Online Learning with Markov Sampling[*]

Steve Smale
Toyota Technological Institute at Chicago
1427 East 60th Street, Chicago, IL 60637, USA
E-mail: smale@math.berkeley.edu
Ding-Xuan Zhou
Department of Mathematics, City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, CHINA
E-mail: mazhou@cityu.edu.hk

August 3, 2007

1 Introduction

This paper attempts to give an extension of learning theory to a setting where the assumption of i.i.d. data is weakened. We hypothesize that a sequence of examples (x_t, y_t) in $X \times Y$ for $t = 1, 2, 3, \ldots$ is drawn from a probability distribution ρ_t on $X \times Y$.

The marginal probabilities on X are supposed to converge to a limit probability on X.

Two main examples for this time process are discussed. The first is a stochastic one which in the special case of a finite space X is defined by a stochastic matrix and more generally by a stochastic kernel.

The second is determined by an underlying discrete dynamical system on the space X.

[*] The first author is partially supported by NSF grant 0325113. The second author is supported by the Research Grants Council of Hong Kong [Project No. CityU 103206]. The main results of the paper were announced in Uruguay in December 2005 at the International Conference on Probability and Dynamical Systems. Here discussions with various participants especially Mike Shub were very helpful.

Our theoretical treatment requires that this dynamics be hyperbolic (or "Axiom A") which still permits a class of chaotic systems (with Sinai-Ruelle-Bowen attractors). Even in the case of a limit Dirac point probability, one needs the measure theory to be defined using Hölder spaces.

Many implications of our work remain unexplored. These include for example the relation to Hidden Markov Models, as well as Markov Chain Monte Carlo methods. It seems reasonable that further work consider the push forward of the process from $X \times Y$ by some kind of observable function to a data space.

2 General Setting of Learning in RKHS

Let X be a compact metric space with metric d. Each $x \in X$ is assigned a probability measure ρ_x on $Y = \mathbb{R}$. The regression function is defined as

$$f_\rho(x) = \int_Y y d\rho_x, \qquad x \in X.$$

Our standing hypothesis for $\{\rho_x\}_{x \in X}$ is that for some $M > 0$, $|y| \leq M$ almost surely for each $x \in X$.

We study an online algorithm learning f_ρ in reproducing kernel Hilbert spaces from random samples (not i.i.d) drawn according to a sequence of probability measures.

Let $K : X \times X \to \mathbb{R}$ be a Mercer kernel and $\mathcal{H}_K$ the corresponding Reproducing Kernel Hilbert Space (RKHS) completed by the set of functions $\{K_x = K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ given by $\langle K_x, K_y \rangle_K = K(x, y)$. Denote $\kappa = \sup_{x \in X} \sqrt{K(x, x)}$.

Assume that there is a sequence of probability measures $\{\rho^{(t)}\}_{t=1,2,\cdots}$ on $Z = X \times Y$ such that the conditional distribution of each $\rho^{(t)}$ at $x \in X$ is ρ_x, independent of t.

We consider an online algorithm for learning f_ρ in $\mathcal{H}_K$. It is defined as $f_1 = 0$ and

$$f_{t+1} = f_t - p_t\big((f_t(x_t) - y_t)K_{x_t} + \lambda_t f_t\big), \qquad \text{for } t \in \mathbb{N} \tag{2.1}$$

where $\lambda_t > 0$ is the regularization parameter and $p_t > 0$ is the step size. For each t, $z_t = (x_t, y_t)$ is a random sample drawn according to $\rho^{(t)}$.

The i.i.d. case corresponding to the special choice of $\rho^{(t)} = \rho$ for each t was studied in [2, 9, 19, 28, 24].

Let $\rho_X^{(t)}$ be the marginal distribution of $\rho^{(t)}$ on X. Unlike the i.i.d. case, the convergence depends largely on the sequence $\{\rho_X^{(t)}\}$. The analysis for the algorithm (2.1) will be done in this paper under the assumption that the sequence $\{\rho_X^{(t)}\}$ converges exponentially fast in the dual of the Hölder space $C^s(X)$. Here the Hölder space $C^s(X)$, with $0 \leq s \leq 1$, is defined as the space of all continuous functions on X with the following norm finite:

$$\|f\|_{C^s(X)} = \|f\|_\infty + |f|_{C^s(X)}, \quad \text{where} \quad |f|_{C^s(X)} := \sup_{x \neq y \in X} \frac{|f(x) - f(y)|}{(d(x,y))^s}.$$

It is a Banach space [11] and each probability measure μ on X can be regarded as a bounded linear functional on $C^s(X)$, i.e., $\mu \in (C^s(X))^*$.

Definition 1. *Let $0 \leq s \leq 1$. We say that the sequence $\{\rho_X^{(t)}\}$ converges exponentially fast in $(C^s(X))^*$ to a probability measure ρ_X on X, or* **converges exponentially** *in short, if there exist $C > 0$ and $0 < \alpha < 1$ such that*

$$\|\rho_X^{(t)} - \rho_X\|_{(C^s(X))^*} \leq C\alpha^t, \qquad t \in \mathbb{N}. \tag{2.2}$$

By the definition of the dual space $(C^s(X))^*$, the decay (2.2) can be expressed as

$$\left| \int_X f(x) d\rho_X^{(t)} - \int_X f(x) d\rho_X \right| \leq C\alpha^t \|f\|_{C^s(X)}, \qquad \forall f \in C^s(X), t \in \mathbb{N}. \tag{2.3}$$

We shall verify the above exponential convergence in Sections 4 and 5 for the case when marginal distributions are generated by iterates of a linear operator acting on an initial probability measure.

Let $\mathcal{M}(X)$ be the space of (signed) bounded measures on X. It is the dual of the Banach space $C(X)$ of continuous functions on X with the norm $\|\cdot\|_{C(X)} = \|\cdot\|_\infty$: $\mathcal{M}(X) = C(X)^*$. (But the dual of $\mathcal{M}(X)$ is $L^\infty(X)$, the space of essentially bounded functions on X, not $C(X)$.) For more details, see [11]. Note that $C^0(X)$ and $C(X)$ are equivalent.

Denote $\mathcal{M}_+(X)$ as the subset of $\mathcal{M}(X)$ consisting of positive bounded measures, and $\mathcal{P}(X)$ the set of all probability measures on X. Then

$$\mathcal{M}_+(X) = \{c\mu : \mu \in \mathcal{P}(X), c \geq 0\}, \quad \mathcal{M}(X) = \{\mu_+ - \mu_- : \mu_+, \mu_- \in \mathcal{M}_+(X)\}.$$

A bounded linear operator $\mathcal{A}$ on $\mathcal{M}(X)$ is *positive* if $\mathcal{M}_+(X)$ is invariant under $\mathcal{A}$, that is, $\mathcal{A}\mu \in \mathcal{M}_+(X)$ for any $\mu \in \mathcal{M}_+(X)$. We call $\mathcal{A}$ *stochastic* if $\mathcal{A}$ maps $\mathcal{P}(X)$ into itself. In this special case, we know that $\mathcal{M}^o := \{\mu \in \mathcal{M}(X) : \int_X d\mu = 0\}$, the subspace of $\mathcal{M}(X)$

consisting of measures with total measure zero, is invariant under $\mathcal{A}$ and the restriction $\mathcal{A}|_{\mathcal{M}^o}$ is well defined. In fact, $\mathcal{M}^o$ the annillator in $\mathcal{M}(X) = C(X)^*$ of the constant function with value 1.

We study the sequence of measures $\{\mathcal{A}^t\gamma\}_{t\in\mathbb{N}}$ generated by iterations of a stochastic linear operator $\mathcal{A}$ on $\mathcal{M}(X)$ where $\gamma \in \mathcal{P}(X)$, and consider the exponential convergence (2.2) of the sequence $\{\rho_X^{(t)} = \mathcal{A}^t\gamma\}$ in two settings.

The first setting is when the stochastic linear operator $\mathcal{A}$ is compact. Here we may take $s = 0$, hence the exponential convergence (2.2) is that of measures in the space $\mathcal{M}(X)$ and it implies the exponential convergence for any $0 < s \leq 1$ according to (2.3).

The second setting is when $\mathcal{A}$ is induced by a contracting (or more general) dynamics on X where the Hölder space $C^s(X)$ cannot be replaced by $C(X)$ (i.e., $s > 0$). In both settings, we shall verify the exponential converge (2.2), that is,

$$\|\mathcal{A}^t\gamma - \rho_X\|_{(C^s(X))^*} \leq C\alpha^t, \qquad t \in \mathbb{N}.$$

Here the index α is a characteristic of the stochastic operator $\mathcal{A}$ and is independent of γ. The constant C also depends (largely) on the initial measure γ. They are related to "mixing time" studied in the literature of dynamical systems and the smallest integer $t_0 \in \mathbb{N}$ with $C\alpha^{t_0} \leq 1$ is critical for the bound $C\alpha^t$ to be less than 1.

3 Main Results on Learning Rates

Let $0 \leq s \leq 1$ be a fixed Hölder exponent used in the exponential convergence of measures.

Definition 2. *We say that the Mercer kernel K satisfies the* **kernel condition** *(of order s) if for some constant $\kappa_{2s} > 0$, $K \in C^s(X \times X)$ and for all $u_1, u_2, v_1, v_2 \in X$*

$$|K(u_1, v_1) - K(u_2, v_1) - K(u_1, v_2) + K(u_2, v_2)| \leq \kappa_{2s} \left(d(u_1, u_2)\right)^s \left(d(v_1, v_2)\right)^s. \tag{3.1}$$

When X is a domain of $\mathbb{R}^n$ with smooth boundary and K is C^2, we know from [30] that the kernel condition holds.

We shall take the parameters $\{\lambda_t\}_t$ and $\{p_t\}_t$ for the algorithm (2.1) as

$$\lambda_t = \lambda_1 t^{-\beta}, p_t = p_1 t^{-\theta} \quad \forall t \in \mathbb{N}, \text{where } \lambda_1 > 0, p_1 > 0, 0 < \theta < 1 \text{ and } 0 \leq \beta \leq 1 - \theta. \tag{3.2}$$

Besides K, $\{\lambda_t\}$, $\{p_t\}$, and the measures $\{\rho_x : x \in X\}$ determining the function f_ρ, the algorithm (2.1) involves the sequence of marginal distributions $\{\rho_X^{(t)}\}$ which is a special feature of this online process. Let us mention two examples which satisfy the condition of exponential convergence in Definition 1 to illustrate the main results on learning rates. See more details in Sections 4 and 5 respectively. In both examples, we assume the kernel condition (3.1) for K and the parameter form (3.2) with $\theta = \frac{2}{3}$, $\beta = \frac{1}{3}$ and $p_1\lambda_1 > \frac{1}{6}$. We say that $\nu \in \mathcal{P}(X)$ is strict positivity if $\mu(\Gamma) > 0$ for any nonempty open subset Γ of X.

Example 1. *Let $\nu \in \mathcal{P}(X)$ be strictly positive and $\psi \in C(X \times X)$ be strictly positive satisfying $\int_X \psi(x,u)d\nu(u) = 1$ for each $x \in X$. Define the sequence $\{\rho_X^{(t)}\}$ by*

$$\rho_X^{(t+1)}(\Gamma) = \int_\Gamma \left\{ \int_X \psi(x,u)d\rho_X^{(t)}(x) \right\} d\nu(u), \qquad t \in \mathbb{N}, \text{ and Borel set } \Gamma \subseteq X.$$

Then $\{\rho_X^{(t)}\}$ converges exponentially to some $\rho_X \in \mathcal{P}(X)$. If $f_\rho = \int_X K_v g_\rho(v) d\rho_X(v)$ for some $g_\rho \in L^2_{\rho_X}(X)$, then we have by Theorem 1 below

$$\mathbb{E}_{z_1,\ldots,z_t}\left(\|f_{t+1} - f_\rho\|_K\right) \leq C^* t^{-\frac{1}{6}},$$

where C^* is a constant independently of t.

In the special case of a finite space X, the function ψ above is a positive stochastic matrix (a stochastic density kernel in general).

Example 2. *Let X be a Riemannian manifold and $\mathcal{S} : X \to X$ be a C^2 diffeomorphism with an Axiom A attractor X_0. Let X^* be the basin of attraction of X_0. If $\rho_X^{(1)} \in \mathcal{P}(X^*)$ is absolutely continuous on the transversal to the stable manifolds of the attractor, and the sequence $\{\rho_X^{(t)}\}$ is given by*

$$\rho_X^{(t+1)}(\Gamma) = \rho_X^{(t)}(\mathcal{S}^{-1}\Gamma), \qquad t \in \mathbb{N}, \text{ and Borel set } \Gamma \subseteq X,$$

then $\{\rho_X^{(t)}\}$ converges exponentially to some $\rho_X \in \mathcal{P}(X_0)$. If $f_\rho = \int_{X_0} K_v g_\rho(v) d\rho_X(v)$ for some $g_\rho \in L^2_{\rho_X}(X_0)$, then we have by Theorem 1 below

$$\mathbb{E}_{z_1,\ldots,z_t}\left(\|f_{t+1} - f_\rho\|_K\right) \leq C^* t^{-\frac{1}{6}}.$$

The above learning rates $t^{-\frac{1}{6}}$ can be improved to $t^{-\frac{2r-1}{4r+2}}$ with an exponent $\frac{1}{2} < r \leq \frac{3}{2}$ used in the following regularity condition for f_ρ.

Definition 3. *For a probability measure μ, we define an integral operator $L_{K,\mu} : L^2_\mu \to L^2_\mu$ as*

$$L_{K,\mu} f = \int_X K_v f(v) d\mu(v).$$

It is a compact operator and its power $L^r_{K,\mu}$ is well-defined. The function f_ρ is said to satisfy the **regularity condition** *(of order r) if*

$$f_\rho = L^r_{K,\rho_X}(g_\rho) \text{ for some } g_\rho \in L^2_{\rho_X}(X). \tag{3.3}$$

Now we can state our main result of the paper about learning rates of the online algorithm (2.1) under the assumption of exponential convergence of the marginal distributions. Denote the norm in $L^2_{\rho_X}(X)$ as $\|\cdot\|_{\rho_X}$.

Theorem 1. *Define $\{f_t\}_t$ by (2.1) with parameters (3.2). Assume the exponential convergence (2.2) for $\{\rho_X^{(t)}\}$, the kernel condition (3.1) for K, and the regularity condition (3.3) for f_ρ. The following bounds hold for $t \in \mathbb{N}$,*

$$\begin{aligned}
&\mathbb{E}_{z_1,\ldots,z_t}\left(\|f_{t+1}-f_\rho\|_K\right) \leq \|g_\rho\|_{\rho_X}\lambda_1^{r-\frac{1}{2}} t^{-\beta(r-\frac{1}{2})} \\
&+ \begin{cases}
\left(p_1C_1^* + (C\lambda_1^{r-\frac{3}{2}} + \lambda_1^{r-\frac{1}{2}})C_2^*\right) t^{-\min\{\beta(r-\frac{1}{2}), \frac{\theta-\beta}{2}\}}, & \text{if } 0<\beta<1-\theta, \alpha<1, \\
\left(p_1C_1^* + (C\lambda_1^{r-\frac{3}{2}} + \lambda_1^{r-\frac{1}{2}})C_2^*\right) t^{-\min\{\beta(r-\frac{1}{2}), \theta-\frac{1}{2}, p_1\lambda_1\}} \log(t+1), & \text{if } \beta=1-\theta, \alpha<1, \\
\left(p_1C_1^* + C\lambda_1^{r-\frac{3}{2}} C_2^*\right) t^{-\frac{\theta}{2}}, & \text{if } \beta=0, \alpha<1, \\
p_1C_1^* t^{-\frac{\theta}{2}} + C\lambda_1^{r-\frac{3}{2}} C_2^* t^{\theta}, & \text{if } \beta=0, \alpha=1
\end{cases}
\end{aligned} \tag{3.4}$$

Here C_1^*, C_2^* are constants independent of t or C, λ_1. They depend on $\kappa, \kappa_{2s}, \alpha, \beta, \theta, r, C_K$ and $p_1\lambda_1$ and will be given explicitly in the proof.

Remark 1. *We can see from the proof of Theorem 1 given in Section 6 that the factor* $\log(t+1)$ *can be omitted when $p_1\lambda_1 \neq \beta(r-\frac{1}{2})$ or $\theta - \frac{1}{2}$ in the case $\beta = 1-\theta, \alpha < 1$. Note that the constants C_1^* and C_2^* depend on the product $p_1\lambda_1$, but not λ_1 meaning that they are the same for different pairs (p_1, λ_1) as long as the product $p_1\lambda_1$ is invariant.*

3.1 Optimal learning rate

In contrast to the exponents θ and β, the exponent r in the regularity condition (3.3) needs information on f_ρ and is not involved in the algorithm (2.1). When $r \geq \frac{\theta}{2-2\theta}$, we get the following rate from Theorem 1 and Remark 1.

Corollary 1. *Let $\theta = \frac{2r}{2r+1}$, $\beta = \frac{1}{2r+1}$ and $p_1\lambda_1 > \frac{2r-1}{4r+2}$. Under the kernel condition (3.1), regularity condition (3.3), and the exponential convergence (2.2) with $\alpha < 1$, we have*

$$\mathbb{E}_{z_1,\ldots,z_t}\left(\|f_{t+1}-f_\rho\|_K\right) \leq \left(\|g_\rho\|_{\rho_X}\lambda_1^{r-\frac{1}{2}} + p_1C_1^* + (C\lambda_1^{r-\frac{3}{2}} + \lambda_1^{r-\frac{1}{2}})C_2^*\right) t^{-\frac{2r-1}{4r+2}}. \tag{3.5}$$

In the i.i.d. case, $C = 0$ and the learning rate for $p_1 = \lambda_1 = 1$ can be stated as follows.

Corollary 2. *Let $p_t = t^{-\frac{2r}{2r+1}}$, $\lambda_t = t^{-\frac{1}{2r+1}}$ and $\rho_X^{(t)} \equiv \rho_X \in \mathcal{P}(X)$ for $t \in \mathbb{N}$ in (2.1). If the kernel condition (3.1) and the regularity condition (3.3) hold, then (3.5) is true.*

This result in the i.i.d. case was given by Tarrés and Yao [24]. Their work led us to generalize the form of the regularization parameters from $\lambda_t = \lambda_1$ (for $t = 1, \dots, T$ with T depending on the approximation error) in our original version to $\lambda_t = \lambda_1 t^{-\beta}$ in the current version.

In the special case of $r = \frac{3}{2}$, the learning rate in Corollary 2 is $\mathbb{E}_{z_1,\dots,z_t}(\|f_{t+1} - f_\rho\|_K) = O\big(t^{-\frac{1}{4}}\big)$, the same as those in the literature [22, 24].

In the case when ρ_X is a Dirac measure δ_{x^*} with $x^* \in X$, Corollary 1 seems odd: the locations of the points $\{x_t\}$ used in the learning algorithm (2.1) tend to a single point x^* and one expects to learn the function f_ρ well only near x^*, not on the whole region X. However, the assumption (3.3) takes a special form in this case. In fact, $L_{K,\rho_X}(f) = \int_X K_x f(x) d\rho_X = f(x^*)K_{x^*}$. Hence L_{K,ρ_X} is a rank one operator mapping K_{x^*} to $K(x^*, x^*)K_{x^*}$. The function $g_\rho \in L^2_{\rho_X}$ equals $a_\rho K_{x^*}$ for some $a_\rho \in \mathbb{R}$ since $L^2_{\rho_X}$ has dimension one. We find that $L^r_{K,\rho_X}(g_\rho) = a_\rho (K(x^*, x^*))^r K_{x^*}$. This is the function f_ρ. For this function with only one parameter a_ρ, its value at the single point x^* is sufficient for learning.

3.2 Mixing time

Our approach assumes one has a sample drawn from $\rho^{(t)}$ for each t. In the literature of Markov Chain Monte Carlo methods or geometric random walks [25], many algorithms such as ones for volume computation take only one sample at "mixing time" and repeat this sampling taking process for m times, then the m samples are used in the computation. There the emphasis is to find a good sample from a measure close to an equilibrium. In our setting, it can be expressed in the following where T_t is a mixing time.

Definition 4. *For each t, $T_t \in \mathbb{N}$, and we take a sample $z_t = (x_{T_t}^{(t)}, y_{T_t}^{(t)})$ where $\{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^{T_t}$ is a random sample drawn drawn according to $\{\rho^{(i)}\}_{i=1}^{T_t}$. Then we define $\{f_t\}_{t\in\mathbb{N}}$ by (2.1).*

By a mixing time T_t, we mean that $\rho_X^{(T_t)}$ is close to ρ_X within a threshold. So we can take $\{\rho_X^{(T_t)}\}_{t\in\mathbb{N}}$ as the sequence of measures in Theorem 1, $C = \epsilon$ and get the following error bounds in the case $\alpha = 1$, $\beta = 0$.

Theorem 2. *Define $\{f_t\}_t$ by Definition 4 with parameters $\lambda_t \equiv \lambda_1 > 0$ and $p_t = \frac{1}{\lambda_1}t^{-\theta}$ for some $0 < \theta < 1$. Assume the kernel condition (3.1) and the regularity condition (3.3). If for some $\epsilon > 0$,*

$$\|\rho_X^{(T_t)} - \rho_X\|_{(C^s(X))^*} \le \epsilon, \qquad \forall t \in \mathbb{N}, \tag{3.6}$$

then we have

$$\mathbb{E}_{z_1,\dots,z_t}\left(\|f_{t+1} - f_\rho\|_K\right) \le \|g_\rho\|_{\rho_X}\lambda_1^{r-\frac{1}{2}} + C_1^*\lambda_1^{-1}t^{-\frac{\theta}{2}} + \epsilon\lambda_1^{r-\frac{3}{2}}C_2^*t^{\theta}, \tag{3.7}$$

where C_1^*, C_2^* are constants independent of t, ϵ or λ_1. In particular, if we take $\lambda_1 = \epsilon^{\frac{1}{2+2r}}$ and $\epsilon^{-\frac{1+2r}{(2+2r)\theta}} \le t = t_\epsilon < \epsilon^{-\frac{1+2r}{(2+2r)\theta}} + 1$, we have

$$\mathbb{E}_{z_1,\dots,z_{t_\epsilon}}\left(\|f_{t_\epsilon+1} - f_\rho\|_K\right) \le \left(\|g_\rho\|_{\rho_X} + C_1^* + 2C_2^*\right)\epsilon^{\frac{2r-1}{4+4r}}.$$

Let us compare the above error with that for (2.1) in the special case $r = \frac{3}{2}$ stated in Corollary 1. The rate in Theorem 2 is $O(\epsilon^{\frac{1}{5}})$ with $\epsilon^{-\frac{4}{5\theta}} \le t_\epsilon < \epsilon^{-\frac{4}{5\theta}} + 1$ samples. According to Corollary 1, the same number of samples used in the algorithm (2.1) with $\theta = \frac{2r}{1+2r} = \frac{3}{4}$ would yield a learning rate $O(t_\epsilon^{-\frac{1}{4}}) = O(\epsilon^{\frac{4}{15}})$. So it seems that the convergence of the algorithm in Definition 4, using mixing time samples only, is slower than that of algorithm (2.1) using all samples. Moreover, each of the t_ϵ samples involves a long sampling process.

3.3 Main difficulty in the non i.i.d. case

To demonstrate the essential difference between our study and the i.i.d. setting, we recall that the off-line version of the online algorithm (2.1) is

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}_K}\left\{\frac{1}{T}\sum_{t=1}^{T}(f(x_t) - y_t)^2 + \lambda\|f\|_K^2\right\}.$$

Its noise-free limit in the i.i.d. case with $\rho_X^{(t)} \equiv \mu$ is

$$f_{\lambda,\mu} = \arg\min_{f \in \mathcal{H}_K}\left\{\int_X (f(x) - f_\rho(x))^2 d\mu(x) + \lambda\|f\|_K^2\right\}. \tag{3.8}$$

In our setting, the measures $\{\rho_X^{(t)}\}$ vary and an essential error is caused by the change of these marginal distributions. When the sequence $\{\rho_X^{(t)}\}$ converges exponentially, we can use the following bound, to be proved in Section 6, to estimate this essential error.

Proposition 1. *Let μ and μ' be two probability measures on X, and $0 \leq s \leq 1$. If $f_\rho \in C^s(X)$ and K satisfies the kernel condition (3.1), then the functions defined by (3.8) satisfy*

$$\|f_{\lambda,\mu} - f_{\lambda,\mu'}\|_K \leq \frac{C_K}{\lambda} \|\mu - \mu'\|_{(C^s(X))^*} \|f_{\lambda,\mu'} - f_\rho\|_{C^s(X)}, \tag{3.9}$$

where C_K is a constant depending only on K given by $C_K = \sqrt{\kappa^2 + 2|K|_{C^s(X\times X)} + \kappa_{2s}}$.

In particular, when $\{\rho_X^{(t)}\}$ converges exponentially and the regularity condition (3.3) holds, we shall derive $\|f_{\lambda_t,\rho_X^{(t)}} - f_{\lambda_t,\rho_X}\|_K = O(\alpha^t \lambda_t^{r-3/2})$.

4 Measures Induced by Compact Operators

In this section we consider the convergence of the measure sequence $\{\mathcal{A}^t\gamma\}_{t\in\mathbb{N}}$ when $\mathcal{A}$ is compact. Under some mild conditions, we show the exponential convergence in the space $\mathcal{M}(X) = (C(X))^*$ which implies the exponential convergence in the space $(C^s(X))^*$ according to the expression (2.3) and $\|f\|_{C(X)} \leq \|f\|_{C^s(X)}$. The following spectral theorem can be found in [16].

Spectral Theorem. *Let a linear operator $\mathcal{A}$ on $\mathcal{M}(X)$ be stochastic and compact. Then $\mathcal{M}(X)$ can be decomposed as a direct sum of a closed subspace W and a finite set of finite dimensional closed subspaces $\{\Delta_\lambda\}_{\lambda\in\Lambda}$, all invariant under $\mathcal{A}$, such that the following hold:*

(1) The spectral radius of $\mathcal{A}|_W$ is less than 1. There are $C > 0, 0 < \alpha < 1$ satisfying

$$\|\mathcal{A}^t w\|_{\mathcal{M}(X)} \leq C\alpha^t \|w\|_{\mathcal{M}(X)}, \qquad \forall t \in \mathbb{N}, w \in W. \tag{4.1}$$

(2) The set Λ contains all eigenvalues of $\mathcal{A}$ with modulus 1. It is a finite set containing 1 and each $\lambda \in \Lambda$ satisfies $\lambda^k = 1$ for some $k \in \mathbb{N}$. Moreover, Δ_λ is the eigenspace of $\mathcal{A}$ associated with the eigenvalue $\lambda \in \Lambda$.

Applying the spectral theorem to iterates of a stochastic linear operator yields the following convergence result.

Corollary 3. *Let $\mathcal{A}$ be a stochastic and compact linear operator on $\mathcal{M}(X)$, and $\gamma \in \mathcal{P}(X)$. Set W and $\{\Delta_\lambda\}_{\lambda\in\Lambda}$ as in the spectral theorem. Then the sequence $\{\mathcal{A}^t\gamma\}$ converges in $\mathcal{M}(X)$ as $t \to \infty$ if (and only if) $\gamma \in \Delta_1 + W$. The following statements hold true.*

(1) If $\gamma = \gamma^{(1)} + w^{(1)}$ with $\gamma^{(1)} \in \Delta_1$ and $w^{(1)} \in W$, then $\mathcal{A}^t\gamma \to \gamma^{(1)}$ and

$$\|\mathcal{A}^t\gamma - \gamma^{(1)}\|_{\mathcal{M}(X)} \leq C\alpha^t \|w^{(1)}\|_{\mathcal{M}(X)}, \qquad t \in \mathbb{N}.$$

(2) If 1 *is a simple eigenvalue with eigenvector* $\rho_X \in \mathcal{P}(X)$ *and all the other eigenvalues of* $\mathcal{A}$ *are less than one in modulus, then* $W = \mathcal{M}^o$ *and for any* $\gamma \in \mathcal{P}(X)$,

$$\|\mathcal{A}^t\gamma - \rho_X\|_{\mathcal{M}(X)} \le C\alpha^t\|\gamma - \rho_X\|_{\mathcal{M}(X)}, \qquad t \in \mathbb{N}.$$

4.1 Special setting of a finite input space

Consider the special case of a finite input space $X = \{1, 2, \dots, k\}$. In this setting, stochastic linear operators can be represented by stochastic matrices.

Let $P = [p_{i,j}]_{i,j=1}^k$ be a stochastic matrix, that is,

$$p_{i,j} \ge 0, \qquad \sum_{j=1}^k p_{i,j} = 1, \qquad \forall i, j = 1, \dots, k. \tag{4.2}$$

Then it induces a stochastic linear operator $\mathcal{A}$ on $\mathcal{M}(X)$ as

$$\mathcal{A}(\sum_{i=1}^k \gamma_i\delta_i) = \sum_{j=1}^k \left(\sum_{i=1}^k p_{i,j}\gamma_i\right)\delta_j \tag{4.3}$$

since each measure $\gamma \in \mathcal{M}(X)$ can be written as $\gamma = \sum_{i=1}^k \gamma_i\delta_i$. If we denote the measure γ as a vector $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_k) \in \mathbb{R}^k$, then $\mathcal{A}^t\gamma = (P^t)^T\gamma = (\gamma^T P^t)^T$. To study the convergence of the sequence $\{\mathcal{A}^t\gamma\}_t$, we need a canonical form for the stochastic matrix P.

Definition 5. *A* $k \times k$ *stochastic matrix* P *is called* irreducible *if there exists no permutation matrix* $\mathcal{T}$ *such that* $\mathcal{T}^{-1}P\mathcal{T} = \begin{bmatrix} P_1 & 0 \\ B & P_2 \end{bmatrix}$ *where* P_1, P_2 *are square matrices of size less than* k*. An irreducible matrix* P *of size* $k \ge 2$ *is called* primitive *if* 1 *is the only eigenvalue of modulus* 1.

The primitivity of an irreducible matrix P is equivalent to that the matrix P^p is strictly positive for some integer p or that $\lim_{p\to\infty} P^p$ exists.

The canonical form of a stochastic matrix P by a suitable renumbering of states in X is

$$P = \begin{bmatrix} R_0 & S_1 & \cdots & S_m \\ 0 & P_1 & \cdots & 0 \\ \vdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & P_m \end{bmatrix}, \tag{4.4}$$

where each square matrix P_j with $1 \leq j \leq m$ is irreducible stochastic indexed by a subset X_j of X and the square matrix R_0, indexed by a (possibly empty) subset X_0 of X, has spectral radius less than 1. Assume further that $P_1, \ldots, P_\ell$ are primitive ($0 \leq \ell \leq m$) and $P_{\ell+1}, \ldots, P_m$ are not primitive. Then we know that $\lim_{t\to\infty} R_0^t = 0$ and $\lim_{t\to\infty} P_j^t$ exists if and only if $1 \leq j \leq \ell$. For $1 \leq j \leq \ell$, let V_j be the only eigenvector of P_j^T whose components sum to 1, then $\lim_{t\to\infty} P_j^t$ is the matrix with each row being V_j^T. Moreover, for each $1 \leq j \leq m$, the limit $Q_j := \lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^t P_j^s$ exists and its transpose Q_j^T has an only eigenvector V_j whose components sum to 1. (This vector is the same as the one defined above when $j \leq \ell$.) The limit $\lim_{t\to\infty} S_j^t$ exists with each row being a multiple of V_j^T. So it can be written as $\lim_{t\to\infty} S_j^t = \widetilde{S}_j V_j^T$ where $\widetilde{S}_j \in \mathbb{R}_+^{T_0}$. All these facts are well known and can be found in e.g. [16]. From these we get the following convergence.

Corollary 4. *Let $X = \{1, \ldots, k\}$ and P be a $k \times k$ stochastic matrix taking the canonical form (4.4). Let $\mathcal{A}$ be the stochastic linear operator on $\mathcal{M}(X)$ defined by (4.3) and $\gamma \in \mathcal{P}(X)$. Then the sequence $\{\mathcal{A}^t\gamma\}$ converges in $\mathcal{M}(X)$ as $t \to \infty$ if and only if $\gamma_i = 0$ for $i \in \cup_{j=\ell+1}^m X_j$ and the vector $\gamma|_{X_0}$ is orthogonal to each $\widetilde{S}_j$ with $\ell+1 \leq j \leq m$. In this case, the exponential decay (2.2) holds with $\rho_X \in \mathcal{P}(X)$ supported on $\cup_{j=1}^\ell X_j$ and given by*

$$\rho_X|_{X_j} = \left(\sum_{i\in X_j} \gamma_i + \gamma|_{X_0} \cdot \widetilde{S}_j \right) V_j, \qquad 1 \leq j \leq \ell.$$

4.2 Connection to hidden Markov models

The special setting described in subsection 4.1 is related to the standard hidden Markov model (HMM) which was first applied in speech recognition e.g. [14] and is now widely used in various fields in many different forms e.g. [10]. In the HMM terminology, $X = \{1, \ldots, k\}$ is the state space with k states, P is the transition probability matrix with $p_{i,j}$ being the probability to go from state i to state j, and $\{\rho_x = \rho_x(y) : x \in X\}$ is the emission probabilities characterizing the likelihood of a certain observation $y \in Y$ if the model is in the state x. The sample $\{(x_t, y_t)\}_{t=1}^T$ corresponds to the state sequence $\{x_1, x_2, \ldots, x_T\}$ and the observation sequence $\{y_1, y_2, \ldots, y_T\}$. In our learning process, the state sequence is part of the sampling. But in HMM, it is hidden, only the observation sequence is available. It would be interesting to develop some learning algorithms which use the observation sequence $\{y_1, y_2, \ldots, y_T\}$ only.

4.3 Measures generated by stochastic kernels

Let us turn to a setting generated by stochastic density kernels.

Definition 6. *Let ν be a probability measure on X. A function $\psi \in C(X \times X)$ is called a stochastic density kernel with respect to ν if $\int_X \psi(x,u)d\nu(u) = 1$ for each $x \in X$. The integral operator L_ψ associated with the pair (ψ, ν) is defined on $C(X)$ by*

$$L_\psi f(x) = \int_X \psi(x,u)f(u)d\nu(u), \qquad x \in X, f \in C(X). \tag{4.5}$$

The operator L_ψ is compact. Its dual $\mathcal{A}_\psi$ is defined on $\mathcal{M}(X)$, therefore is also compact. We have for Borel sets Γ of X

$$\mathcal{A}_\psi \mu(\Gamma) = \int_\Gamma \left\{ \int_X \psi(x,u)d\mu(x) \right\} d\nu(u). \tag{4.6}$$

Consider now the sequence $\{\mathcal{A}_\psi^t \gamma\}$. Recall the following definition of ergodicity e.g. [8].

Definition 7. *The operator $\mathcal{A}_\psi$ associated with the pair (ψ, ν) is* **ergodic** *if there exists a strictly positive probability measure ρ_X on X such that $\mathcal{A}_\psi^t \gamma \to \rho_X$ (as $t \to \infty$) for any initial probability measure γ. The convergence $\mathcal{A}_\psi^t \gamma \to \rho_X$ means $\lim_{t\to\infty}(\mathcal{A}_\psi^t \gamma)(\Gamma) = \rho_X(\Gamma)$ for any open set Γ whose boundary has ρ_X-measure zero.*

By [8] (p. 249), the convergence $\mathcal{A}_\psi^t \gamma \to \rho_X$ is equivalent to that $\lim_{t\to\infty} \int_X f(x)d(\mathcal{A}_\psi^t \gamma) = \int_X f(x)d\rho_X$ for any $f \in C(X)$. That is, $\mathcal{A}_\psi^t \gamma$ converges to ρ_X in the weakly* topology.

Remark 2. *If the measure ν is strictly positive, and the kernel function ψ is strictly positive on $X \times X$, then the associated operator $\mathcal{A}_\psi$ is ergodic. See [8].*

Proposition 2. *If the operator $\mathcal{A}_\psi$ associated with the pair (ψ, ν) is ergodic, then there exists a strictly positive probability measure ρ_X on X and some constants $C > 0, 0 < \alpha < 1$ such that*

$$\left| \int_X f(x)d(\mathcal{A}_\psi^t \gamma) - \int_X f(x)d\rho_X \right| \le C\alpha^t \|f\|_{C(X)}, \qquad \forall t \in \mathbb{N}, f \in C(X). \tag{4.7}$$

Hence $\{\mathcal{A}_\psi^t \gamma\}_{t\in\mathbb{N}}$ converges exponentially fast in $(C^s(X))^*$ to ρ_X for every $0 \le s \le 1$.

Proof. By the definition of ergodicity, there is a strictly positive probability measure ρ_X on X such that $\mathcal{A}_\psi^t \gamma \to \rho_X$ for any initial probability measure γ. That is, $\mathcal{A}_\psi^t \gamma$ converges to ρ_X

in the weakly* topology and $\mathcal{A}_\psi \rho_X = \rho_X$. By Corollary 3, we know that $\mathcal{M}(X) = \Delta_1 + W$, and Δ_1 is a one dimensional subspace spanned by ρ_X. So Corollary 3 (2) tells us that $W = \mathcal{M}^o$, and that

$$\|\mathcal{A}_\psi^t \gamma - \rho_X\|_{\mathcal{M}(X)} \leq C\alpha^t \|\gamma - \rho_X\|_{\mathcal{M}(X)}.$$

Since $\mathcal{M}(X)$ is the dual of the Banach space $C(X)$, our conclusion follows from $\|\gamma - \rho_X\|_{\mathcal{M}(X)} \leq 2$ by replacing $2C$ by C. □

5 Measures Induced by Dynamical Systems

We turn to the second setting when the stochastic linear operator $\mathcal{A}$ on $\mathcal{M}(X)$ is induced by a continuous map $\mathcal{S} : X \to X$.

The continuous map $\mathcal{S}$ on X induces a linear operator $\mathcal{S}^\# : C^s(X) \to C^s(X)$ by

$$\mathcal{S}^\#(f)(x) = f(\mathcal{S}x), \qquad f \in C^s(X), x \in X.$$

The dual $\mathcal{A} = \mathcal{A}_\mathcal{S}$ of $\mathcal{S}^\#$ is a linear operator on $(C^s(X))^*$. These two operators are dual satisfying

$$\langle f, \mathcal{A}\rho\rangle = \langle \mathcal{S}^\# f, \rho\rangle, \qquad f \in C^s(X), \rho \in (C^s(X))^*.$$

Note that $\mathcal{M}(X) \subset (C^s(X))^*$. We see that $\mathcal{A}(\mathcal{M}(X)) \subseteq \mathcal{M}(X)$ satisfying

$$\mathcal{A}\rho(B) = \rho(\mathcal{S}^{-1}B)$$

for any Borel set $B \subseteq X$ and Borel measure ρ on X.

One example of dynamical system is the northpole southpole system (e.g. [15]) given on the unit circle by the map $\mathcal{S}$ with some $0 < \epsilon < 1/(2\pi)$ as $\mathcal{S}(\theta) = \theta + \epsilon \sin(2\pi\theta)$ mod 1. It has two fixed points, one attracting $\theta = 1/2$ and the other repelling $\theta = 0$. We can state this example as follows.

Example 3. *Let $X = \{e^{i\pi(x+1/2)}\}$ be parameterized by the parameter $x \in [0,1)$, $0 < \epsilon < 1/(2\pi)$ and the map $\mathcal{S} : X \to X$ be gievn by*

$$\mathcal{S}(x) = x + \epsilon \sin(2\pi x), \qquad x \in [0,1).$$

Note that $\mathcal{S}'(x) = 1 + 2\pi\epsilon \cos(2\pi x) > 0$ and $\mathcal{S}$ has two fixed points: i and $-i$. Consider the fixed point $-i$ corresponding to the parameter value $x^* = \frac{1}{2} \in [0,1)$ and the open subset $X^* = (0,1)$ of $[0,1)$. There holds

$$|\mathcal{S}(x) - x^*| \leq \big(1 - 4\epsilon \cos(\pi|x - x^*|)\big)|x - x^*|.$$

If we choose $t_0 \in \mathbb{N}$ such that $\big(1-4\epsilon\cos(\pi|x-x^*|)\big)^{t_0} \le \frac{2}{3}$, then $|\mathcal{S}^{t_0}(x)-x^*| \le \frac{1}{3}$ and

$$|\mathcal{S}^t(x)-x^*| \le \big(1-4\epsilon\cos(\pi|\mathcal{S}^{t_0}(x)-x^*|)\big)^{t-t_0}|\mathcal{S}^{t_0}(x)-x^*| \le \big(1-2\epsilon\big)^{t-t_0}|x-x^*|$$

for $t > t_0$. Hence with $\alpha_0 = 1-2\epsilon < 1$ we have

$$|\mathcal{S}^t(x)-x^*| \le \widetilde{C}(x)\alpha_0^t|x-x^*|, \qquad \forall x \in X^*, t \in \mathbb{N}. \tag{5.1}$$

Here $\widetilde{C} : X^* \to \mathbb{R}_+$ is a continuous function given from an explicit expression for t_0 as

$$\widetilde{C}(x) = (1-2\epsilon)^{-1-\log\frac{2}{3}/\log(1-4\epsilon\cos(\pi|x-x^*|))}, \qquad x \in X^*.$$

A class of examples is given by contractive maps of complete matric spaces. A map $\mathcal{S} : X \to X$ is (strictly) *contractive* if there exist $0 < \alpha < 1$ and $C_0 > 0$ such that $d(\mathcal{S}^t x, \mathcal{S}^t y) \le C_0\alpha^t d(x,y)$ for any $x, y \in X$ and all $t \in \mathbb{N}$. It follows that the map $\mathcal{S}$ has a unique fixed point $x^* \in X$.

Example 4. *Let $\mathcal{S}$ be a contractive map on X with the unique fixed point x^*. Then for any $\gamma \in \mathcal{P}(X)$ and $0 < s \le 1$, the sequence of measures $\{\mathcal{A}^t\gamma\}$ converges exponentially fast in $(C^s(X))^*$ to the Dirac measure $\rho_X = \delta_{x^*}$.*

Note that if X_0 is an $\mathcal{S}$-invariant compact subset of X, then the linear operator $\mathcal{A}_{\mathcal{S}|_{X_0}} : \mathcal{M}(X_0) \to \mathcal{M}(X_0)$ induced by the restriction of $\mathcal{S}$ onto X_0 is well defined. When X_0 is finite, $\mathcal{A}_{\mathcal{S}|_{X_0}}$ can be represented by a stochastic matrix.

Example 5. *Let $\mathcal{S}$ be a continuous map on the compact metric space X and X^* be an open subset of X. Assume that $X_0 \subset X^*$ is an $\mathcal{S}$-invariant finite subset of X such that $\mathcal{A}_{\mathcal{S}|_{X_0}}$ has a simple eigenvalue 1 with eigenvector $\rho_X \in \mathcal{P}(X_0)$ and all the other eigenvalues are less than one in modulus. Suppose there exists a measurable function $C : X^* \to \mathbb{R}_+$ such that each $u \in X^*$ corresponds to some $x \in X_0$ satisfying*

$$d(\mathcal{S}^t u, \mathcal{S}^t x) \le C(u)\alpha^t, \qquad \forall t \in \mathbb{N}. \tag{5.2}$$

If $\gamma \in \mathcal{P}(X^*)$ satisfies $\int_{X^*} C(u)d\gamma(u) < \infty$, then for any $0 < s \le 1$, the sequence of measures $\{\mathcal{A}^t\gamma\}$ converges exponentially fast in $(C^s(X))^*$ to ρ_X.

Proof. For each $x \in X_0$, define $\mathcal{F}_x$ to be the set of all points $u \in X^*$ such that (5.2) holds. Define a (marginal) measure γ_{X_0} on X_0 by $\gamma_{X_0}(\{x\}) = \gamma(\mathcal{F}_x)$. Then $\gamma_{X_0} \in \mathcal{P}(X_0)$. Moreover, there is a probability (conditional) measure $\gamma(u|x)$ on $\mathcal{F}_x$ such that $\int_{X^*} g(u)d\gamma(u) = \int_{X_0}\int_{\mathcal{F}_x} g(u)d\gamma(u|x)d\gamma_{X_0}(x)$ for any measurable function g on X^*.

Let $f \in C^s(X)$. Since γ vanishes on $X \setminus X^*$, we have

$$\langle f, \mathcal{A}^t \gamma \rangle = \int_X f(u) d\big(\mathcal{A}^t \gamma\big) = \int_X f(\mathcal{S}^t u) d\gamma = \int_{X^*} f(\mathcal{S}^t u) d\gamma.$$

This equals $\int_{X_0} \left\{ \int_{\mathcal{F}_x} f(\mathcal{S}^t u) d\gamma(u|x) \right\} d\gamma_{X_0}(x)$. The definition of the seminorm $\|\cdot\|_{C^s(X)}$ gives

$$\begin{aligned}\left| \langle f, \mathcal{A}^t \gamma \rangle - \int_{X_0} f(\mathcal{S}^t x) d\gamma_{X_0}(x) \right| &= \left| \int_{X_0} \left\{ \int_{\mathcal{F}_x} f(\mathcal{S}^t u) - f(\mathcal{S}^t x) d\gamma(u|x) \right\} d\gamma_{X_0}(x) \right| \\ &\leq \int_{X_0} |f|_{C^s(X)} \left\{ \int_{\mathcal{F}_x} \big(d(\mathcal{S}^t u, \mathcal{S}^t x)\big)^s d\gamma(u|x) \right\} d\gamma_{X_0}(x).\end{aligned}$$

Using the condition (5.2), we see that the above expression can be bounded by

$$|f|_{C^s(X)} \int_{X_0} \left\{ \int_{\mathcal{F}_x} \big(C(u)\alpha^t\big)^s d\gamma(u|x) \right\} d\gamma_{X_0}(x).$$

This in connection with the Hölder inequality tells us that

$$\left| \langle f, \mathcal{A}^t \gamma \rangle - \int_{X_0} f(\mathcal{S}^t x) d\gamma_{X_0}(x) \right| \leq |f|_{C^s(X)} \alpha^{ts} \left(\int_{X^*} C(u) d\gamma(u) \right)^s.$$

Observe that

$$\int_{X_0} f(\mathcal{S}^t x) d\gamma_{X_0} = \int_{X_0} f|_{X_0} \left((\mathcal{S}|_{X_0})^t x \right) d\gamma_{X_0} = \langle f|_{X_0}, (\mathcal{A}_{\mathcal{S}|_{X_0}})^t \gamma_{X_0} \rangle = \left((\mathcal{A}_{\mathcal{S}|_{X_0}})^t \gamma_{X_0} \right) (f|_{X_0}).$$

According to Corollary 3, the eigenvalue condition on $\mathcal{A}_{\mathcal{S}|_{X_0}}$ tells us that there exist some $0 < \alpha < 1$ and $C_0 > 0$ such that

$$\|(\mathcal{A}_{\mathcal{S}|_{X_0}})^t \gamma_{X_0} - \rho_X\|_{(C^s(X_0))^*} \leq C_0 \alpha^t, \qquad \forall t \in \mathbb{N}.$$

Also, $(\mathcal{A}_{\mathcal{S}|_{X_0}})^t \rho_X = \rho_X$ for each $t \in \mathbb{N}$. Therefore,

$$\begin{aligned}&\left| \int_{X_0} f(\mathcal{S}^t x) d\gamma_{X_0} - \int_{X_0} f(x) d\rho_X \right| = \left| \left((\mathcal{A}_{\mathcal{S}|_{X_0}})^t \gamma_{X_0} \right) (f|_{X_0}) - \rho_X(f|_{X_0}) \right| \\ &\leq \left\| (\mathcal{A}_{\mathcal{S}|_{X_0}})^t \gamma_{X_0} - \rho_X \right\|_{(C^s(X_0))^*} \|f|_{X_0}\|_{C^s(X_0)} \leq C_0 \alpha^t \|f\|_{C^s(X)}.\end{aligned}$$

If we denote the trivial extension of ρ_X onto X as ρ_X, then $\int_{X_0} f(x) d\rho_X = \int_X f(x) d\rho_X = \langle f, \rho_X \rangle$. So we have

$$\|\langle f, \mathcal{A}^t \gamma - \rho_X \rangle\| \leq \left\{ C_0 + \left(\int_{X^*} C(u) d\gamma(u) \right)^s \right\} \|f\|_{C^s(X)} \alpha^{ts}.$$

Since f is arbitrary in $C^s(X)$, we have

$$\|\mathcal{A}^t \gamma - \rho_X\|_{(C^s(X))^*} \leq \left\{ C_0 + \left(\int_{X^*} C(u) d\gamma(u) \right)^s \right\} \alpha^{ts}.$$

This proves the exponential convergence. □

Remark 3. *Let $\mathcal{S}$ be a continuous map on X with a fixed point x^*. Let X^* be an open subset of X which contains x^*. Assume there are $\alpha \in (0,1)$ and a measurable function $C : X^* \to \mathbb{R}_+$ such that*

$$d(\mathcal{S}^t x, x^*) \le C(x)\alpha^t, \qquad \forall x \in X^*, t \in \mathbb{N}. \tag{5.3}$$

If $\gamma \in \mathcal{P}(X^*)$ satisfies $\int_{X^*} C(u)d\gamma(u) < \infty$, then for any $0 < s \le 1$, the sequence of measures $\{\mathcal{A}^t\gamma\}$ converges exponentially fast in $(C^s(X))^*$ to δ_{x^*}.

When $s = 0$, the exponential convergence (2.3) does not hold in general. For example when X_0 is an $\mathcal{S}$-invariant subset of X containing the fixed point x^* and $\mathcal{S}|_{X_0}$ is injective, we take $\gamma = \delta_{x_0}$ with $x_0 \ne x^*$, then $\mathcal{A}^t\gamma$ is the Dirac at the point $\mathcal{S}^t x_0 \ne x^*$ and $\|\mathcal{A}^t\gamma - \delta_{x^*}\|_{\mathcal{M}(X_0)} = 2$ for each $t \in \mathbb{N}$.

Finally we turn to a general dynamical system. The following result about the exponential convergence of the sequence $\{\mathcal{A}^t\gamma\}_t$ was forwarded to us by Mike Shub [17]:

"I asked Jasha Pesin the following question.

Let D be a disc in the basin of attraction of an axiom a attractor of a C^2 diffeomorphism. Suppose that D is transversal to the stable manifolds of the attractor. Then is it true that the push forwards of a normalized (to have probability one) smooth volume on D converge to the SRB measure on the attractor? Moreover, is the convergence exponential in the dual space to Hölder functions?

Jasha said yes and you can quote him. On the other hand he nor Viana nor anyoneone else who all thought the statement very reasonable could give a direct reference. Jasha said he thought that the best approximation would be his paper with Sinai entitled something like "Gibbs Measures for partially hyperbolic..."[1]. The Gibbs measure for the log of the unstable Jacobian is the SRB measure of the attractor."

Some discussions on the Sinai-Ruelle-Bowen measure of $\mathcal{S}$ can be found in [29].

6 Error Analysis for Online Learning in RKHS

We proceed to the proof of Theorem 1.

[1] This no doubt is the reference: Ya. B. Pesin and Ya. G. Sinai, Gibbs measures for partially hyperbolic attractors, Ergodic Theory Dynam. Systems **2** (1982), 417-438.

Since the range of $L_{K,\mu}$ is in $\mathcal{H}_K$, it can also be regarded as an operator on $\mathcal{H}_K$ or an operator from L^2_μ to $\mathcal{H}_K$. We shall use the same notion for these operators.

The noise-free limit function $f_{\lambda,\mu}$ defined by (3.8) can be expressed in terms of the integral operator $L_{K,\mu}$ as

$$f_{\lambda,\mu} = \big(L_{K,\mu} + \lambda I\big)^{-1} L_{K,\mu} f_\rho. \tag{6.1}$$

We shall estimate the error $f_{t+1} - f_\rho$ by decomposing into three parts:

$$\begin{aligned} f_{t+1} - f_\rho \quad &= \quad \text{first term} \quad + \quad \text{middle term} \quad + \ \text{last term} \\ &= \Big\{ f_{t+1} - f_{\lambda_t, \rho_X^{(t)}} \Big\} + \Big\{ f_{\lambda_t, \rho_X^{(t)}} - f_{\lambda_t, \rho_X} \Big\} + \Big\{ f_{\lambda_t, \rho_X} - f_\rho \Big\}. \end{aligned} \tag{6.2}$$

The last term above is easy to deal with and will be bounded in subsection 6.1 because the measure ρ_X is fixed. The middle term involves the varying marginal distributions $\{\rho_X^{(t)}\}$ for which the exponential convergence is needed. It can be bounded by Proposition 1 proved in subsection 6.2. The first term is the most difficult part. It involves the change of the regularization parameter λ_t and the marginal distribution $\rho_X^{(t)}$. We shall decompose this term further by (6.10) below and make the detailed analysis in subsection 6.4.

6.1 The regularization error

The last term of (6.2) is incurred by the regularization parameter and called the *approximation error.*

Proposition 3. *If f_ρ satisfies the condition (3.3), then for any $\lambda > 0$ we have*

$$\|f_{\lambda,\rho_X} - f_\rho\|_K \leq \|g_\rho\|_{\rho_X} \lambda^{r-\frac{1}{2}}. \tag{6.3}$$

Proof. Observe that $f_{\lambda,\rho_X} - f_\rho = -\lambda\big(L_{K,\rho_X} + \lambda I\big)^{-1} f_\rho$. Write $f_\rho = L^r_{K,\rho_X}(g_\rho)$ as $L^{r-\frac{1}{2}}_{K,\rho_X} L^{\frac{1}{2}}_{K,\rho_X} g_\rho$ where g_ρ comes from the condition (3.3), and split the power -1 of $L_{K,\rho_X} + \lambda I$ in two factors with powers $r - \frac{3}{2}$ and $\frac{1}{2} - r$. We find the expression

$$f_{\lambda,\rho_X} - f_\rho = -\lambda\big(L_{K,\rho_X} + \lambda I\big)^{r-\frac{3}{2}} \Big\{ \big(L_{K,\rho_X} + \lambda I\big)^{\frac{1}{2}-r} L^{r-\frac{1}{2}}_{K,\rho_X} \Big\} L^{\frac{1}{2}}_{K,\rho_X} g_\rho.$$

This together with the positivity of the operator L_{K,ρ_X} and the norm equality $\|L^{\frac{1}{2}}_{K,\rho_X} g_\rho\|_K = \|g_\rho\|_{\rho_X}$ (see e.g. [4]) implies (6.3). □

6.2 Error caused by measure differences

Recall the reproducing property of $\mathcal{H}_K$:

$$\langle f, K_x \rangle_K = f(x), \qquad f \in \mathcal{H}_K, x \in X. \tag{6.4}$$

It follows that (6.4) implies

$$|f(x)| \le \|f\|_{C(X)} \le \kappa \|f\|_K, \qquad f \in \mathcal{H}_K, x \in X. \tag{6.5}$$

When the condition (3.1) is valid, it was proved in [30] that $\mathcal{H}_K$ is included in $C^s(X)$ with the inclusion bounded:

$$\|f\|_{C^s(X)} \le (\kappa + \kappa_{2s}) \|f\|_K, \qquad \forall f \in \mathcal{H}_K. \tag{6.6}$$

Now we can prove Proposition 1 concerning the error $f_{\lambda,\mu} - f_{\lambda,\mu'}$ caused by the difference of measures. It will provide bounds for the middle term of (6.2).

Proof of Proposition 1. Since $(L_{K,\mu'} + \lambda I) f_{\lambda,\mu'} = L_{K,\mu'} f_\rho$, we have

$$f_{\lambda,\mu} - f_{\lambda,\mu'} = (L_{K,\mu} + \lambda I)^{-1} \left\{ (L_{K,\mu} - L_{K,\mu'}) f_\rho + L_{K,\mu'} f_{\lambda,\mu'} - L_{K,\mu} f_{\lambda,\mu'} \right\}.$$

It follows that

$$f_{\lambda,\mu} - f_{\lambda,\mu'} = (L_{K,\mu} + \lambda I)^{-1} (L_{K,\mu} - L_{K,\mu'})(f_\rho - f_{\lambda,\mu'})$$

and

$$\|f_{\lambda,\mu} - f_{\lambda,\mu'}\|_K \le \frac{1}{\lambda} \|(L_{K,\mu} - L_{K,\mu'})(f_\rho - f_{\lambda,\mu'})\|_K. \tag{6.7}$$

Note that $f_{\lambda,\mu'} \in \mathcal{H}_K \subseteq C^s(X)$. Denote $f = f_\rho - f_{\lambda,\mu'} \in C^s(X)$. We can express the norm square $\|(L_{K,\mu} - L_{K,\mu'}) f\|_K^2$ as

$$\int_X f(u) \left\{ \int_X f(v) K(u,v) d(\mu - \mu')(v) \right\} d(\mu - \mu')(u).$$

Since $\mathcal{M}(X) \subseteq (C^s(X))^*$, according to the definition of the norm in $(C^s(X))^*$, we have

$$\|(L_{K,\mu} - L_{K,\mu'}) f\|_K^2 \le \|\mu - \mu'\|_{(C^s(X))^*} \|g\|_{C^s(X)} = \|\mu - \mu'\|_{(C^s(X))^*} \left(\|g\|_{C(X)} + |g|_{C^s(X)} \right).$$

Here g denotes the function

$$g(u) = f(u) \left\{ \int_X f(v) K(u,v) d(\mu - \mu')(v) \right\}, \qquad u \in X.$$

We need to estimate the norm for g.

First, by the definition of the norm in $(C^s(X))^*$, there holds

$$\|g\|_{C(X)} \leq \|f\|_{C(X)} \sup_{u\in X} \left| \int_X f(v)K(u,v)d(\mu-\mu')(v) \right|$$

Second, consider g as the product of two functions. We have

$$\begin{aligned} |g|_{C^s(X)} \leq & |f|_{C^s(X)} \sup_{u\in X} \left| \int_X f(v)K(u,v)d(\mu-\mu')(v) \right| \\ & + \|f\|_{C(X)} \left| \int_X f(v)K(u,v)d(\mu-\mu')(v) \right|_{C^s(X)} . \end{aligned}$$

For the first term above, we have for any $u \in X$,

$$\begin{aligned} & \left| \int_X f(v)K(u,v)d(\mu-\mu')(v) \right| \leq \|\mu-\mu'\|_{(C^s(X))^*} \|f(\cdot)K(u,\cdot)\|_{C^s(X)} \\ \leq & \|\mu-\mu'\|_{(C^s(X))^*} \left\{ \kappa^2 \|f\|_{C(X)} + |f|_{C^s(X)}\kappa^2 + \|f\|_{C(X)}\kappa_{2s} \right\} . \end{aligned}$$

The second term above $\left| \int_X f(v)K(u,v)d(\mu-\mu')(v) \right|_{C^s(X)}$ equals

$$\sup_{u_1\neq u_2\in X} \left| \int_X f(v) \frac{K(u_1,v)-K(u_2,v)}{(d(u_1,u_2))^s} d(\mu-\mu')(v) \right| .$$

Using the definition of the norm in $(C^s(X))^*$ again, the above quantity is bounded by

$$\sup_{u_1\neq u_2\in X} \|\mu-\mu'\|_{(C^s(X))^*} \left\| f(v) \frac{K(u_1,v)-K(u_2,v)}{(d(u_1,u_2))^s} \right\|_{C^s(X)}$$

while the last $C^s(X)$ norm has an upper bound as

$$\begin{aligned} & \|f\|_{C(X)}|K|_{C^s(X\times X)} + |f|_{C^s(X)}|K|_{C^s(X\times X)} \\ & + \|f\|_{C(X)} \sup_{u_1\neq u_2\in X} \frac{|K(u_1,v_1)-K(u_2,v_1)-K(u_1,v_2)+K(u_2,v_2)|}{(d(u_1,u_2))^s (d(v_1,v_2))^s} \\ \leq & \|f\|_{C^s(X)}|K|_{C^s(X\times X)} + \|f\|_{C(X)}\kappa_{2s} \leq \|f\|_{C^s(X)} \left(|K|_{C^s(X\times X)} + \kappa_{2s} \right) . \end{aligned}$$

Combining all the above analysis, we have

$$\|g\|_{C^s(X)} \leq \|f\|^2_{C^s(X)} \|\mu-\mu'\|_{(C^s(X))^*} \left\{ \kappa^2 + 2|K|_{C^s(X\times X)} + \kappa_{2s} \right\} .$$

It follows that

$$\|(L_{K,\mu}-L_{K,\mu'})f\|^2_K \leq \|f\|^2_{C^s(X)} \|\mu-\mu'\|^2_{(C^s(X))^*} \left\{ \kappa^2 + 2|K|_{C^s(X\times X)} + \kappa_{2s} \right\} .$$

This in connection with (6.7) implies that

$$\|f_{\lambda,\mu} - f_{\lambda,\mu'}\|_K \leq \frac{1}{\lambda}\sqrt{\kappa^2 + 2|K|_{C^s(X\times X)} + \kappa_{2s}}\|\mu - \mu'\|_{(C^s(X))^*}\|f_\rho - f_{\lambda,\mu'}\|_{C^s(X)}.$$

This proves Proposition 1. □

6.3 The error caused by varying regularization parameters

Since the regularization parameter changes with the step, we need a result which was stated by Tarrés and Yao [24] with $\mu = \rho_X$. Their proof yields

Proposition 4. *Let $\rho_X \in \mathcal{P}(X)$ and $\lambda, \lambda' > 0$. If f_ρ satisfies (3.3), then*

$$\|f_{\lambda,\rho_X} - f_{\lambda',\rho_X}\|_K \leq \left|\lambda^{r-\frac{1}{2}} - (\lambda')^{r-\frac{1}{2}}\right| \frac{\|g_\rho\|_{\rho_X}}{r - 1/2}. \tag{6.8}$$

Remark 4. *If we only assume $f_\rho \in \mathcal{H}_K$, then there holds*

$$\|f_{\lambda,\mu} - f_{\lambda',\mu}\|_K \leq \frac{|\lambda - \lambda'|}{\lambda}\|f_\rho\|_K. \tag{6.9}$$

6.4 Proof of the error bounds

Now we turn to the first term of (6.2). It will be studied by iteration. Going from f_t to f_{t+1} involves the change of the function $f_{\lambda_{t-1},\rho_X^{(t-1)}}$ to $f_{\lambda_t,\rho_X^{(t)}}$, hence an error caused by the change of the regularization parameter

$$f_{\lambda_{t-1},\rho_X^{(t-1)}} - f_{\lambda_t,\rho_X^{(t)}} = \left\{f_{\lambda_{t-1},\rho_X^{(t-1)}} - f_{\lambda_{t-1},\rho_X} + f_{\lambda_t,\rho_X} - f_{\lambda_t,\rho_X^{(t)}}\right\} + \left\{f_{\lambda_{t-1},\rho_X} - f_{\lambda_t,\rho_X}\right\}. \tag{6.10}$$

The following bounds for (6.10) and the middle term of (6.2) are obtained by combining (6.6) with Propositions 3, 1 and 4.

Lemma 1. *Let $0 \leq s \leq 1$. Assume (3.2), (2.2) and (3.3). If $K \in C^s(X \times X)$ satisfies (3.1), then*

$$\|f_{\lambda_t,\rho_X^{(t)}} - f_{\lambda_t,\rho_X}\|_K \leq C'C\|g_\rho\|_{\rho_X}\alpha^t t^{\beta(\frac{3}{2}-r)}$$

and

$$\|f_{\lambda_{t-1},\rho_X^{(t-1)}} - f_{\lambda_t,\rho_X^{(t)}}\|_K \leq \begin{cases} 4\|g_\rho\|_{\rho_X}\left\{C'C\alpha^{t-1}t^{\beta(\frac{3}{2}-r)} + \lambda_1^{r-\frac{1}{2}}t^{-\beta(r-\frac{1}{2})-1}\right\}, & \textit{if } \beta > 0, \\ 2\|g_\rho\|_{\rho_X}C'C\alpha^{t-1}, & \textit{if } \beta = 0, \end{cases}$$

where C' is the constant $C' := \frac{C_K(\kappa+\kappa_{2s})}{\lambda_1^{3/2-r}}$.

Remark 5. *If we only assume $f_\rho \in \mathcal{H}_K$, then we have $\|f_{\lambda_t,\mu} - f_{\lambda_{t+1},\mu}\|_K \leq \frac{\beta\|f_\rho\|_K}{t}$.*

We shall use the following elementary inequalities in our proof.

Lemma 2. *(a) For $c, a > 0$, there holds*

$$\exp\{-cx\} \leq \left(\frac{a}{ec}\right)^a x^{-a}, \qquad \forall x > 0, \tag{6.11}$$

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

$$\sum_{i=1}^{t-1} i^{-q_2} \exp\left\{-c \sum_{j=i+1}^{t} j^{-q_1}\right\} \leq \left(\frac{2^{q_1+q_2}}{c} + + \left(\frac{1+q_2}{ec(1-2^{q_1-1})}\right)^{\frac{1+q_2}{1-q_1}}\right) t^{q_1-q_2}. \tag{6.12}$$

For $q_1 = 1$, we have

$$\sum_{i=1}^{t-1} i^{-q_2} \exp\left\{-c \sum_{j=i+1}^{t} j^{-1}\right\} \leq \begin{cases} \frac{2^{q_2}}{|c-q_2+1|} t^{-\min\{c, q_2-1\}}, & \textit{if } c \neq q_2 - 1, \\ 2^{q_2} t^{-c} \log t, & \textit{if } c = q_2 - 1. \end{cases} \tag{6.13}$$

Proof. The first inequality (6.11) is an elementary one and can be found in [19].

(b) First consider the case $q_1 < 1$. In this case, we observe that $\sum_{j=i+1}^{t} j^{-q_1} \geq \int_{i+1}^{t+1} x^{-q_1} dx = \frac{1}{1-q_1}\left((t+1)^{1-q_1} - (i+1)^{1-q_1}\right)$. Then we have

$$\begin{aligned} I_1 &:= \sum_{t/2 \leq i < t} i^{-q_2} \exp\left\{-c \sum_{j=i+1}^{t} j^{-q_1}\right\} \\ &\leq \sum_{t/2 \leq i < t} \left(\frac{t}{2}\right)^{-q_2} \exp\left\{-\frac{c(t+1)^{1-q_1}}{1-q_1}\left(1 - \left(1 - \frac{t-i}{t+1}\right)^{1-q_1}\right)\right\}. \end{aligned}$$

We use an elementary inequality

$$(1-x)^{1-q_1} \leq 1 - (1-q_1)x \qquad \forall 0 \leq x < 1$$

which is proved by considering the function $f(x) = (1-x)^{1-q_1} - 1 + (1-q_1)x$ on $[0,1)$ satisfying $f(0) = 0$ and $f'(x) < 0$ on $(0,1)$. Then we have

$$\begin{aligned} I_1 &\leq \sum_{t/2 \leq i < t} \left(\frac{t}{2}\right)^{-q_2} \exp\left\{-c(t+1)^{1-q_1}\frac{t-i}{t+1}\right\} = \sum_{1 \leq i \leq t/2} \left(\frac{t}{2}\right)^{-q_2} \exp\left\{-c(t+1)^{-q_1} i\right\} \\ &\leq \left(\frac{t}{2}\right)^{-q_2} \int_0^{\frac{t}{2}} \exp\left\{-c(t+1)^{-q_1} x\right\} dx. \end{aligned}$$

Since

$$\int_0^{\frac{t}{2}} \exp\Big\{-c(t+1)^{-q_1}x\Big\}dx = \frac{(t+1)^{q_1}}{c}\left[1-\exp\Big\{-c(t+1)^{-q_1}\frac{t}{2}\Big\}\right],$$

we see that

$$I_1 \leq \frac{2^{q_1+q_2}}{c} t^{q_1-q_2}.$$

Consider the part $i < \frac{t}{2}$. We have $i+1 \leq (t+1)/2$ and $\sum_{j=i+1}^{t} j^{-q_1} \geq \frac{1-2^{q_1-1}}{1-q_1}(t+1)^{1-q_1}$. It follows that

$$I_2 := \sum_{1\leq i<t/2} i^{-q_2}\exp\Big\{-c\sum_{j=i+1}^{t} j^{-q_1}\Big\} \leq \frac{t}{2}\exp\Big\{-\frac{c(1-2^{q_1-1})}{1-q_1}(t+1)^{1-q_1}\Big\}.$$

Applying the inequality in part (a) with $a = \frac{1+q_2}{1-q_1}$ and $x = (t+1)^{1-q_1}$ we know that

$$I_2 \leq \frac{1}{2}\left(\frac{1+q_2}{ec(1-2^{q_1-1})}\right)^{\frac{1+q_2}{1-q_1}} t^{-q_2}.$$

Thus (6.12) is verified.

If $q_1 = 1$, then $\sum_{j=i+1}^{t} j^{-1} \geq \int_{i+1}^{t+1} x^{-1}dx = \log\frac{t+1}{i+1}$. Hence

$$\sum_{i=1}^{t-1} i^{-q_2}\exp\Big\{-c\sum_{j=i+1}^{t} j^{-1}\Big\} \leq \sum_{i=1}^{t-1} i^{-q_2}\left(\frac{i+1}{t+1}\right)^c \leq 2^{q_2}(t+1)^{-c}\sum_{i=1}^{t-1}(i+1)^{c-q_2}.$$

But

$$\sum_{i=1}^{t-1}(i+1)^{c-q_2} \leq \begin{cases} \frac{1}{q_2-c-1}, & \text{if } c-q_2 < -1, \\ \log t, & \text{if } c-q_2 = -1, \\ \frac{(t+1)^{c-q_2+1}}{c-q_2+1}, & \text{if } c-q_2 > -1. \end{cases}$$

Then the inequality (6.13) follows. □

For $x \in X$, we use the sampling operator $S_x : \mathcal{H}_K \to \mathbb{R}$ defined by $S_x f = f(x)$. See [21, 22]. Its adjoint $S_x^T : \mathbb{R} \to \mathcal{H}_K$ is given by $S_x(c) = cK_x$. Then $S_x^T S_x : \mathcal{H}_K \to \mathcal{H}_K$ is given by $S_x^T S_x(f) = f(x)K_x = \langle f, K_x\rangle_K K_x$. It is a rank-one positive operator bounded by κ^2. This is an approximation of the integral operator $L_{K,\rho_X^{(t)}}$ and $\mathbb{E}_{\rho_X^{(t)}}(S_x^T S_x) = L_{K,\rho_X^{(t)}}$.

We are in a position to prove our main result on learning rates of the online algorithm (2.1).

Proof of Theorem 1. Denote the first term of the error decomposition (6.2) as

$$W_{t+1} = f_{t+1} - f_{\lambda_t,\rho_X^{(t)}}.$$

The *first step of the proof* is to establish a simple expression for W_{t+1}, (6.14) below, by iterating a one-step recursion.

In the definition (2.1), we notice that $y_t K_{x_t} = S_{x_t}^T y_t$ and $f_t(x_t)K_{x_t} = S_{x_t}(f_t)K_{x_t} = S_{x_t}^T S_{x_t}(f_t)$. Then we know that

$$\begin{aligned} W_{t+1} &= f_t - f_{\lambda_t,\rho_X^{(t)}} - p_t\Big\{S_{x_t}^T S_{x_t}\big(f_t\big) - S_{x_t}^T y_t + \lambda_t f_t\Big\} \\ &= f_t - f_{\lambda_t,\rho_X^{(t)}} - p_t\Big\{S_{x_t}^T S_{x_t}\big(f_t - f_{\lambda_t,\rho_X^{(t)}}\big) + S_{x_t}^T S_{x_t} f_{\lambda_t,\rho_X^{(t)}} - S_{x_t}^T y_t + \lambda_t f_t\Big\}. \end{aligned}$$

To group in terms of $f_t - f_{\lambda_t,\rho_X^{(t)}}$, we write $\lambda_t f_t$ as $\lambda_t(f_t - f_{\lambda_t,\rho_X^{(t)}}) + \lambda_t f_{\lambda_t,\rho_X^{(t)}}$. The definition (6.1) with the measure $\rho_X^{(t)}$ yields $\lambda_t f_{\lambda_t,\rho_X^{(t)}} = L_{K,\rho_X^{(t)}}(f_\rho - f_{\lambda_t,\rho_X^{(t)}})$. Therefore, we have

$$W_{t+1} = \Big((1-p_t\lambda_t)I - p_t S_{x_t}^T S_{x_t}\Big)\big(f_t - f_{\lambda_t,\rho_X^{(t)}}\big) - p_t\Big\{S_{x_t}^T S_{x_t} f_{\lambda_t,\rho_X^{(t)}} - S_{x_t}^T y_t + L_{K,\rho_X^{(t)}}(f_\rho - f_{\lambda_t,\rho_X^{(t)}})\Big\}.$$

Denote $A_t = (1-p_t\lambda_t)I - p_t S_{x_t}^T S_{x_t}$ and $\chi_t = p_t\Big\{S_{x_t}^T S_{x_t} f_{\lambda_t,\rho_X^{(t)}} - S_{x_t}^T y_t + L_{K,\rho_X^{(t)}}(f_\rho - f_{\lambda_t,\rho_X^{(t)}})\Big\}$. Observe that $W_t = f_t - f_{\lambda_{t-1},\rho_X^{(t-1)}}$ and

$$f_t - f_{\lambda_t,\rho_X^{(t)}} = W_t + \Big\{f_{\lambda_{t-1},\rho_X^{(t-1)}} - f_{\lambda_t,\rho_X^{(t)}}\Big\}.$$

If we denote $f_{\lambda_0,\rho_X^{(0)}} = 0$, then there holds

$$W_{t+1} = A_t W_t + A_t\big(f_{\lambda_{t-1},\rho_X^{(t-1)}} - f_{\lambda_t,\rho_X^{(t)}}\big) - \chi_t, \qquad \forall t \in \mathbb{N}.$$

Denote $\Pi_i = A_t A_{t-1} \dots A_i$ and $\Pi_{t+1} = I$. Since $f_1 = 0$ gives $W_1 = 0$, by iteration we obtain

$$W_{t+1} = \sum_{i=1}^{t} \Pi_i\big(f_{\lambda_{i-1},\rho_X^{(i-1)}} - f_{\lambda_i,\rho_X^{(i)}}\big) - \sum_{i=1}^{t} \Pi_{i+1}\chi_i, \qquad \forall t \in \mathbb{N}. \tag{6.14}$$

The operator $p_i\lambda_i I + p_i S_{x_i}^T S_{x_i}$ is positive and bounded by $(p_i\lambda_i + p_i\kappa^2)I$. So for $i \geq t_0$, the smallest integer greater than $(p_1\lambda_1 + p_1\kappa^2)^{1/\theta}$, the operator $A_i : \mathcal{H}_K \to \mathcal{H}_K$ is positive and bounded by $(1 - p_i\lambda_i)I$, hence $\|A_i\|_{\mathcal{H}_K\to\mathcal{H}_K} \leq 1 - p_i\lambda_i \leq \exp\{-p_i\lambda_i\}$. For $i < t_0$, $\|A_i\|_{\mathcal{H}_K\to\mathcal{H}_K} \leq 1 + p_i\lambda_i + p_i\kappa^2$. It follows that the operator norm of Π_i satisfies

$$\|\Pi_i\|_{\mathcal{H}_K\to\mathcal{H}_K} \leq \widetilde{C}_0 \exp\big\{-p_1\lambda_1 \sum_{j=i}^{t} j^{-\beta-\theta}\big\} \qquad \forall 1 \leq i \leq t, \tag{6.15}$$

where $\widetilde{C}_0$ is the constant given by

$$\widetilde{C}_0 = (1 + p_1\lambda_1 + p_1\kappa^2)^{(p_1\lambda_1 + p_1\kappa^2)^{1/\theta}} \exp\left\{p_1\lambda_1(p_1\lambda_1 + p_1\kappa^2)^{1/\theta}\right\}.$$

The *second step of the proof* is to bound the first term in (6.14). Apply Lemma 1 and (6.15). We find that

$$\begin{aligned}&\left\|\sum_{i=1}^{t} \Pi_i\big(f_{\lambda_{i-1},\rho_X^{(i-1)}} - f_{\lambda_i,\rho_X^{(i)}}\big)\right\|_K \\ &\leq \begin{cases} 4\|g_\rho\|_{\rho_X}\widetilde{C}_0 \sum_{i=1}^t \exp\big\{-p_1\lambda_1 \sum_{j=i}^t j^{-\beta-\theta}\big\}\Big\{C'C\alpha^{i-1} i^{\beta(\frac{3}{2}-r)} + \lambda_1^{r-\frac{1}{2}} i^{-\beta(r-\frac{1}{2})-1}\Big\}, & \text{if } \beta > 0, \\ 2\|g_\rho\|_{\rho_X}\widetilde{C}_0 \sum_{i=1}^t \exp\big\{-p_1\lambda_1 \sum_{j=i}^t j^{-\theta}\big\} C'C\alpha^{i-1}, & \text{if } \beta = 0. \end{cases}\end{aligned}$$

Consider the case $\beta > 0$ and $\alpha < 1$. Because the exponential decay is faster than any polynomial decay, we know that the term with $\alpha^i i^{\beta(\frac{3}{2}-r)}$ is dominated by the polynomial term $i^{-\beta(r-\frac{1}{2})-1}$. In fact, by Lemma 2 (a) with $c = \log(1/\alpha)$ and $a = 2$, we have

$$\alpha^i = \exp\big\{-i\log(1/\alpha)\big\} \leq \left(\frac{2}{e\log(1/\alpha)}\right)^2 i^{-2}. \tag{6.16}$$

So for each $i \in \mathbb{N}$,

$$\alpha^{i-1} i^{\beta(\frac{3}{2}-r)} \leq \left(\frac{4}{e\log(1/\alpha)}\right)^2 i^{-\beta(r-\frac{1}{2})-1}.$$

It follows that

$$\left\|\sum_{i=1}^{t} \Pi_i\big(f_{\lambda_{i-1},\rho_X^{(i-1)}} - f_{\lambda_i,\rho_X^{(i)}}\big)\right\|_K \leq C'' \sum_{i=1}^t i^{-\beta(r-\frac{1}{2})-1} \exp\big\{-p_1\lambda_1 \sum_{j=i}^t j^{-\beta-\theta}\big\},$$

where C'' is the constant

$$C'' = 4\|g_\rho\|_{\rho_X}\widetilde{C}_0\bigg\{C'C\Big(\frac{4}{e\log(1/\alpha)}\Big)^2 + \lambda_1^{r-\frac{1}{2}}\bigg\}.$$

Applying Lemma 2 (b) with $c = p_1\lambda_1$, $q_2 = \beta(r-\frac{1}{2}) + 1$ and $q_1 = \beta + \theta$, we obtain a bound for the first term of (6.14) as

$$\left\|\sum_{i=1}^{t} \Pi_i\big(f_{\lambda_{i-1},\rho_X^{(i-1)}} - f_{\lambda_i,\rho_X^{(i)}}\big)\right\|_K \leq \begin{cases} \widetilde{C}' t^{-\beta(r-\frac{1}{2})-1+\beta+\theta}, & \text{if } \beta+\theta < 1, \\ \widetilde{C}' t^{-\min\{\beta(r-\frac{1}{2}), p_1\lambda_1\}}, & \text{if } \beta+\theta = 1 \text{ and } p_1\lambda_1 \neq \beta(r-\frac{1}{2}), \\ \widetilde{C}' t^{-\beta(r-\frac{1}{2})}\log(t+1), & \text{if } \beta+\theta = 1 \text{ and } p_1\lambda_1 = \beta(r-\frac{1}{2}), \end{cases} \tag{6.17}$$

where $\widetilde{C}'$ is the constant given by $\widetilde{C}' = C''C'''$ with

$$C''' := \begin{cases} \frac{8}{p_1\lambda_1} + 1 + \left(\frac{2+\beta(r-\frac{1}{2})}{ep_1\lambda_1(1-2^{\beta+\theta-1})}\right)^{\frac{2+\beta(r-\frac{1}{2})}{1-\beta-\theta}}, & \text{if } \beta+\theta<1, \\ \frac{4}{|p_1\lambda_1-\beta(r-\frac{1}{2})|} + 1, & \text{if } \beta+\theta=1 \text{ and } p_1\lambda_1 \neq \beta(r-\frac{1}{2}), \\ 5, & \text{if } \beta+\theta=1 \text{ and } p_1\lambda_1 = \beta(r-\frac{1}{2}). \end{cases}$$

The case $\beta = 0$ is easier. We apply (6.16) when $\alpha < 1$. Lemma 2 (b) with $c = p_1\lambda_1$ and $q_1 = \theta$ yields

$$\left\| \sum_{i=1}^{t} \Pi_i \big(f_{\lambda_{i-1},\rho_X^{(i-1)}} - f_{\lambda_i,\rho_X^{(i)}} \big) \right\|_K \leq \begin{cases} \widetilde{C}' t^{-1}, & \text{if } \beta=0, \alpha<1, \\ \widetilde{C}' t^{\theta}, & \text{if } \beta=0, \alpha=1, \end{cases}$$

where the constant $\widetilde{C}'$ is given by $\widetilde{C}' = 2\|g_\rho\|_{\rho_X} \widetilde{C}_0 C' C C'''$ with

$$C''' := \begin{cases} \left(\frac{4}{e\log(1/\alpha)}\right)^2 \left\{ \frac{8}{p_1\lambda_1} + 1 + \left(\frac{3}{ep_1\lambda_1(1-2^{\theta-1})}\right)^{\frac{3}{1-\theta}} \right\}, & \text{if } \beta=0, \alpha<1, \\ \frac{2}{p_1\lambda_1} + 1 + \left(\frac{1}{ep_1\lambda_1(1-2^{\theta-1})}\right)^{\frac{1}{1-\theta}}, & \text{if } \beta=0, \alpha=1. \end{cases}$$

The *third step of the proof* is to estimate the second term of (6.14). Write

$$\left\| \sum_{i=1}^{t} \Pi_{i+1}\chi_i \right\|_K^2 = \sum_{i=1}^{t} \sum_{\ell=1}^{t} \langle \Pi_{i+1}\chi_i, \Pi_{\ell+1}\chi_\ell \rangle_K.$$

Observe that $\mathbb{E}_{z_i}(S_{x_i}^T y_i) = \mathbb{E}_{z_i}(y_i K_{x_i}) = L_{K,\rho_X^{(i)}} f_\rho$, and χ_i depends only on z_i while Π_{i+1} depends only on $z_t, z_{t-1}, \ldots, z_{i+1}$. So $\mathbb{E}_{z_i|z_t,z_{t-1},\ldots,z_{i+1}}\big(\Pi_{i+1}\chi_i\big) = 0$. It follows that for $\ell > i$, the expected value $\mathbb{E}_{z_1,z_2,\ldots,z_t}\left(\langle \Pi_{i+1}\chi_i, \Pi_{i+1}\chi_\ell \rangle_K\right)$ equals

$$\mathbb{E}_{z_t,z_{t-1},\ldots,z_{i+1}} \langle \mathbb{E}_{z_i|z_t,z_{t-1},\ldots,z_{i+1}} \left(\Pi_{i+1}\chi_i\right), \Pi_{\ell+1}\chi_\ell \rangle_K = 0.$$

Hence

$$\mathbb{E}_{z_1,z_2,\ldots,z_t} \left(\left\| \sum_{i=1}^{t} \Pi_{i+1}\chi_i \right\|_K^2 \right) = \sum_{i=1}^{t} \mathbb{E}_{z_1,z_2,\ldots,z_t} \left(\|\Pi_{i+1}\chi_i\|_K^2 \right).$$

It follows from (6.15) that

$$\mathbb{E}_{z_1,z_2,\ldots,z_t} \left(\left\| \sum_{i=1}^{t} \Pi_{i+1}\chi_i \right\|_K^2 \right) \leq \sum_{i=1}^{t} \widetilde{C}_0^2 \exp\Big\{-2p_1\lambda_1 \sum_{j=i+1}^{t} j^{-\beta-\theta}\Big\} \mathbb{E}_{z_i}\left(\|\chi_i\|_K^2 \right).$$

Since $\chi_i = p_i \left\{ (f_{\lambda_i,\rho_X^{(i)}}(x_i) - y_i)K_{x_i} + L_{K,\rho_X^{(i)}}(f_\rho - f_{\lambda_i,\rho_X^{(i)}}) \right\}$, we see that

$$\|\chi_i\|_K^2 \le 2p_i^2\kappa^2 \left\{ (f_{\lambda_i,\rho_X^{(i)}}(x_i) - y_i)^2 + \|f_\rho - f_{\lambda_i,\rho_X^{(i)}}\|^2_{\rho_X^{(i)}} \right\}.$$

Then

$$\mathbb{E}_{z_i}\left(\|\chi_i\|_K^2\right) \le 4p_i^2\kappa^2 \left\{ \|f_\rho - f_{\lambda_i,\rho_X^{(i)}}\|^2_{\rho_X^{(i)}} + M^2 \right\}.$$

To bound the norm, we take $f = 0$ in (3.8) with $\lambda = \lambda_i$ and $\mu = \rho_X^{(i)}$, and find that

$$\|f_{\lambda_i,\rho_X^{(i)}} - f_\rho\|^2_{\rho_X^{(i)}} + \lambda_i \|f_{\lambda,\rho_X^{(i)}}\|_K^2 \le \|f_\rho\|^2_{\rho_X^{(i)}} \le M^2.$$

Hence $\|f_\rho - f_{\lambda_i,\rho_X^{(i)}}\|^2_{\rho_X^{(i)}} \le M^2$ and $\mathbb{E}_{z_i}\left(\|\chi_i\|_K^2\right) \le 8p_i^2\kappa^2M^2$. Therefore,

$$\mathbb{E}_{z_1,z_2,\ldots,z_t}\left(\left\| \sum_{i=1}^t \Pi_{i+1}\chi_i \right\|_K^2 \right) \le 8p_1^2\kappa^2M^2\widetilde{C}_0^2 \sum_{i=1}^t i^{-2\theta} \exp\left\{ -2p_1\lambda_1 \sum_{j=i+1}^t j^{-\beta-\theta} \right\}.$$

Applying Lemma 2 (b) with $c = 2p_1\lambda_1$, $q_2 = 2\theta$ and $q_1 = \beta+\theta$ and the Schwarz inequality, we know that

$$\mathbb{E}_{z_1,z_2,\ldots,z_t}\left(\left\| \sum_{i=1}^t \Pi_{i+1}\chi_i \right\|_K \right) \le \begin{cases} 3p_1\kappa M\widetilde{C}_0\widetilde{C}'' t^{\frac{\beta-\theta}{2}}, & \text{if } \beta+\theta < 1, \\ 3p_1\kappa M\widetilde{C}_0\widetilde{C}'' t^{-\min\{\theta-\frac{1}{2},p_1\lambda_1\}}, & \text{if } \beta+\theta = 1, p_1\lambda_1 \ne \theta - \frac{1}{2}, \\ 3p_1\kappa M\widetilde{C}_0\widetilde{C}'' t^{\frac{1}{2}-\theta}\sqrt{\log(t+1)}, & \text{if } \beta+\theta = 1, p_1\lambda_1 = \theta - \frac{1}{2}, \end{cases}$$

where $\widetilde{C}''$ is the constant given by

$$\widetilde{C}'' = \begin{cases} \frac{2}{\sqrt{p_1\lambda_1}} + 1 + \left(\frac{2}{ep_1\lambda_1(1-2^{\beta+\theta-1})}\right)^{\frac{2}{1-\beta-\theta}}, & \text{if } \beta+\theta < 1, \\ \frac{2}{\sqrt{|2p_1\lambda_1 - 2\theta + 1|}} + 1, & \text{if } \beta+\theta = 1, p_1\lambda_1 \ne \theta - \frac{1}{2}, \\ 3, & \text{if } \beta+\theta = 1, p_1\lambda_1 = \theta - \frac{1}{2}, \end{cases}$$

This in connection with (6.17) provides a bound for the first term of (6.2).

The *last step of the proof* is to estimate the total error $\|f_{t+1} - f_\rho\|_K$ by applying the triangle inequality to the error decomposition (6.2). The first term of (6.2) is estimated in Proposition 3 as

$$\|f_{\lambda_t,\rho_X} - f_\rho\|_K \le \|g_\rho\|_{\rho_X}\lambda_1^{r-\frac{1}{2}} t^{-\beta(r-\frac{1}{2})}$$

while the middle term is bounded in Lemma 1 as

$$\|f_{\lambda_t,\rho_X^{(t)}} - f_{\lambda_t,\rho_X}\| \le C'C\|g_\rho\|_{\rho_X}\alpha^t t^{\beta(\frac{3}{2}-r)}.$$

Note that when $\alpha < 1$, we have $\alpha^t = \exp\{-\log\frac{1}{\alpha}t\} \leq \frac{1}{e\log\frac{1}{\alpha}}t^{-1}$ by Lemma 2 (a) with $a = 1$ and $c = \log\frac{1}{\alpha}$. Adding bounds for the three terms verifies the error estimate (3.4) with the constants C_1^*, C_2^* given explicitly by

$$\begin{aligned} C_1^* &= 3\kappa M\widetilde{C}_0\widetilde{C}'', \\ C_2^* &= \|g_\rho\|_{\rho_X}\begin{cases} \frac{C_K(\kappa+\kappa_{2s})}{e\log\frac{1}{\alpha}} + 4C'''\widetilde{C}_0\Big\{C_K(\kappa+\kappa_{2s})\Big(\frac{4}{e\log(1/\alpha)}\Big)^2+1\Big\}, & \text{if } \beta > 0, \alpha < 1, \\ \frac{C_K(\kappa+\kappa_{2s})}{e\log\frac{1}{\alpha}} + 2C'''\widetilde{C}_0C_K(\kappa+\kappa_{2s}), & \text{if } \beta = 0, \alpha < 1, \\ C_K(\kappa+\kappa_{2s})\left(1+2C'''\widetilde{C}_0\right), & \text{if } \beta = 0, \alpha = 1. \end{cases} \end{aligned}$$

The proof of Theorem 1 is completed.

References

[1] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. **68** (1950), 337–404.

[2] N. Cesa-Bianchi, P. Long and M. Warmuth, Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. IEEE Trans. Neural Networks **7** (1996), 604–619.

[3] F. Cucker and S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. **39** (2001), 1–49.

[4] F. Cucker and D. X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, 2007.

[5] E. De Vito, A. Caponnetto, and L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, Found. Comput. Math. **5** (2005), 59–85.

[6] L. Devroye, L. Györfi and G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer-Verlag, New York, 1997.

[7] T. Evgeniou, M. Pontil and T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. **13** (2000), 1–50.

[8] W. Feller, An Introduction to Probability Theory and Its Applications, Vol. II, 2nd Ed., John Wiley & Sons, 1971.

[9] J. Kivinen, A. J. Smola and R. C. Williamson, Online learning with kernels, IEEE Trans. Signal Processing **52** (2004), 2165–2176.

[10] T. Koski, Hidden Markov Models for Bioinformatics, Kluwer, Dordrecht, 2001.

[11] P. D. Lax, Functional Analysis, John Wiley & Sons, Now York, 2002.

[12] P. Niyogi and F. Girosi, On the relationships between generalization error, hypothesis complexity and sample complexity for radial basis functions, Neural Comp. **8** (1996), 819–842.

[13] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, Ann. Probab. **22** (1994), 1679–1706.

[14] L. R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proc. IEEE **77** (1989), 257–286.

[15] C. Robinson, Dynamical Systems: Stability, Symbolic Dynamics, and Chaos, CRC Press, 1999.

[16] H. H. Schaefer, Banach Lattices and Positive Operators, Springer-Verlag, Berlin/Heidelberg/New York, 1974.

[17] M. Shub, Personal comminucation, 2006.

[18] S. Smale, Differentiable dynamical systems, Bull. Amer. Math. Soc. **73** (1967), 747–817.

[19] S. Smale and Y. Yao, Online learning algorithms, *Found. Comp. Math.* **6** (2006), 145–170.

[20] S. Smale and D. X. Zhou, Estimating the approximation error in learning theory, Anal. Appl. **1** (2003), 17–41.

[21] S. Smale and D. X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Amer. Math. Soc. **41** (2004) 279–305.

[22] S. Smale and D. X. Zhou, Shannon sampling II: Connection to learning theory, Appl. Comput. Harmonic Anal. **19** (2005), 285–302.

[23] S. Smale and D. X. Zhou, Learning theory estimates via integral operators and their applications, Constr. Approx. (2007) DOI: 10.1007/s00365-006-0659-y

[24] P. Tarrés and Y. Yao, Online learning as stochastic approximations of regularization paths, preprint, 2006.

[25] S. Vempala, Geometric random walks: a survey, in Combinatorial and Computational Geometry, Math. Sci. Res. Inst. Publ., **52**, Cambridge Univ. Press, Cambridge, 2005, pp. 577–616,

[26] Y. Yao, On complexity issue of online learning algorithms, IEEE Trans. Inform. Theory, to appear.

[27] G. B. Ye and D. X. Zhou, Fully online classification by regularization, Appl. Comput. Harmonic Anal. (2007), DOI: 10.1016/j.acha.2006.12.001

[28] Y. Ying and D. X. Zhou, Online regularized classification algorithms, IEEE Trans. Inform. Theory **52** (2006), 4775–4788.

[29] L. S. Young, What are SRB measures, and which dynamical systems have them? J. Statist. Phys. **108** (2002), 733–754.

[30] D. X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory **49** (2003), 1743–1752.