_0 = 0$

for $t\in\mathbb{T}$ **do**
- Receive sample pair (x_t,y_t)
- $\widetilde{\mathcal{D}}_t \leftarrow \mathcal{D}_{t-1}\cup(x_t,y_t)$
- Compute $\widetilde{W}_t$ based on $\widetilde{\mathcal{D}}_t$ via (13)
- Compute

$$
\Delta_t \leftarrow \min_Z\left\|\sum_{i,j\in\mathcal{I}_{t-1}}Z^{ij}\kappa_Y(y_i,\cdot)\otimes\kappa_X(x_j,\cdot) - \sum_{i=1,j=1}^{t}\widetilde{W}_t^{ij}\kappa_Y(y_i,\cdot)\otimes\kappa_X(x_j,\cdot)\right\|_{\mathrm{HS}}^2
$$

- **if** $\Delta_t<\varepsilon_t$ **then**
  - $\mathcal{D}_t\leftarrow\mathcal{D}_{t-1}$,$W_t\leftarrow Z_\star$
- **else**
  - $\mathcal{D}_t\leftarrow\widetilde{\mathcal{D}}_t$, $W_t\leftarrow\widetilde{W}_t$
- **end**
- Compute U_t according to (18) and $\mu_t(\cdot)\leftarrow U_t\phi_X(\cdot)$

end

Appendix F. Proof of Results in Section 4

F.1. Proof of Lemma 5

We first recall the definition of total variation distance for distributions over a standard measurable space $(\mathbb{X},\mathcal{B}_X)$ where $\mathcal{B}_X$ is a sigma-field. Consider two distributions P,Q defined on $(\mathbb{X},\mathcal{B}_X)$ with Radon-Nikodym derivatives p and q with respect to a reference measure ν on $\mathbb{X}$, respectively. Then, using Scheffe's Lemma (see e.g., (Takezawa, 2005, Lemma 2.1)), the total variation distance between P and Q denoted by $\|P-Q\|_{\mathrm{TV}}$ is given by

$$\|P-Q\|_{\mathrm{TV}} = \frac{1}{2}\int_{\mathbb{X}} |p(x)-q(x)| \mathrm{d}\nu(x). \tag{95}$$

In order to prove Lemma 5, we apply generalized Jensen's inequality for HS-valued random variables. Let p_t, q be the Radon-Nikodym derivatives of $P_t(\cdot|\mathcal{F}_0)$ and $\rho(\cdot)$ with respect to the reference measure ν on $\mathbb{X}\times\mathbb{Y}$. For $t \geq \tau_\delta$, we write the Bochner conditional expectation as Bochner integral w.r.t $P_t(\cdot|\mathcal{F}_0)$, $\rho(\cdot)$ and obtain

$$\begin{aligned}
&\left\|\mathbb{E}\left[\widetilde{\nabla} R_\lambda(x_t, y_t, U)\,|\mathcal{F}_0\right] - \nabla R_\lambda(U)\right\|_{\mathrm{HS}} \\
=&\left\|\int_{\mathbb{X}\times\mathbb{Y}} \widetilde{\nabla} R_\lambda(x,y,U)\, P_t(x,y|\mathcal{F}_0) - \int_{\mathbb{X}\times\mathbb{Y}} \widetilde{\nabla} R_\lambda(x,y,U)\,\mathrm{d}\rho(x,y)\right\|_{\mathrm{HS}} \\
=&\left\|\int_{\mathbb{X}\times\mathbb{Y}} \widetilde{\nabla} R_\lambda(x,y,U)\, p_t(x,y)\,\mathrm{d}\nu(x,y) - \int_{\mathbb{X}\times\mathbb{Y}} \widetilde{\nabla} R_\lambda(x,y,U)\, q(x,y)\,\mathrm{d}\nu(x,y)\right\|_{\mathrm{HS}} \\
\overset{(a)}{\leq}& \int_{\mathbb{X}\times\mathbb{Y}} \left\|\widetilde{\nabla} R_\lambda(x,y,U)\right\|_{\mathrm{HS}} |p_t(x,y)-q(x,y)|\,\mathrm{d}\nu(x,y),
\end{aligned} \tag{96}$$

where (a) holds since $\widetilde{\nabla} R_\lambda(x,y,U)$ is Bochner integrable.

By the affine scaling property stated in Lemma 8 and (26), we have for any $t \geq \tau_\delta$

$$\begin{aligned}
\left\|\mathbb{E}\left[\widetilde{\nabla} R_\lambda(x_t, y_t, U)\,|\mathcal{F}_0\right] - \nabla R_\lambda(U)\right\|_{\mathrm{HS}} \leq& B_\kappa\left(\|U\|_{\mathrm{HS}}+1\right)\int_{\mathbb{X}\times\mathbb{Y}} |p_t(x,y)-q(x,y)|\,\mathrm{d}\nu(x,y) \\
\overset{(a)}{=}& 2B_\kappa\left(\|U\|_{\mathrm{HS}}+1\right)\|P_t(\cdot|\mathcal{F}_0)-\rho(\cdot)\|_{\mathrm{TV}} \\
\leq& 2B_\kappa\delta\left(\|U\|_{\mathrm{HS}}+1\right),
\end{aligned} \tag{97}$$

where (a) follows from (95).

F.2. Proof of Lemma 6

Before proving Lemma 6, we present two lemmas which will be useful later.

Lemma 10 *(Uniform boundedness) Let Assumptions 1 and 2 hold. Then U_λ and the iterates $\{U_t\}_{t\in\mathbb{T}}$ are uniformly bounded as*

$$\|U_t\|_{HS} \leq \frac{K}{\lambda}, \quad \|U_\lambda\|_{HS} \leq \frac{K}{\lambda}, \quad \forall t \in \mathbb{T}. \tag{98}$$

Proof To begin with, notice that once the dictionary $\mathcal{D}_t$ and the coefficient matrix W_t have been updated, (18) can be interpreted as implementing orthogonal projection of $\widetilde{U}_{t+1}$ onto the closed subspace defined by $\mathcal{D}_t$, i.e.,

$$U_{t+1} = \Pi_{\mathcal{D}_t}[\widetilde{U}_{t+1}], \quad t \in \mathbb{T}. \tag{99}$$

We next establish (10) by induction. At time $t = 1$, we have

$$\|U_1\|_{\text{HS}} = \|\Pi_{\mathcal{D}_0}[U_1]\|_{\text{HS}} \overset{(a)}{\leq} \|U_1\|_{\text{HS}} = \left\|\eta_1 \widetilde{C}_{YX}(1)\right\|_{\text{HS}} \leq \eta_1 K \overset{(b)}{\leq} K/\lambda, \tag{100}$$

where (a) follows from the non-expansive property of the projection operator and (b) follows from Assumption 2. Thus the base case holds. Now assume $\|U_k\|_{\text{HS}} \leq \frac{K}{\lambda}$ for $k = 1, \ldots, t$. Then, at time $t+1$, using the non-expansive property of the projection operator again, we have

$$\|U_{t+1}\|_{\text{HS}} = \left\|\Pi_{\mathcal{D}_t}\left[\widetilde{U}_{t+1}\right]\right\|_{\text{HS}} \leq \left\|\widetilde{U}_{t+1}\right\|_{\text{HS}}. \tag{101}$$

We then expand $\widetilde{U}_{t+1}$ using (11) and we have

$$\begin{aligned}
\|U_{t+1}\|_{\text{HS}} &= \left\|(\text{Id} - \lambda\eta_t) U_t - \eta_t U_t \widetilde{C}_{XX}(t) + \eta_t \widetilde{C}_{YX}(t)\right\|_{\text{HS}} \\
&= \left\|U_t\left(\text{Id} - \eta_t\left(\lambda\text{Id} + \widetilde{C}_{XX}(t)\right)\right) + \eta_t \widetilde{C}_{YX}(t)\right\|_{\text{HS}} \\
&\leq \|U_t\|_{\text{HS}} \left\|\text{Id} - \eta_t\left(\lambda\text{Id} + \widetilde{C}_{XX}(t)\right)\right\|_{\text{op}} + \eta_t \left\|\widetilde{C}_{YX}(t)\right\|_{\text{HS}},
\end{aligned} \tag{102}$$

where the last line holds due to triangle inequality and the relation $\|AB\|_{\text{HS}} \leq \|A\|_{\text{HS}} \|B\|_{\text{op}}$. Furthermore, recall that the operator norm of a self-adjoint operator coincides with its maximum eigenvalue. Since $\widetilde{C}_{XX}(t)$ is self-adjoint, denote $\kappa_{x_t} := \kappa_X(x_t, \cdot)$ and we have

$$\begin{aligned}
\|\text{Id} - \eta_t(\kappa_{x_t} \otimes \kappa_{x_t} + \lambda\text{Id})\|_{\text{op}} =& \sigma_{\max}\left((\text{Id} - \eta_t(\kappa_{x_t} \otimes \kappa_{x_t} + \lambda\text{Id}))\right) \\
\leq& 1 - \eta_t \sigma_{\min}(\kappa_{x_t} \otimes \kappa_{x_t} + \lambda I) \\
\leq& 1 - \eta_t \lambda.
\end{aligned} \tag{103}$$

Hence, we conclude

$$\|U_{t+1}\|_{\text{HS}} \leq \|U_t\|_{\text{HS}}(1 - \eta_t\lambda) + \eta_t \left\|\widetilde{C}_{YX}(t)\right\|_{\text{HS}} \leq \frac{K}{\lambda}(1 - \eta_t\lambda) + \eta_t K = \frac{K}{\lambda}. \tag{104}$$

In addition, U_λ satisfies

$$\begin{aligned}
\|U_\lambda\|_{\text{HS}} &= \left\|C_{YX}(C_{XX} + \lambda\text{Id})^{-1}\right\|_{\text{HS}} \\
&\overset{(a)}{=} \left\|\left(C_{YX}(C_{XX} + \lambda\text{Id})^{-1}\right)^*\right\|_{\text{HS}} \\
&= \left\|(C_{XX} + \lambda\text{Id})^{-1} C_{YX}^*\right\|_{\text{HS}} \\
&\overset{(b)}{\leq} \left\|(C_{XX} + \lambda\text{Id})^{-1}\right\|_{\text{op}} \|C_{YX}^*\|_{\text{HS}} \\
&\overset{(c)}{=} \left\|(C_{XX} + \lambda\text{Id})^{-1}\right\|_{\text{op}} \|C_{YX}\|_{\text{HS}} \\
&\overset{(d)}{\leq} \frac{\|C_{YX}\|_{\text{HS}}}{\lambda} \\
&\overset{(e)}{\leq} \frac{K}{\lambda},
\end{aligned} \tag{105}$$

where (a) and (c) holds since $\|A\|_{\mathrm{HS}} = \|A^*\|_{\mathrm{HS}}$, (b) follows from the fact that $\|BA\|_{\mathrm{HS}} \leq \|B\|_{\mathrm{op}} \|A\|_{\mathrm{HS}}$ for Hilbert-Schmidt operator A and bounded linear operator B, (d) holds since $\left\|(C_{XX} + \lambda \mathrm{Id})^{-1}\right\|_{\mathrm{op}} \leq 1/\lambda$ and (e) follows from an upper bound on the HS-norm of C_{YX} given by

$$\begin{aligned}
\|C_{YX}\|_{\mathrm{HS}}^2 =& \|\mathbb{E}\left[\kappa_Y(y,\cdot) \otimes \kappa_X(x,\cdot)\right]\|_{\mathrm{HS}}^2 \\
=& \left\langle \mathbb{E}_{YX}\left[\kappa_Y(y,\cdot) \otimes \kappa_X(x,\cdot)\right], \mathbb{E}_{Y'X'}\left[\kappa_Y(y',\cdot) \otimes \kappa_X(x',\cdot)\right]\right\rangle_{\mathrm{HS}} \\
=& \mathbb{E}_{YX,Y'X'}\left[\left\langle \kappa_Y(y,\cdot) \otimes \kappa_X(x,\cdot), \kappa_Y(y',\cdot) \otimes \kappa_X(x',\cdot)\right\rangle_{\mathrm{HS}}\right] \\
=& \mathbb{E}_{YX,Y'X'}\left[\kappa_Y(y,y')\kappa_Y(x,x')\right] \\
\leq& K^2
\end{aligned}$$

with (X', Y') being an independent copy of (X, Y). ■

The next lemma characterizes the difference between two iterates via the sum of stepsizes and the norm of an iterate and will be used in the proof of Lemma 6. A similar result for stochastic approximation in finite-dimensional Euclidean space appeared in (Srikant and Ying, 2019, Lemma 3) and Chen et al. (2022). Here, we consider *nonlinear* stochastic recursion in the space of *Hilbert-Schmidt operators* which is *infinite-dimensional* and make use of properties of operator-valued gradients presented in Lemma 8.

Lemma 11 *Let Assumptions 1 and 2 hold. Let $s < r$ such that $(r-s)\eta \leq 1/4B$, where $B = B_\kappa + B_{cmp}$. Then:*

(a) $\|U_s - U_r\|_{HS} \leq 2B(r-s)\eta\left(\|U_s\|_{HS} + 1\right)$,

(b) $\|U_s - U_r\|_{HS} \leq 4B(r-s)\eta\left(\|U_r\|_{HS} + 1\right)$.

Proof Recall from Lemma 8, the operator-valued stochastic gradient scales affinely with respect to the current iterates. We leverage this property to provide a bound for $\|U_{t+1}\|_{\mathrm{HS}}$ in terms of $\|U_t\|_{\mathrm{HS}}$, and repeatedly apply this results to bound $U_s - U_r$. To begin with, since $\eta \in (0,1)$, by Assumption 2(b), we have

$$\varepsilon \leq B_{\mathrm{cmp}}\eta^2 \leq B_{\mathrm{cmp}}\eta \leq B_{\mathrm{cmp}}\eta + B_{\mathrm{cmp}}\eta \|U_t\|_{\mathrm{HS}} = B_{\mathrm{cmp}}\eta\left(\|U_t\|_{\mathrm{HS}} + 1\right). \tag{106}$$

Let $t \in [s, r]$, and we have

$$\begin{aligned}
\|U_{t+1} - U_t\|_{\mathrm{HS}} =& \eta\left\| -\widetilde{\nabla} R_\lambda(x_t, y_t, U_t) + \frac{E_t}{\eta}\right\|_{\mathrm{HS}} \\
\leq& \eta\left\| -\widetilde{\nabla} R_\lambda(x_t, y_t, U_t)\right\|_{\mathrm{HS}} + \|E_t\|_{\mathrm{HS}}
\end{aligned} \tag{107}$$

By Lemma 8 (c) and condition $\|E_t\|_{\mathrm{HS}} \leq \varepsilon$, we have

$$\begin{aligned}
\|U_{t+1} - U_t\|_{\mathrm{HS}} \leq& \eta B_\kappa\left(\|U_t\|_{\mathrm{HS}} + 1\right) + \varepsilon \\
\leq& \eta B_\kappa\left(\|U_t\|_{\mathrm{HS}} + 1\right) + B_{\mathrm{cmp}}\eta\left(\|U_t\|_{\mathrm{HS}} + 1\right) \\
=& \eta\left(B_\kappa + B_{\mathrm{cmp}}\right)\left(\|U_t\|_{\mathrm{HS}} + 1\right) \\
\leq& \eta B\left(\|U_t\|_{\mathrm{HS}} + 1\right).
\end{aligned} \tag{108}$$

Triangle inequality gives

$$\|U_{t+1}\|_{\text{HS}} \leq \|U_t\|_{\text{HS}} + \|U_{t+1} - U_t\|_{\text{HS}} \leq (\eta B + 1) \|U_t\|_{\text{HS}} + \eta B. \tag{109}$$

As a result, we have

$$\|U_{t+1}\|_{\text{HS}} + 1 \leq (\eta B + 1) \left(\|U_t\|_{\text{HS}} + 1\right). \tag{110}$$

By recursively applying the above inequality, we have

$$\|U_t\|_{\text{HS}} + 1 \leq \Pi_{i=s}^{t-1} (\eta B + 1) \left(\|U_s\|_{\text{HS}} + 1\right). \tag{111}$$

Using $1 + x \leq e^x$ for $x \in \mathbb{R}$ and $e^x \leq 1 + 2x$ for $x \in [0, 1/2]$, we obtain

$$\begin{aligned} \|U_t\|_{\text{HS}} + 1 \leq& \exp\left(B(r-s)\eta\right) \left(\|U_s\|_{\text{HS}} + 1\right) \\ \leq& \left(1 + 2B(r-s)\eta\right) \left(\|U_s\|_{\text{HS}} + 1\right) \\ \leq& 2\left(\|U_s\|_{\text{HS}} + 1\right), \end{aligned} \tag{112}$$

where the last inequality follows from assumption $(r-s)\eta \leq 1/4B$, and we conclude

$$\|U_{t+1} - U_t\|_{\text{HS}} \leq 2\eta B \left(\|U_s\|_{\text{HS}} + 1\right). \tag{113}$$

Therefore, we have

$$\|U_r - U_s\|_{\text{HS}} \leq \sum_{t=s}^{r-1} \|U_{t+1} - U_t\|_{\text{HS}} \leq 2B \sum_{t=s}^{r-1} \eta \left(\|U_s\|_{\text{HS}} + 1\right) = 2B(r-s)\eta \left(\|U_s\|_{\text{HS}} + 1\right). \tag{114}$$

This proves the first claim.

Next, applying triangle inequality and the assumption $(r-s)\eta \leq 1/4B$ gives

$$\begin{aligned} \|U_r - U_s\|_{\text{HS}} \leq& 2B(r-s)\eta \left(\|U_s\|_{\text{HS}} + 1\right) \\ \leq& 2B(r-s)\eta \left(\|U_r\|_{\text{HS}} + \|U_r - U_s\|_{\text{HS}} + 1\right) \\ \leq& \frac{1}{2} \|U_r - U_s\|_{\text{HS}} + 2B(r-s)\eta \left(\|U_r\|_{\text{HS}} + 1\right). \end{aligned} \tag{115}$$

Rearranging terms gives the second claim. ■

We now return to the proof of Lemma 6. We use an analogous argument as that used in the proof of (Chen et al., 2022, Theorem 2.1) by decomposing the one-time-step drift of the Lyapunov function, which tracks the difference of the RKHS norm distance to a nominal element U_λ at two different iterates U_{t+1} and U_t. By definition, we can write $W(\cdot)$ as an inner product in HS, i.e., $W(U_t) = \langle U_t - U_\lambda, U_t - U_\lambda \rangle_{\text{HS}}$. For $t \geq \tau_\eta$, we aim to express the drift of the Lyapunov function $\mathbb{E}\left[\|U_{t+1} - U_\lambda\|_{\text{HS}}^2 - \|U_t - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right]$ in terms of $U_{t+1} - U_t$ and then derive an upper by invoking Lemma 8 and the mixing property in Assumption 3. To this end, first notice that

$$\begin{aligned} &\mathbb{E}\left[\|U_{t+1} - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] - \mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] \\ =& \mathbb{E}\left[\|(U_{t+1} - U_t) + (U_t - U_\lambda)\|_{\text{HS}}^2 - \|U_t - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] \\ =& \mathbb{E}\left[2\left\langle U_{t+1} - U_t, U_t - U_\lambda\right\rangle_{\text{HS}} + \|U_{t+1} - U_t\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right]. \end{aligned} \tag{116}$$

Expanding $U_{t+1} - U_t$ using recursion (25) gives

$$
\begin{aligned}
&\mathbb{E}\left[\langle U_t - U_\lambda, U_{t+1} - U_t\rangle_{\text{HS}} \,|\mathcal{F}_{t-\tau_t}\right] + \mathbb{E}\left[\|U_{t+1} - U_t\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] \\
=&2\mathbb{E}\left[\left\langle U_t - U_\lambda, \eta\left(-\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_t\right) + \frac{E_t}{\eta}\right)\right\rangle_{\text{HS}} |\mathcal{F}_{t-\tau_t}\right] \\
&+ \mathbb{E}\left[\left\|\eta\left(-\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_t\right) + \frac{E_t}{\eta}\right)\right\|_{\text{HS}}^2 |\mathcal{F}_{t-\tau_t}\right] \\
=&2\underbrace{\eta\mathbb{E}\left[\langle U_t - U_\lambda, -\nabla R_\lambda\left(U_t\right)\rangle_{\text{HS}} |\mathcal{F}_{t-\tau_t}\right]}_{T_1} + 2\underbrace{\eta\mathbb{E}\left[\left\langle U_t - U_\lambda, \frac{E_t}{\eta}\right\rangle_{\text{HS}} |\mathcal{F}_{t-\tau_t}\right]}_{T_2} \\
&+2\underbrace{\eta\mathbb{E}\left[\left\langle U_t - U_\lambda, -\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_t\right) + \nabla R_\lambda\left(U_t\right)\right\rangle_{\text{HS}} |\mathcal{F}_{t-\tau_t}\right]}_{T_3} \\
&+\underbrace{\eta^2\mathbb{E}\left[\left\|-\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_t\right) + \frac{E_t}{\eta}\right\|_{\text{HS}}^2 |\mathcal{F}_{t-\tau_t}\right]}_{T_4}
\end{aligned} \tag{117}
$$

In the above decomposition, T_1 corresponds to the negative drift. This term can be easily bounded by applying Lemma 11. T_2 results from the error due to compression and depends on a proper choice of compression budget ε. T_3 is a consequence of Markovian sampling, and if we were to collect IID samples, T_3 equals zero. Thanks to Lemma 5, T_3 can be bounded by invoking the mixing property. Lastly, T_4 collects the error due to the discretization of ODE and compression. It can be controlled under a proper choice of stepsizes and compression budget.

We next provide an upper bound for each term above in five steps.

(Step 1) To bound T_1, we first apply Lemma 11 to obtain

$$
\langle U_t - U_\lambda, -\nabla R_\lambda\left(U_t\right)\rangle_{\text{HS}} \leq -\lambda\left\|U_t - U_\lambda\right\|_{\text{HS}}^2. \tag{118}
$$

Thus, we have

$$
T_1 \leq -\eta\lambda\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right]. \tag{119}
$$

(Step 2) To bound T_2, recall that our compression rule imposes $\|E_t\|_{\text{HS}} \leq \varepsilon$, and we have

$$
\begin{aligned}
T_2 =&\mathbb{E}\left[\langle U_t - U_\lambda, E_t\rangle_{\text{HS}} \,|\mathcal{F}_{t-\tau_t}\right] \\
\leq&\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}} \|E_t\|_{\text{HS}} \,|\mathcal{F}_{t-\tau_t}\right] \\
\leq&\varepsilon\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}} \,|\mathcal{F}_{t-\tau_t}\right]
\end{aligned} \tag{120}
$$

To further bound $\|U_t - U_\lambda\|_{\text{HS}}$, we use triangle inequality and Lemma 10 to obtain

$$
\|U_t - U_\lambda\|_{\text{HS}} \leq \|U_t\|_{\text{HS}} + \|U_\lambda\|_{\text{HS}} = 2K/\lambda. \tag{121}
$$

Thus, we get

$$
T_2 \leq 2\varepsilon K/\lambda. \tag{122}
$$

(Step 3) To bound T_3, we invoke the mixing property, and we rearrange T_3 as follows

$$\begin{aligned}
&\mathbb{E}\left[\left\langle U_t - U_\lambda, -\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_t\right) + \nabla R_\lambda\left(U_t\right)\right\rangle_{\mathrm{HS}} | \mathcal{F}_{t-\tau_t}\right] \\
=&\underbrace{\mathbb{E}\left[\left\langle U_t - U_{t-\tau_t}, -\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_t\right) + \nabla R_\lambda\left(U_t\right)\right\rangle_{\mathrm{HS}} | \mathcal{F}_{t-\tau_t}\right]}_{T_{3,1}} \\
&+\underbrace{\mathbb{E}\left[\left\langle U_{t-\tau_t} - U_\lambda, -\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_{t-\tau_t}\right) + \nabla R_\lambda\left(U_{t-\tau_t}\right)\right\rangle_{\mathrm{HS}} | \mathcal{F}_{t-\tau_t}\right]}_{T_{3,2}} \\
&+\underbrace{\mathbb{E}\left[\left\langle U_{t-\tau_t} - U_\lambda, -\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_t\right) + \widetilde{\nabla} R_\lambda\left(x_t, y_t, U_{t-\tau_t}\right) - \nabla R_\lambda\left(U_{t-\tau_t}\right) + \nabla R_\lambda\left(U_t\right)\right\rangle_{\mathrm{HS}} | \mathcal{F}_{t-\tau_t}\right]}_{T_{3,3}}
\end{aligned} \tag{123}$$

For the first term, notice that we can apply Lemma 11 to bound $\left\|U_t - U_{t-\tau_t}\right\|_{\mathrm{HS}}$ and Lemma 8 to bound the norm of gradients. Specifically, Cauchy-Schwartz inequality gives

$$T_{3,1} \leq \mathbb{E}\left[\left\|U_t - U_{t-\tau_t}\right\|_{\mathrm{HS}} \left\|-\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_t\right) + \nabla R_\lambda\left(U_t\right)\right\|_{\mathrm{HS}} | \mathcal{F}_{t-\tau_t}\right]. \tag{124}$$

We then use Lemma 11 to bound $\left\|U_t - U_{t-\tau_t}\right\|_{\mathrm{HS}}$ as

$$\left\|U_t - U_{t-\tau_t}\right\|_{\mathrm{HS}} \leq 4B\tau_\eta\eta\left(\left\|U_t\right\|_{\mathrm{HS}} + 1\right). \tag{125}$$

On the other hand, triangle inequality gives

$$\left\|-\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_t\right) + \nabla R_\lambda\left(U_t\right)\right\|_{\mathrm{HS}} \leq \left\|-\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_t\right)\right\|_{\mathrm{HS}} + \left\|-\nabla R_\lambda\left(U_t\right)\right\|_{\mathrm{HS}}. \tag{126}$$

We can then use Lemma 8 to bound the gradients and we have

$$\begin{aligned}
T_{3,1} \leq& 8B^2\tau_\eta\eta\mathbb{E}\left[\left(\left\|U_t\right\|_{\mathrm{HS}} + 1\right)^2 | \mathcal{F}_{t-\tau_t}\right] \\
\leq& 8B^2\tau_\eta\eta\mathbb{E}\left[\left(\left\|U_t - U_\lambda\right\|_{\mathrm{HS}} + \left\|U_\lambda\right\|_{\mathrm{HS}} + 1\right)^2 | \mathcal{F}_{t-\tau_t}\right] \\
\leq& 16B^2\tau_\eta\eta\left(\mathbb{E}\left[\left\|U_t - U_\lambda\right\|_{\mathrm{HS}}^2 | \mathcal{F}_{t-\tau_t}\right] + \Xi_\lambda^2\right).
\end{aligned} \tag{127}$$

In order to bound $T_{3,2}$, Cauchy-Schwatz inequality gives

$$\begin{aligned}
T_{3,2} =& \left\langle U_{t-\tau_t} - U_\lambda, \mathbb{E}\left[-\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_{t-\tau_t}\right) | \mathcal{F}_{t-\tau_t}\right] + \nabla R_\lambda\left(U_{t-\tau_t}\right)\right\rangle_{\mathrm{HS}} \\
\leq& \left\|U_{t-\tau_t} - U_\lambda\right\|_{\mathrm{HS}} \left\|\mathbb{E}\left[-\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_{t-\tau_t}\right) | \mathcal{F}_{t-\tau_t}\right] + \nabla R_\lambda\left(U_{t-\tau_t}\right)\right\|_{\mathrm{HS}}.
\end{aligned} \tag{128}$$

Applying Lemma 5 to bound the bias of operator-valued stochastic gradients, we have

$$T_{3,2} \leq 2B_\kappa\eta\left\|U_{t-\tau_t} - U_\lambda\right\|_{\mathrm{HS}}\left(\left\|U_{t-\tau_t}\right\|_{\mathrm{HS}} + 1\right). \tag{129}$$

We next attempt to obtain a bound in terms of $\left\|U_t - U_\lambda\right\|_{\mathrm{HS}}$ as follows. To this end, triangle inequality gives

$$\left\|U_{t-\tau_t} - U_\lambda\right\|_{\mathrm{HS}} \leq \left\|U_t - U_{t-\tau_t}\right\|_{\mathrm{HS}} + \left\|U_t - U_\lambda\right\|_{\mathrm{HS}}. \tag{130}$$

From Lemma 11, we have

$$\|U_t - U_{t-\tau_t}\|_{\mathrm{HS}} \leq 4B_\kappa \tau_\eta \eta \left(\|U_t\|_{\mathrm{HS}} + 1\right) \overset{(a)}{\leq} \|U_t\|_{\mathrm{HS}} + 1, \tag{131}$$

where (a) holds by assumption $\tau_\eta \eta \leq 1/4B$. We next decompose $\|U_t\|_{\mathrm{HS}}$ as $\|U_t - U_\lambda + U_\lambda\|_{\mathrm{HS}}$ and using triangle inequality again to obtain

$$\begin{aligned} \|U_{t-\tau_t} - U_\lambda\|_{\mathrm{HS}} \leq & \|U_\lambda\|_{\mathrm{HS}} + \|U_t - U_\lambda\|_{\mathrm{HS}} + 1 + \|U_t - U_\lambda\|_{\mathrm{HS}} \\ =& 2\|U_t - U_\lambda\|_{\mathrm{HS}} + \|U_\lambda\|_{\mathrm{HS}} + 1. \end{aligned} \tag{132}$$

Likewise, we can bound $\|U_{t-\tau_t}\|_{\mathrm{HS}} + 1$ in terms of $\|U_t - U_\lambda\|$. First notice that

$$\|U_{t-\tau_t}\|_{\mathrm{HS}} + 1 \leq \|U_{t-\tau_t} - U_t\|_{\mathrm{HS}} + \|U_t - U_\lambda\|_{\mathrm{HS}} + \|U_\lambda\|_{\mathrm{HS}} + 1, \tag{133}$$

where the first term can be bounded as in (131). We further write U_t as $U_t - U_\lambda + U_\lambda$ and use triangle inequality to obtain

$$\begin{aligned} \|U_{t-\tau_t}\|_{\mathrm{HS}} + 1 \leq & \|U_t\|_{\mathrm{HS}} + 1 + \|U_t - U_\lambda\|_{\mathrm{HS}} + \|U_\lambda\|_{\mathrm{HS}} + 1 \\ \leq & \left(\|U_t - U_\lambda\|_{\mathrm{HS}} + \|U_\lambda\|_{\mathrm{HS}} + 1\right) + \|U_t - U_\lambda\|_{\mathrm{HS}} + \|U_\lambda\|_{\mathrm{HS}} + 1 \\ =& 2\left(\|U_t - U_\lambda\|_{\mathrm{HS}} + \|U_\lambda\|_{\mathrm{HS}} + 1\right). \end{aligned} \tag{134}$$

Notice that $U_{t-\tau_t}$ is $\mathcal{F}_{t-\tau_t}$-adapted. Substituting (132),(134) into (129) gives

$$\begin{aligned} T_{3,2} \leq & 2B_\kappa \eta \mathbb{E}\left[\|U_{t-\tau_t} - U_\lambda\|_{\mathrm{HS}} \left(\|U_{t-\tau_t}\| + 1\right) | \mathcal{F}_{t-\tau_t}\right] \\ \leq & 2B_\kappa \eta \mathbb{E}\left[4\left(\|U_t - U_\lambda\|_{\mathrm{HS}} + \|U_\lambda\|_{\mathrm{HS}} + 1\right)^2 | \mathcal{F}_{t-\tau_t}\right] \\ =& 8B_\kappa \eta \mathbb{E}\left[\left(\|U_t - U_\lambda\|_{\mathrm{HS}} + \Xi_\lambda\right)^2 | \mathcal{F}_{t-\tau_t}\right] \\ \leq & 16B_\kappa \eta \mathbb{E}\left[\left(\|U_t - U_\lambda\|_{\mathrm{HS}}^2\right) + \Xi_\lambda^2 | \mathcal{F}_{t-\tau_t}\right] \\ =& 16B_\kappa \eta \left(\mathbb{E}\left[\left(\|U_t - U_\lambda\|_{\mathrm{HS}}^2\right) | \mathcal{F}_{t-\tau_t}\right] + \Xi_\lambda^2]\right), \end{aligned} \tag{135}$$

To bound $T_{3,3}$, we leverage Lemma 8 to bound the difference between gradients. To this end, Cauchy-Schwarz inequality and Lemma 8 gives

$$\begin{aligned} T_{3,3} \leq & \mathbb{E}\left[\|U_{t-\tau_t} - U_\lambda\|_{\mathrm{HS}} \left\{\left\|-\widetilde{\nabla} R_\lambda\left(x_t, y_t, U_t\right) + \widetilde{\nabla} R_\lambda\left(x_t, y_t, U_{t-\tau_t}\right)\right\|_{\mathrm{HS}}\right.\right. \\ & \left.\left. + \|-\nabla R_\lambda\left(U_{t-\tau_t}\right) + \nabla R_\lambda\left(U_t\right)\|_{\mathrm{HS}}\right\} | \mathcal{F}_{t-\tau_t}\right] \\ \leq & 2B_\kappa \mathbb{E}\left[\|U_{t-\tau_t} - U_\lambda\|_{\mathrm{HS}} \|U_t - U_{t-\tau_\eta}\|_{\mathrm{HS}} | \mathcal{F}_{t-\tau_t}\right]. \end{aligned} \tag{136}$$

Again, we next aim to bound the preceding term using $\|U_t - U_\lambda\|_{\mathrm{HS}}$. To this end, we apply Lemma 11 to bound $\|U_t - U_{t-\tau_t}\|_{\mathrm{HS}}$ and we have

$$T_{3,3} \leq 8B_\kappa B \tau_\eta \eta \mathbb{E}\left[\|U_{t-\tau_t} - U_\lambda\|_{\mathrm{HS}} \left(\|U_t\|_{\mathrm{HS}} + 1\right) | \mathcal{F}_{t-\tau_t}\right]. \tag{137}$$

To study the first term inside the conditional expectation, we have

$$\begin{aligned} \|U_{t-\tau_t} - U_\lambda\|_{\mathrm{HS}} \leq & \|U_{t-\tau_t} - U_t\|_{\mathrm{HS}} + \|U_t - U_\lambda\|_{\mathrm{HS}} \\ \leq & \|U_t\|_{\mathrm{HS}} + 1 + \|U_t - U_\lambda\|_{\mathrm{HS}}, \end{aligned} \tag{138}$$

where the last line follows from (131). We further bound $\|U_t\|_{\text{HS}}$ using $\|U_t - U_\lambda\|_{\text{HS}}$ as below.

$$\|U_t\|_{\text{HS}} + 1 \leq \|U_t - U_\lambda\|_{\text{HS}} + \|U_\lambda\|_{\text{HS}} + 1. \tag{139}$$

Thus, (138) becomes

$$\|U_{t-\tau_t} - U_\lambda\|_{\text{HS}} \leq 2\|U_t - U_\lambda\|_{\text{HS}} + \|U_\lambda\|_{\text{HS}} + 1. \tag{140}$$

We then have

$$\begin{aligned} T_{3,3} \leq & 8B^2\tau_\eta\eta\mathbb{E}\left[(2\|U_t - U_\lambda\|_{\text{HS}} + \|U_\lambda\|_{\text{HS}} + 1)(\|U_t - U_\lambda\|_{\text{HS}} + \|U_\lambda\|_{\text{HS}} + 1)\,|\mathcal{F}_{t-\tau_t}\right] \\ \leq & 8B^2\tau_\eta\eta\mathbb{E}\left[2(\|U_t - U_\lambda\|_{\text{HS}} + \|U_\lambda\|_{\text{HS}} + 1)(\|U_t - U_\lambda\|_{\text{HS}} + \|U_\lambda\|_{\text{HS}} + 1)\,|\mathcal{F}_{t-\tau_t}\right] \\ = & 16B^2\tau_\eta\eta\mathbb{E}\left[(\|U_t - U_\lambda\|_{\text{HS}} + \|U_\lambda\|_{\text{HS}} + 1)^2\,|\mathcal{F}_{t-\tau_t}\right] \\ \leq & 32B^2\tau_\eta\eta\left(\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2\,|\mathcal{F}_{t-\tau_t}\right] + (\|U_\lambda\|_{\text{HS}} + 1)^2\right). \end{aligned} \tag{141}$$

Combing the previous three bounds we have

$$\begin{aligned} T_3 \leq & \left(16B^2\tau_\eta\eta^2 + 16B_\kappa\eta^2 + 32B^2\tau_\eta\eta^2\right)\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2\,|\mathcal{F}_{t-\tau_t}\right] \\ & + \left(16B^2\tau_\eta\eta^2 + 16B_\kappa\eta^2 + 32B^2\tau_\eta\eta^2\right)\Xi_\lambda^2 \end{aligned} \tag{142}$$

(Step 4) Finally, by (106), we apply affine scaling of gradients in Lemma 8 to bound $\left\|-\widetilde{\nabla} R_\lambda(x_t, y_t, U_t)\right\|_{\text{HS}}$. Together with the bound on compression error E_t, we have

$$\begin{aligned} T_4 \leq & \eta^2\mathbb{E}\left[\left(\left\|-\widetilde{\nabla} R_\lambda(x_t, y_t, U_t)\right\|_{\text{HS}} + \left\|\frac{E_t}{\eta}\right\|_{\text{HS}}\right)^2 |\mathcal{F}_{t-\tau_t}\right] \\ \leq & \eta^2\mathbb{E}\left[\left(\left\|-\widetilde{\nabla} R_\lambda(x_t, y_t, U_t)\right\|_{\text{HS}} + \varepsilon/\eta\right)^2 |\mathcal{F}_{t-\tau_t}\right] \\ \leq & \eta^2\mathbb{E}\left[(B_\kappa(\|U_t\|_{\text{HS}} + 1) + B_{\text{cmp}}(\|U_t\|_{\text{HS}} + 1))^2\,|\mathcal{F}_{t-\tau_t}\right] \\ = & \eta^2\mathbb{E}\left[B^2(\|U_t\|_{\text{HS}} + 1)^2\,|\mathcal{F}_{t-\tau_t}\right], \end{aligned} \tag{143}$$

where $B = B_\kappa + B_{\text{cmp}}$. We further bound the term $\|U_t\|_{\text{HS}}$ via $\|U_t - U_\lambda\|_{\text{HS}}$ as follows.

$$\begin{aligned} T_4 \leq & \eta^2\mathbb{E}\left[B^2(\|U_t - U_\lambda\|_{\text{HS}} + \|U_\lambda\|_{\text{HS}} + 1)^2\,|\mathcal{F}_{t-\tau_t}\right] \\ \leq & 2\eta^2B^2\left(\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2\,|\mathcal{F}_{t-\tau_t}\right] + \Xi_\lambda^2\right). \end{aligned} \tag{144}$$

(Step 5) Combing T_1 to T_4 and merging terms that involve $\|U_t - U_\lambda\|_{\text{HS}}^2$, we have

$$\begin{aligned}
&\mathbb{E}\left[\|U_{t+1} - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] - \mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] \\
\leq& - 2\eta\lambda\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] + 4\varepsilon K/\lambda \\
&+ 2\left(16B^2\tau_\eta\eta^2 + 16B_\kappa\eta^2 + 32B^2\tau_\eta\eta^2\right)\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] \\
&+ 2\left(16B^2\tau_\eta\eta^2 + 16B_\kappa\eta^2 + 32B^2\tau_\eta\eta^2\right)\Xi_\lambda^2 \\
&+ 2\eta^2B^2\left(\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] + \Xi_\lambda^2\right) \\
=&\left(-2\eta\lambda + 32B^2\tau_\eta\eta^2 + 32B_\kappa\eta^2 + 64B^2\tau_\eta\eta^2 + 2\eta^2B^2\right)\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] \\
&+ \left(32B^2\tau_\eta\eta^2 + 32B_\kappa\eta^2 + 64B^2\tau_\eta\eta^2 + 2\eta^2B^2\right)\Xi_\lambda^2 + 4\varepsilon K/\lambda \\
\leq&\left(-2\eta\lambda + \left(32B^2 + 32B + 64B^2 + 2B^2\right)\tau_\eta\eta^2\right)\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] \\
&+ \left(32B^2 + 32B + 64B^2 + 2B^2\right)\tau_\eta\eta^2\Xi_\lambda^2 + 4\varepsilon K/\lambda,
\end{aligned} \tag{145}$$

where the last line holds since $B_\kappa \leq B := B_\kappa + B_{\text{cmp}}$. Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[\|U_{t+1} - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] \leq& \left(1 - 2\eta\lambda + \left(98B^2 + 32B\right)\tau_\eta\eta^2\right)\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] \\
&+ \left(98B^2 + 32B\right)\tau_\eta\eta^2\Xi_\lambda^2 + 4\varepsilon K/\lambda.
\end{aligned} \tag{146}$$

Moreover, under the assumption that $\tau_\eta\eta \leq \lambda/\left(98B^2 + 32B\right)$, we have for $t \geq \tau_t$,

$$\begin{aligned}
\mathbb{E}\left[\|U_{t+1} - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] \leq& \left(1 - \eta\lambda + \left(\left(98B^2 + 32B\right)\tau_\eta\eta^2 - \eta\lambda\right)\right)\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] \\
&+ \left(98B^2 + 32B\right)\tau_\eta\eta^2\Xi_\lambda^2 + 4\varepsilon K/\lambda \\
\leq& \left(1 - \eta\lambda\right)\mathbb{E}\left[\|U_t - U_\lambda\|_{\text{HS}}^2 \,|\mathcal{F}_{t-\tau_t}\right] + \left(98B^2 + 32B\right)\tau_\eta\eta^2\Xi_\lambda^2 \\
&+ 4\varepsilon K/\lambda.
\end{aligned} \tag{147}$$

Taking the expectation on both sides and applying the law of total expectation concludes the proof.

F.3. Proof of Theorem 7

By triangle inequality, we have

$$\begin{aligned}
\mathbb{E}\left\|[\mu_t] - \mu\right\|_\gamma^2 \leq& \mathbb{E}\left(\left\|[\mu_t - \mu_\lambda]\right\|_\gamma + \left\|[\mu_\lambda] - \mu\right\|_\gamma\right)^2 \\
\leq& 2\mathbb{E}\left\|[\mu_t - \mu_\lambda]\right\|_\gamma^2 + 2\mathbb{E}\left\|[\mu_\lambda] - \mu\right\|_\gamma^2 \\
\overset{(a)}{\leq}& 2\lambda^{-(\gamma+1)}\left(K + \lambda\right)^2\mathbb{E}\left\|U_t - U_\lambda\right\|_{\text{HS}}^2 + 2\lambda^{\beta-\gamma}\left\|\mu\right\|_\beta^2,
\end{aligned} \tag{148}$$

where (a) follows from Lemma 9 and (Li et al., 2022, Lemma 1). We next proceed to bound the term $\mathbb{E}[\|U_t - U_\lambda\|_{\text{HS}}^2]$. Define

$$\Theta_0\left(\eta^2, B_{\text{cmp}}, \lambda\right) := \check{B}\tau_\eta\eta^2\Xi_\lambda^2 + 4B_{\text{cmp}}\eta^2K/\lambda.$$

By Lemma 6, for $t \geq \tau_\eta$, we have

$$\begin{aligned}\mathbb{E}\left[\|U_{t+1}-U_\lambda\|_{\text{HS}}^2\right] \leq & (1-\lambda\eta)\,\mathbb{E}\left[\|U_t-U_\lambda\|_{\text{HS}}^2\right]+\Theta_0\left(\eta^2,B_{\text{cmp}},\lambda\right)\\ =&\mathbb{E}\left[\|U_{\tau_\eta}-U_\lambda\|_{\text{HS}}^2\right](1-\lambda\eta)^{t-\tau_\eta}+\sum_{i=\tau_\eta}^{t-1}\Theta_0\left(\eta^2,B_{\text{cmp}},\lambda\right)\left(\Pi_{j=i+1}^{t-1}(1-\lambda\eta)\right)\\ =&\mathbb{E}\left[\|U_{\tau_\eta}-U_\lambda\|_{\text{HS}}^2\right](1-\lambda\eta)^{t-\tau_\eta}+\sum_{i=\tau_\eta}^{t-1}(1-\lambda\eta)^{t-i-1}\,\Theta_0\left(\eta^2,B_{\text{cmp}},\lambda\right).\end{aligned} \tag{149}$$

Using Lemma 11 and the assumption that $U_0=0$ and $\tau_\eta\eta \leq 1/4B$, we have

$$\begin{aligned}\mathbb{E}\left[\|U_{\tau_\eta}-U_\lambda\|_{\text{HS}}^2\right] \leq&\mathbb{E}\left[\left(\|U_{\tau_\eta}\|_{\text{HS}}+\|U_\lambda\|_{\text{HS}}\right)^2\right]\\ \leq&\mathbb{E}\left[\left(\|U_{\tau_\eta}-U_0\|_{\text{HS}}+\|U_0\|_{\text{HS}}+\|U_\lambda\|_{\text{HS}}\right)^2\right]\\ \overset{(a)}{\leq}&\mathbb{E}\left[\left(2B\tau_\eta\eta\left(\|U_0\|_{\text{HS}}+1\right)+\|U_\lambda\|\right)^2\right]\\ \leq&\mathbb{E}\left[\left(4B\tau_\eta\eta+\|U_\lambda\|\right)^2\right]\\ \leq&\Xi_\lambda^2,\end{aligned} \tag{150}$$

where (a) follows from Lemma 11. Therefore, we conclude

$$\mathbb{E}\left[\|U_{t+1}-U_\lambda\|_{\text{HS}}^2\right] \leq\Xi_\lambda^2(1-\lambda\eta)^{t-\tau_\eta}+\sum_{i=\tau_\eta}^{t-1}\Theta_0\left(\eta^2,B_{\text{cmp}},\lambda\right)\left(\Pi_{j=\tau_\eta}^{t-1}(1-\lambda\eta)\right). \tag{151}$$

In addition, we have

$$\begin{aligned}\sum_{i=\tau_\eta}^{t-1}(1-\lambda\eta)^{t-i-1}\,\Theta_0\left(\eta^2,B_{\text{cmp}},\lambda\right)=&\left(\sum_{k=0}^{t-\tau_\eta-1}(1-\lambda\eta)^k\right)\Theta_0\left(\eta^2,B_{\text{cmp}},\lambda\right)\\ \leq&\frac{1}{\lambda\eta}\Theta_0\left(\eta^2,B_{\text{cmp}},\lambda\right).\end{aligned} \tag{152}$$

Therefore, it holds that

$$\mathbb{E}\left[\|U_t-U_\lambda\|_{\text{HS}}^2\right] \leq\Xi_\lambda^2(1-\lambda\eta)^{t-\tau_\eta}+\Theta_0\left(\eta^2,B_{\text{cmp}},\lambda\right)/(\lambda\eta). \tag{153}$$

Substituting this into (148), we have

$$
\begin{aligned}
\mathbb{E}\left\|[\mu_t]-\mu\right\|_\gamma^2 \leq & 2\lambda^{-(\gamma+1)}(K+\lambda)^2\left(\Xi_\lambda^2(1-\lambda\eta)^{t-\tau_\eta}+\frac{\Theta_0\left(\eta^2, B_{\text{cmp}}, \lambda\right)}{\lambda\eta}\right)+2\lambda^{\beta-\gamma}\left\|\mu\right\|_\beta^2 \\
= & 2\lambda^{-(\gamma+1)}(K+\lambda)^2\,\Xi_\lambda^2(1-\lambda\eta)^{t-\tau_\eta} \\
& +2\lambda^{-(\gamma+1)}(K+\lambda)^2\,\frac{\check{B}\tau_\eta\eta^2\Xi_\lambda^2+4B_{\text{cmp}}\eta^2K/\lambda}{\lambda\eta} \\
& +2\lambda^{\beta-\gamma}\left\|\mu\right\|_\beta^2 \\
= & 2\lambda^{-(\gamma+1)}(K+\lambda)^2\,\Xi_\lambda^2(1-\lambda\eta)^{t-\tau_\eta} \\
& +2\lambda^{-(\gamma+2)}(K+\lambda)^2\left(\check{B}\tau_\eta\Xi_\lambda^2+4\lambda^{-1}B_{\text{cmp}}\varepsilon K\right)\eta \\
& +2\lambda^{\beta-\gamma}\left\|\mu\right\|_\beta^2 .
\end{aligned}
\tag{154}
$$

This completes the proof of Theorem 7.

Appendix G. Details of Implementing Compressed Online Learning of the CMEs

In this section, we show that the compression procedure introduced in Section 3.2 can be implemented efficiently using finite-dimensional kernel matrices. Let $d_t=|\mathcal{I}_t|$. After receiving new samples (x_{t+1}, y_{t+1}), define feature matrices and Gram matrices as

$$
\begin{aligned}
&\Phi_{X,t+1}=\left[\phi_X(x_1),\ldots,\phi_X(x_{d_t})\right], \quad \Psi_{Y,t}=\left[\phi_Y(y_1),\ldots,\phi_Y(y_{d_t})\right], \\
&\widetilde{\Phi}_{X,t+1}=\left[\phi_X(x_1),\ldots,\phi_X(x_{d_t}),\phi_X(x_{t+1})\right], \quad \widetilde{\Psi}_{Y,t+1}=\left[\phi_Y(y_1),\ldots,\phi_Y(y_{d_t}),\phi_Y(y_{t+1})\right], \\
&G_{X,t+1}=\Phi_{X,t+1}^\top\Phi_{X,t+1}, \quad G_{Y,t+1}=\Psi_{Y,t+1}^\top\Psi_{Y,t} \\
&\widetilde{G}_{X,t+1}=\widetilde{\Phi}_{X,t+1}^\top\widetilde{\Phi}_{X,t+1}, \quad \widetilde{G}_{Y,t+1}=\widetilde{\Psi}_{Y,t+1}^\top\widetilde{\Psi}_{Y,t+1} \\
&\bar{G}_{X,t+1}=\widetilde{\Phi}_{X,t+1}^\top\Phi_{X,t+1}, \quad \bar{G}_{Y,t+1}=\widetilde{\Psi}_{Y,t+1}^\top\Psi_{Y,t+1}.
\end{aligned}
\tag{155}
$$

In what follows, we omit the index $t+1$ for simplicity. Then we can write the left hand side of the condition (16) in terms of the decision variable $Z\in\mathbb{R}^{d_t\times d_t}$ as

$$
\begin{aligned}
l(Z) := & \left\|\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{I}_t}Z^{ij}\kappa_Y(y_i,\cdot)\otimes\kappa_X(x_j,\cdot)-\sum_{i=1}^{t+1}\sum_{j=1}^{t+1}\widetilde{W}^{ij}\kappa_Y(y_i,\cdot)\otimes\kappa_X(x_j,\cdot)\right\|_{\text{HS}}^2 \\
= & \left\|\Psi_Y Z\Phi_X^\top-\widetilde{\Psi}_Y\widetilde{W}\widetilde{\Phi}_X^\top\right\|_{\text{HS}}^2 \\
\overset{(a)}{=} & \operatorname{Tr}\Big(\Phi_X^\top Z^\top\Psi_Y^\top\Psi_Y Z\Phi_X^\top-\Phi_X^\top Z^\top\Psi_Y^\top\widetilde{\Psi}_Y^\top\widetilde{W}\widetilde{\Phi}_X^\top-\widetilde{\Phi}_X\widetilde{W}^\top\widetilde{\Psi}_Y^\top\Psi_Y Z\Phi_X^\top \\
& +\widetilde{\Phi}_X\widetilde{W}^\top\widetilde{\Psi}_Y^\top\widetilde{\Psi}_Y^\top\widetilde{W}\widetilde{\Phi}_X^\top\Big) \\
\overset{(b)}{=} & \operatorname{Tr}\left(\Phi_X^\top Z^\top\Psi_Y^\top\Psi_Y Z\Phi_X^\top\right)-2\operatorname{Tr}\left(\widetilde{\Phi}_X\widetilde{W}^\top\widetilde{\Psi}_Y^\top\Psi_Y Z\Phi_X^\top\right)+\operatorname{Tr}\left(\widetilde{\Phi}_X\widetilde{W}^\top\widetilde{\Psi}_Y^\top\widetilde{\Psi}_Y\widetilde{W}\widetilde{\Phi}_X^\top\right) \\
= & \operatorname{Tr}\left(\Phi_X^\top Z^\top G_Y Z\Phi_X^\top-2\widetilde{\Phi}_X\widetilde{W}^\top\bar{G}_Y Z\Phi_X^\top+\widetilde{\Phi}_X\widetilde{W}^\top G_Y\widetilde{W}\widetilde{\Phi}_X^\top\right),
\end{aligned}
\tag{156}
$$

where line (a) follows from $\langle A, B\rangle_{\text{HS}} = \text{Tr}(A^\top B)$ for two HS operators A, B, and line (b) follows from $\text{Tr}(AB) = \text{Tr}(BA)$.

Let $\theta = Z\Phi_X^\top$. By a change of variable, $l(Z)$ can then be viewed as a quadratic function in θ. Thus, by completing the square we have

$$\begin{aligned} l(\theta) =& \text{Tr}\left(\Phi_X^\top Z^\top G_Y Z \Phi_X^\top - 2\widetilde{\Phi}_X \widetilde{W}^\top \bar{G}_Y Z \Phi_X^\top + \widetilde{\Phi}_X \widetilde{W}^\top G_Y \widetilde{W} \widetilde{\Phi}_X^\top\right) \\ =& \text{Tr}\left(\theta^\top G_Y Z \Phi_X^\top - 2\widetilde{\Phi}_X \widetilde{W}^\top \bar{G}_Y \theta + \widetilde{\Phi}_X \widetilde{W}^\top G_Y \widetilde{W} \widetilde{\Phi}_X^\top\right) \\ =& \text{Tr}\left[\left(\theta - G_Y^{-1}\bar{G}_Y^\top \widetilde{W}\widetilde{\Phi}_X^\top\right)^\top G_Y \left(\theta - G_Y^{-1}\bar{G}_Y^\top \widetilde{W}\widetilde{\Phi}_X^\top\right) - \widetilde{\Phi}_X \widetilde{W}^\top \bar{G}_Y G_Y^{-1} \bar{G}_Y^\top \widetilde{W} \widetilde{\Phi}_X^\top \right. \\ & \left. + \widetilde{\Phi}_X \widetilde{W}^\top G_Y \widetilde{W} \widetilde{\Phi}_X^\top\right]. \end{aligned} \tag{157}$$

Hence, by minimizing the above function in θ we have

$$\theta_\star = G_Y^{-1}\bar{G}_Y^\top \widetilde{W}\widetilde{\Phi}_X^\top, \tag{158}$$

and the optimal value $l(\theta_\star)$ is

$$\begin{aligned} l(\theta_\star) =& \text{Tr}\left(-\widetilde{\Phi}_X \widetilde{W}^\top \bar{G}_Y G_Y^{-1}\bar{G}_Y^\top \widetilde{W}\widetilde{\Phi}_X^\top + \widetilde{\Phi}_X \widetilde{W}^\top G_Y \widetilde{W}\widetilde{\Phi}_X^\top\right) \\ =& \text{Tr}\left[\widetilde{\Phi}_X \widetilde{W}^\top \left(G_Y - \bar{G}_Y G_Y^{-1} \bar{G}_Y^\top\right) \widetilde{W} \widetilde{\Phi}_X^\top\right] \\ \overset{(a)}{=}& \text{Tr}\left[\widetilde{W}^\top \left(G_Y - \bar{G}_Y G_Y^{-1} \bar{G}_Y^\top\right) \widetilde{W} \widetilde{\Phi}_X^\top \widetilde{\Phi}_X\right] \\ =& \text{Tr}\left[\widetilde{W}^\top \left(G_Y - \bar{G}_Y G_Y^{-1} \bar{G}_Y^\top\right) \widetilde{W} G_X\right], \end{aligned} \tag{159}$$

where line (a) follows from $\text{Tr}(AB) = \text{Tr}(BA)$. In addition, since $\theta = Z\Phi_X^\top$, $Z_\star$ satisfies

$$Z_\star \Phi_X^\top = G_Y^{-1}\bar{G}_Y^\top \widetilde{W}\widetilde{\Phi}_X^\top. \tag{160}$$

To recover $Z_\star$, we first multiply both side by Φ_X to obtain

$$Z_\star G_X = G_Y^{-1}\bar{G}_Y^\top \widetilde{W}\bar{G}_X. \tag{161}$$

Therefore, we have

$$Z_\star = G_Y^{-1}\bar{G}_Y^\top \widetilde{W}\bar{G}_X G_X^{-1}. \tag{162}$$

In summary, the condition (16) reduces to check whether the trace of a finite-dimensional matrix satisfies

$$\text{Tr}\left[\widetilde{W}^\top \left(G_Y - \bar{G}_Y G_Y^{-1}\bar{G}_Y^\top\right) \widetilde{W} G_X\right] \le \varepsilon_t. \tag{163}$$

And the coefficient matrix can be computed as

$$W = Z_\star = G_Y^{-1}\bar{G}_Y^\top \widetilde{W}\bar{G}_X G_X^{-1}. \tag{164}$$

Moreover, to speed up computation, at each time $t \in \mathbb{T}$, the inversion of Gram matrix $G_{Y,t}^{-1}$ can be computed based on $G_{Y,t-1}^{-1}$ using the Woodbury matrix identity Woodbury (1950). We refer readers to (Engel et al., 2004, Section 4) for details.