Chapter 3: Early Stopping Score Matching Density Estimator

Contents

In this chapter, we introduce the early stopping SM density estimator and discuss its properties. As we have discussed in **Chapter 2**, minimizing $\widehat{J}_{\mathrm{SM}}$ over $\mathcal{H}$ has no solution and a certain kind of regularization has to be imposed. The approach taken by Sriperumbudur et al. (2017) was to add a penalty functional and minimize the penalized SM loss functional and yields the penalized SM density estimator. In this chapter, we consider the early stopping approach to regularize. More precisely, we apply the gradient descent algorithm to minimizing $\widehat{J}_{\mathrm{SM}}$ and terminate the algorithm early, leading to the early stopping SM density estimator.

After an overview of the early stopping regularization and the gradient descent algorithm in Section 3.1, we present our early stopping SM density estimator in Section 3.2. We then study its theoretical properties in Section 3.3. Finally, we compare our early stopping SM density estimator and the penalized SM density estimator proposed by Sriperumbudur et al. (2017) in Section 3.4.

3.1 An Overview

Early stopping is a form of regularization based on choosing when to terminate an iterative optimization algorithm. This form of regularization is often referred to as implicit regularization, in contrast to the penalized approach by explicitly adding a penalty term. The main advantage of early stopping regularization, compared with the penalized approach, is the lower computational complexity.

In the supervised learning setting where a model is estimated via minimizing a loss function, using early stopping can effectively avoid overfitting so that the estimated

model can generalize well. It is essentially a bias-variance tradeoff: stopping the algorithm too early results in a model with a large bias and a low variance, but stopping it too late leads to a model with a low bias but a large variance. Stopping the algorithm early (but not too early) is an approach to solve this bias-variance tradeoff. Early stopping has been systematically investigated in L_2 boosting algorithm (Bühlmann and Yu, 2003), boosting algorithm for a general convex loss function (Zhang and Yu, 2005), and the nonparametric least squares regression in the RKHS setting (Yao, Rosasco, and Caponnetto, 2007; Raskutti, Wainwright, and Yu, 2014), to name a few.

In our development below, we choose the iterative optimization algorithm to be the gradient descent algorithm, which is a first-order optimization algorithm for solving an unconstrained minimization problem. Starting from a point in the feasible set, at each iteration, gradient descent algorithm goes along the direction of the negative gradient of the objective function at the current point. This direction is the one with the steepest descent.

Mathematically, let $g : \mathbb{R}^m \to \mathbb{R}$ be a convex differentiable function and assume there exists a unique $x^* \in \mathbb{R}^m$ such that $\inf_{x\in\mathbb{R}^m} g(x) = g(x^*) > -\infty$. We would like to minimize g over $\mathbb{R}^m$. Gradient descent algorithm starts from $x^{(0)} \in \mathbb{R}^m$ and generates the sequence

$$x^{(t+1)} = x^{(t)} - \tau_t \nabla g(x^{(t)}), \qquad \text{for all } t \in \mathbb{N}_0 := \mathbb{N} \cup \{0\},$$

where $\tau_t > 0$ is the appropriately chosen step size and the subscript "t" indicates

that the step size may differ for different numbers of iterations. With appropriately chosen τ_t, the gradient descent algorithm guarantees g is decreased at each iteration, i.e., $g(x^{(t+1)}) < g(x^{(t)})$ for all $t \in \mathbb{N}_0$, except when g has already achieved the minimum at $x^{(t)}$. One terminates the algorithm when one of the following three criteria does not exceed a pre-specified tolerant parameter

$$\|\nabla g(x^{(t)})\|, \qquad |g(x^{(t+1)}) - g(x^{(t)})|, \qquad \text{and} \qquad \left|\frac{g(x^{(t+1)}) - g(x^{(t)})}{g(x^{(t)})}\right|,$$

where $\|\cdot\|$ is a norm over $\mathbb{R}^m$ chosen by the user. It can be shown that, with appropriately chosen $\{\tau_t\}_{t\in\mathbb{N}}$, the gradient descent algorithm is guaranteed to converge to $g(x^*)$ (see, for example, Chapter 9 in Boyd and Vandenberghe, 2004).

3.2 Early Stopping SM Density Estimator

We present our early stopping SM density estimator in this section.

To start with, we formally state a set of assumptions, in addition to **(A1)** - **(A4)** in **Chapter 2**, that will be used in this section.

(B1) $\mathcal{X}$ is a non-empty open subset of $\mathbb{R}^d$ with a piecewise smooth boundary $\partial\mathcal{X} := \overline{\mathcal{X}}\backslash\mathcal{X}$, where $\overline{\mathcal{X}}$ denotes the closure of $\mathcal{X}$.

(B2) p_0 is continuously differentiable and is continuously extendible to $\overline{\mathcal{X}}$ and satisfies $\int_{\mathcal{X}} p_0(x)\|\nabla \log p_0(x)\|_2^2 \mathrm{d}x < \infty$.

(B3) k is twice continuously differentiable on $\mathcal{X} \times \mathcal{X}$ with continuous extension of $\partial_u\partial_{u+d}k$ and $\partial_u\partial_v\partial_{u+d}\partial_{v+d}k$ to $\overline{\mathcal{X}} \times \overline{\mathcal{X}}$ for all $u, v = 1, 2, \cdots, d$.

(B4) As $x \to \partial\mathcal{X}$, $\sup_{u=1,\cdots,d} \partial_u \partial_{u+d} k(x,x) p_0(x) \to 0$.

(B5) $\kappa_2 := \sup_{u=1,\cdots,d} \sup_{x\in\mathcal{X}} \sqrt{\partial_u \partial_{u+d} k(x,x)} < \infty$,

$\kappa_3 := \sup_{u=1,\cdots,d} \sup_{x\in\mathcal{X}} \sqrt{\partial_u^2 \partial_{u+d}^2 k(x,x)} < \infty$, and

$\kappa_4 := \sup_{u=1,\cdots,d} \sup_{x\in\mathcal{X}} |\partial_u \log \mu(x)| \sqrt{\partial_u \partial_{u+d} k(x,x)} < \infty$.

Under these assumptions, the H-divergence and the SM matching loss take on the forms we have presented in **Chapter 2**, as the following theorem states.

Theorem 1 (Theorem 4 in Sriperumbudur et al. (2017))**.**

*(a) Under **(A1)** - **(A4)** in **Chapter 2** and **(B1)** - **(B5)** above, the H-divergence between p_0 and $q_f \in \mathcal{Q}_{\text{kef}}$ is*

$$J_{\text{SM}}(f) := \frac{1}{2}\langle f, Cf\rangle_{\mathcal{H}} - \langle f, z\rangle_{\mathcal{H}} + \text{const},$$

where $C : \mathcal{H} \to \mathcal{H}$ is a linear positive semidefinite operator given by

$$C := \int_{\mathcal{X}} p_0(x) \sum_{u=1}^{d} \partial_u k(x, \cdot) \otimes \partial_u k(x, \cdot) \mathrm{d}x \tag{3.1}$$

satisfying $Cf = \int_{\mathcal{X}} p_0(x) \sum_{u=1}^{d} \partial_u f(x) \partial_u k(x, \cdot) \mathrm{d}x$ for all $f \in \mathcal{H}$, and $z : \mathcal{X} \to \mathbb{R}$ is a function in $\mathcal{H}$ given by

$$z := -\int_{\mathcal{X}} p_0(x) \sum_{u=1}^{d} \Big(\partial_u^2 k(x, \cdot) + \partial_u \log \mu(x) \partial_u k(x, \cdot) \Big) \mathrm{d}x, \tag{3.2}$$

and const is a term that does not *depend on f and is equal to $\frac{1}{2}\int_{\mathcal{X}} p_0(x) \big\| \nabla \log p_0(x) - \nabla \log \mu(x) \big\|_2^2 \mathrm{d}x$.*

(b) *Let* $X_1, \cdots, X_n$ *be i.i.d samples from* p_0. *Under* ***(A1)*** - ***(A4)*** *in* ***Chapter 2***, ***(B1)*** - ***(B5)*** *above, and* $q_f \in \mathcal{Q}_{\text{kef}}$, *the SM loss functional is* $\widehat{J}_{\text{SM}} : \mathcal{H} \to \mathbb{R}$ *given by*

$$\widehat{J}_{\text{SM}}(f) := \frac{1}{2}\langle f, \widehat{C} f \rangle_{\mathcal{H}} - \langle f, \hat{z} \rangle_{\mathcal{H}} + \text{const},$$

where $\widehat{C} : \mathcal{H} \to \mathcal{H}$ *and* $\hat{z} \in \mathcal{H}$ *are given by* (2.10) *and* (2.11) *in* ***Chapter 2***, *respectively.*

The proof of this theorem can be found in Sriperumbudur et al. (2017) and is omitted here.

As we have discussed in **Section 2.2 in Chapter 2**, minimizing $\widehat{J}_{\text{SM}}$ over $\mathcal{H}$ does *not* have a solution. In order to obtain a solution, we need to impose certain kind of regularization. We apply the gradient descent algorithm with constant step size to minimizing $\widehat{J}_{\text{SM}}$ and terminate it early to regularize. Since $\widehat{J}_{\text{SM}}$ maps from $\mathcal{H}$ to $\mathbb{R}$, we need a notion of the gradient defined for functionals whose input space is a Hilbert space. We again use the Fréchet gradient (see Definition 2 in **Appendix 1**) and derive the Fréchet gradient of $\widehat{J}_{\text{SM}}$ in the following theorem.

Theorem 2 (Fréchet gradient of $\widehat{J}_{\text{SM}}$)**.** *Under the same assumptions in Theorem 1, the Fréchet gradient of* $\widehat{J}_{\text{SM}}$ *is a map from* $\mathcal{H}$ *to* $\mathcal{H}$ *given by*

$$\widehat{J}_{\text{SM}}(f) = \widehat{C} f - \hat{z}, \qquad \text{for all } f \in \mathcal{H}.$$

The proof of Theorem 2 can be found in Section 3.5.1.

With the result of Theorem 2, starting from some $\hat{f}^{(0)} \in \mathcal{H}$, the gradient descent

iterates are

$$\hat{f}^{(t+1)} = \hat{f}^{(t)} - \tau_t \nabla \widehat{J}_{\mathrm{SM}}(\hat{f}^{(t)}) = \hat{f}^{(t)} - \tau_t(\widehat{C}\hat{f}^{(t)} - \hat{z}), \qquad \text{for all } t \in \mathbb{N}_0, \tag{3.3}$$

where $\tau_t \in (0, 1/(d\kappa_2^2))$ is the step size at the t-th iterate.

Using the definition of the second-order Fréchet derivative and gradient (Definition 3 in **Appendix 1**), we have $\widehat{J}_{\mathrm{SM}}$ is twice Fréchet differentiable and $\nabla^2 \widehat{J}_{\mathrm{SM}}(f) = \widehat{C}$, for all $f \in \mathcal{H}$. Using **(B5)**, we have $\|\widehat{C}\| \leq d\kappa_2^2$, where $\|\widehat{C}\|$ denotes the operator norm of $\widehat{C}$. By the standard results on the gradient descent algorithm, with the choice of $\tau_t \in (0, 1/(d\kappa_2^2))$, the value of $\widehat{J}_{\mathrm{SM}}$ is guaranteed to decrease at each iteration.

The following theorem links $\hat{f}^{(t+1)}$ to $\hat{f}^{(0)}$ for an arbitrary $t \in \mathbb{N}_0$ and will help us derive a practical algorithm to compute $\hat{f}^{(t+1)}$ and facilitate our analysis in Section 3.3.

Theorem 3. *Starting the gradient descent algorithm from $\hat{f}^{(0)} \in \mathcal{H}$, the $(t+1)$-st gradient descent iterate is*

$$\hat{f}^{(t+1)} = \prod_{i=0}^{t}(I - \tau_i \widehat{C})\hat{f}^{(0)} + \sum_{j=0}^{t}\left[\prod_{i=j+1}^{t}(I - \tau_i \widehat{C})\right]\tau_j \hat{z}, \qquad \textit{for all } t \in \mathbb{N}_0, \tag{3.4}$$

where $I : \mathcal{H} \to \mathcal{H}$ is the identity operator, and $\prod_{i=t+1}^{t}(I - \tau_i \widehat{C}) = I$ for all $t \in \mathbb{N}_0$. If, in particular, we choose the constant step size, that is, $\tau_t \equiv \tau$ for all $t \in \mathbb{N}_0$, we have

$$\hat{f}^{(t+1)} = (I - \tau \widehat{C})^{t+1}\hat{f}^{(0)} + \tau \sum_{j=0}^{t}(I - \tau \widehat{C})^{t-j}\hat{z}.$$

Here, $(I - \tau\widehat{C})^i\hat{z}$*, for any* $i \in \mathbb{N}_0$*, is defined as*

$$(I - \tau\widehat{C})^0\hat{z} = \hat{z}, \qquad \textit{and} \qquad (I - \tau\widehat{C})^i\hat{z} = (I - \tau\widehat{C})\left[(I - \tau\widehat{C})^{i-1}\hat{z}\right] \textit{ for all } i \geq 1.$$

The proof of Theorem 3 can be found in Section 3.5.2.

In order to compute $\hat{f}^{(t)}$, we have to choose $\hat{f}^{(0)}$. Different choices of $\hat{f}^{(0)}$ leads to different trajectories of gradient descent iterates and, hence, different density estimates. We choose $\hat{f}^{(0)} = 0 \in \mathcal{H}$ for two reasons. First, this choice makes the computation and the theoretical analysis in the sequel easier, since, with $\hat{f}^{(0)} = 0$, (3.4) becomes

$$\hat{f}^{(t+1)} = \sum_{j=0}^{t}\left[\prod_{i=j+1}^{t}(I - \tau_i\widehat{C})\right]\tau_j\hat{z},$$

and we can ignore the term $\prod_{i=0}^{t}(I-\tau_i\widehat{C})\hat{f}^{(0)}$ that may not easy to deal with. Second, this choice makes the early stopping and penalized SM density estimators comparable. In the penalized approach, the penalty term $\frac{1}{2}\|f\|_{\mathcal{H}}^2$ shrinks the natural parameter toward the zero function, and, as the penalty parameter $\rho \to \infty$, $q_{\hat{f}^{(\rho)}} \to \mu$. In the early stopping approach, with $\hat{f}^{(0)} = 0$, we have $q_{\hat{f}^{(0)}} = \mu$, corresponding to the case $\rho \to \infty$ in the penalized approach. As the gradient descent algorithm evolves, the resulting density estimates correspond to those of smaller ρ values.

3.2.1 Computation of $\hat{f}^{(t)}$

Even though Theorem 3 is useful in characterizing $\hat{f}^{(t)}$ for each $t \in \mathbb{N}_0$ and in the subsequent analysis, it is not very directly applicable to produce an implementable algorithm to compute $f^{(t)}$, since (3.4) involves $\widehat{C}$ and $\hat{z}$ that reside in an infinite-dimensional RKHS. With the choice of $\hat{f}^{(0)} = 0 \in \mathcal{H}$, our goal of this section is to drive a practical algorithm to compute $\hat{f}^{(t)}$ for each $t \in \mathbb{N}$.

By the discussion in the preceding section, with $\hat{f}^{(0)} = 0$ and a constant step size $\tau_t = \tau \in (0, 1/(d\kappa_2^2))$ for all $t \in \mathbb{N}_0$, the t-th gradient descent iterate is

$$\hat{f}^{(t+1)} = \tau \sum_{j=0}^{t} (I - \tau \widehat{C})^{t-j} \hat{z} = \tau \sum_{j=0}^{t} (I - \tau \widehat{C})^{j} \hat{z}, \qquad \text{for all } t \in \mathbb{N}_0.$$

The following theorem provides an alternative expression for $\hat{f}^{(t)}$ that helps compute it.

Theorem 4. *Let $\hat{f}^{(0)} = 0 \in \mathcal{H}$ and the step size be the constant $\tau \in (0, 1/(d\kappa_2^2))$. Then, for all $t \in \mathbb{N}_0$, we have*

$$\hat{f}^{(t+1)} = (t+1)\tau\hat{z} + \tau \sum_{i=1}^{n} \sum_{u=1}^{d} \left[\sum_{j=1}^{t} \sum_{\ell=1}^{j} \binom{j}{\ell} \left(\frac{(-\tau)^\ell}{n^\ell} \mathbf{G}^{\ell-1} \mathbf{h} \right) \right]_{(i-1)d+u} \partial_u k(X_i, \cdot) \tag{3.5}$$

$$= (t+1)\tau\hat{z} + \sum_{i=1}^{n} \sum_{u=1}^{d} \alpha^{(t)}_{(i-1)d+u} \partial_u k(X_i, \cdot), \tag{3.6}$$

where $\mathbf{G}$ is an $(nd) \times (nd)$ matrix with the $((i-1)d+u, (j-1)d+v)$-th entry being $\langle \partial_u k(X_i, \cdot), \partial_v k(X_j, \cdot) \rangle_{\mathcal{H}} = \partial_u \partial_{v+d} k(X_i, X_j)$, $\mathbf{h}$ is an $(nd) \times 1$ vector with the

$((i-1)d+u)$-*th entry being*

$$\langle \hat{z}, \partial_u k(X_i, \cdot)\rangle_{\mathcal{H}} = -\frac{1}{n}\sum_{j=1}^{n}\sum_{v=1}^{d}\Big(\partial_u \partial_{v+d}^2 k(X_i, X_j) + \partial_v \log \mu(X_j)\partial_u \partial_{v+d} k(X_i, X_j)\Big),$$

and $\alpha^{(t)}_{(i-1)d+u} = -\frac{\tau^2}{n}\big[\mathbf{Q}\widetilde{\mathbf{\Lambda}}^{(t)}\mathbf{Q}^\top \mathbf{h}\big]_{(i-1)d+u}$, $\mathbf{G} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ *is the eigen-decomposition of the matrix* $\mathbf{G}$ *with* $\mathbf{Q} \in \mathbb{R}^{nd\times nd}$ *being an orthogonal matrix and* $\mathbf{\Lambda} \in \mathbb{R}^{nd\times nd}$ *being the diagonal matrix with all eigenvalues* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{nd} \geq 0$ *on the diagonal, and* $\widetilde{\mathbf{\Lambda}}^{(t)} \in \mathbb{R}^{nd\times nd}$ *is a diagonal matrix with diagonal elements* $\tilde{\lambda}_1^{(t)}, \cdots, \tilde{\lambda}_{nd}^{(t)}$ *being*

$$\tilde{\lambda}_w^{(t)} = \begin{cases} -\left(\dfrac{n}{\tau\lambda_w}\right)^2\left(1-\dfrac{\tau}{n}\lambda_w\right)\left(1-\left(1-\dfrac{\tau}{n}\lambda_w\right)^t\right) + \dfrac{tn}{\tau\lambda_w}, & \text{if } \lambda_w \neq 0, \\ \dfrac{(t+1)t}{2}, & \text{if } \lambda_w = 0, \end{cases}$$

for all $w = 1, \cdots, nd$.

The proof of Theorem 4 is provided in Section 3.5.3.

Remark 1. We discuss an alternative implementation of the gradient descent algorithm.

Recall that, starting from $\hat{f}^{(0)} = 0$, gradient descent iterates are

$$\hat{f}^{(t+1)} = \hat{f}^{(t)} - \tau(\widehat{C}\hat{f}^{(t)} - \hat{z}),$$

which belongs to the union of range$(\widehat{C})$ and $\{z\}$. Since, for an arbitrary $g \in \mathcal{H}$,

$\widehat{C}g = \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\langle g, \partial_u k(X_i, \cdot)\rangle_{\mathcal{H}}\partial_u k(X_i, \cdot)$, we have

$$\text{range}(\widehat{C}) = \text{Span}\Big\{\partial_u k(X_i, \cdot), \text{ for all } i = 1, \cdots, n, \text{ and } u = 1, \cdots, d\Big\}.$$

Therefore, the gradient descent iterates must lie in the linear subspace

$$\text{Span}\Big\{\partial_u k(X_i, \cdot) \text{ for all } i = 1, \cdots, n \text{ and } u = 1, \cdots, d, \text{ and } \hat{z}\Big\}, \tag{3.7}$$

which is of the dimensionality (at most) $nd + 1$.

Since any $\tilde{f}$ belonging to (3.7) can be written as

$$\tilde{f} = \sum_{j=1}^{n}\sum_{v=1}^{d}\beta_{(j-1)d+v}\partial_v k(X_j, \cdot) + \beta_{nd+1}\hat{z}, \tag{3.8}$$

plugging (3.8) into $\widehat{J}_{\text{SM}}$ yields

$$\widetilde{J}_{\text{SM}}(\boldsymbol{\beta}) := \widehat{J}_{\text{SM}}(\tilde{f}) = \frac{1}{2n}\boldsymbol{\beta}^\top\widetilde{\mathbf{G}}^\top\widetilde{\mathbf{G}}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\widetilde{\mathbf{h}},$$

where $\boldsymbol{\beta} := (\beta_1, \cdots, \beta_{nd}, \beta_{nd+1})^\top \in \mathbb{R}^{nd+1}$, and

$$\widetilde{\mathbf{G}} := \Big[\begin{array}{c:c} \mathbf{G} & \mathbf{h} \end{array}\Big] \in \mathbb{R}^{nd\times(nd+1)}, \qquad \text{and} \qquad \widetilde{\mathbf{h}} := \Big[\mathbf{h}^\top \vdots \|\hat{z}\|_{\mathcal{H}}^2\Big]^\top \in \mathbb{R}^{(nd+1)\times 1}.$$

Since the gradient vector of $\widetilde{J}_{\text{SM}}$ at $\boldsymbol{\beta}$ is $\nabla\widetilde{J}_{\text{SM}}(\boldsymbol{\beta}) = \frac{1}{n}\widetilde{\mathbf{G}}^\top\widetilde{\mathbf{G}}\boldsymbol{\beta} - \widetilde{\mathbf{h}}$, starting from

$\boldsymbol{\beta}^{(0)} \in \mathbb{R}^{nd+1}$, the gradient descent iterates are

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \tilde{\tau}\left(\frac{1}{n}\widetilde{\mathbf{G}}^\top\widetilde{\mathbf{G}}\boldsymbol{\beta}^{(t)} - \widetilde{\mathbf{h}}\right), \qquad \text{for all } t \in \mathbb{N}_0,$$

where $\tilde{\tau} \in \left(0, \frac{1}{n}\lambda_{\max}(\widetilde{\mathbf{G}}^\top\widetilde{\mathbf{G}})\right)$ is the constant step size, and $\lambda_{\max}(\widetilde{\mathbf{G}}^\top\widetilde{\mathbf{G}})$ is the largest eigenvalue of $\widetilde{\mathbf{G}}^\top\widetilde{\mathbf{G}}$.

Correspondingly, the t-th gradient descent iterate of the natural parameter is

$$\tilde{f}^{(t)} = \sum_{j=1}^{n}\sum_{v=1}^{d}\beta_{(j-1)d+v}^{(t)}\partial_v k(X_j, \cdot) + \beta_{nd+1}^{(t)}\hat{z},$$

where $\beta_w^{(t)}$ is the w-th component of $\boldsymbol{\beta} \in \mathbb{R}^{nd+1}$, for all $w = 1, \cdots, wd+1$.

3.2.2 Numerical Examples of Early Stopping SM Density Estimates

We now use the `waiting` data in the Old Faithful Geyser dataset introduced in **Chapter 1** to illustrate the early stopping SM density estimates.

We let $\mathcal{X} = (0, \infty)$ and choose the base density μ to be the pdf of Gamma distribution with shape and scale parameters being 26 and 3, respectively. Plots of μ and its logarithm are shown in Figure 1. We choose the kernel function to be the Gaussian kernel with the bandwidth parameter $\sigma = 5$,

$$k(x, y) = \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right), \qquad \text{for all } x, y \in \mathcal{X}.$$

In computing the early stopping density estimates, we use $\tau = 20$. Note that, with this particular choice of the kernel function, $\kappa_2^2 = \frac{1}{25}$ and $\tau = 20 \in (0, 1/(d\kappa_2^2)) = (0, 25)$.

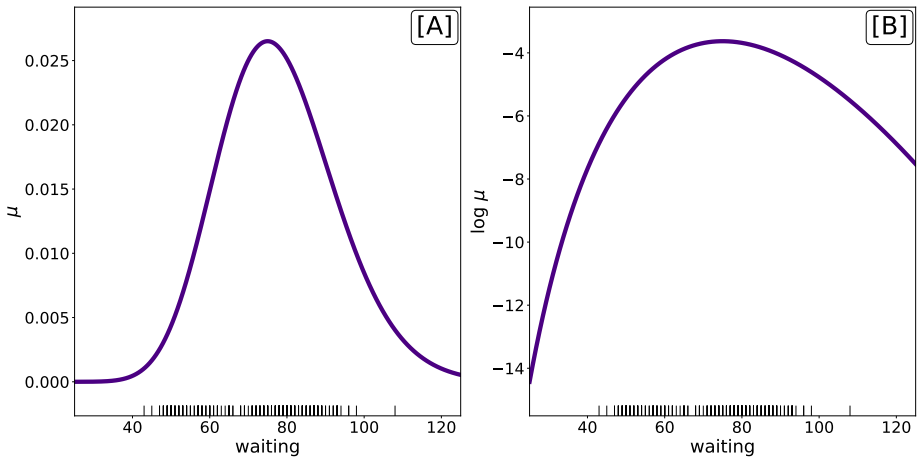

Figure 1: Panel [A] shows μ and [B] shows $\log \mu$. The rug plot indicates the location of data.

The resulting early stopping SM density estimates are shown in Figure 2. Note that when the number of iterations t is small, the resulting density estimates are very close to μ, as we expect. As t increases, the density estimates become reasonable and show the bimodal feature of the data. However, as t becomes very large, the density estimates contain a bump or become a spike at the observation 108. We will return to this phenomenon in Section 3.3.1 and in the subsequent chapters of this dissertation.

3.2.3 When to Terminate the Algorithm

With the numerical examples shown in the preceding section, we see the choice of the number of iterations is critical to producing satisfactory density estimates. The goal of this section is to discuss how to determine when to terminate the gradient descent algorithm in practice. We will provide two methods: the hold-out method and the K-fold cross validation.

The hold-out method randomly partition the entire data into two parts, where one part, denoted by $\mathbf{S}_{\text{train}}$, is used to perform the gradient descent algorithm and

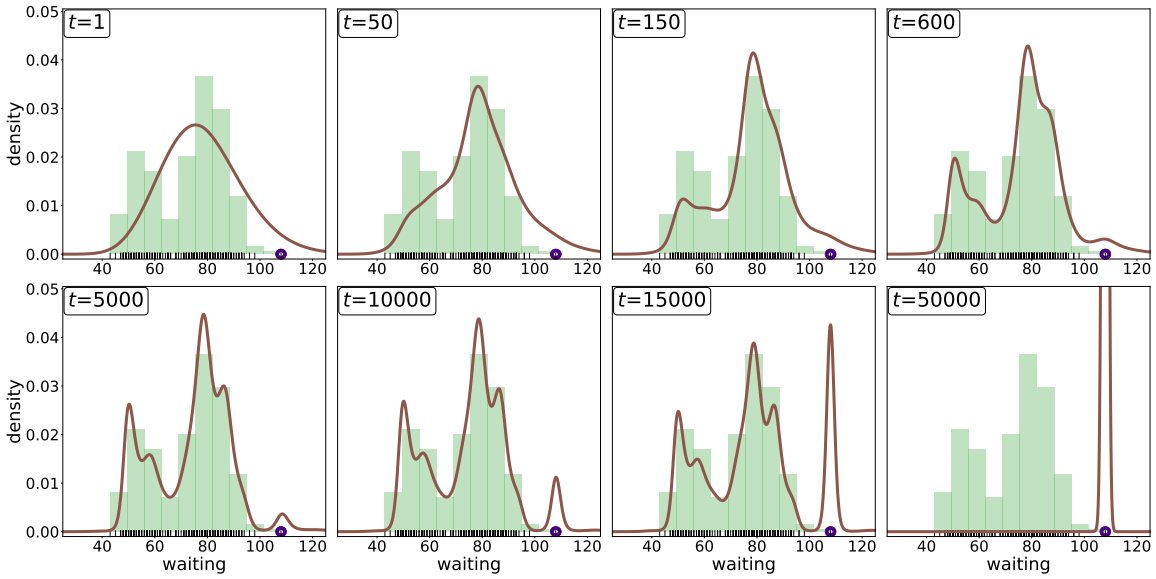

Figure 2: Early stopping SM density estimates for different values of number of iterations labeled at the upper left corner. Histogram of data using the Freedman-Diaconis rule is shown in green. The rug plot indicates the location of data and the purple circle indicates the location of the observation 108.

the other part, denoted by $\mathbf{S}_{\text{test}}$, is used to evaluate the performance of the estimator. Let $\widehat{J}_{\text{SM,train}}$ and $\widehat{J}_{\text{SM,test}}$ be the SM loss functionals constructed using $\mathbf{S}_{\text{train}}$ and $\mathbf{S}_{\text{test}}$, respectively. Starting from $\hat{f}^{(0)} = 0$, we perform the gradient descent algorithm on $\widehat{J}_{\text{SM,train}}$, and after obtaining a gradient descent iterate $\hat{f}^{(t)}$, we compute $\widehat{J}_{\text{SM,test}}(\hat{f}^{(t)})$, and terminate when $\widehat{J}_{\text{SM,test}}(\hat{f}^{(t+1)}) > \widehat{J}_{\text{SM,test}}(\hat{f}^{(t)})$. The algorithm is shown in Algorithm 1.

The major drawback of this hold-out method is that it has to set aside a portion of data for the evaluation purpose, rather than use the entire data for the estimation purpose.

The second method is the K-fold cross validation (CV). We first specify a set of number of iterations candidates to terminate the algorithm, and randomly partition all data into K folds of (roughly) equal size. Let us fix one number of iterations

Algorithm 1 Hold-out method to determine when to terminate

Require:

- $X_1, \cdots, X_n$, the data from p_0;
- $m < n$, the number of observations in $\mathbf{S}_{\text{train}}$;
- τ, the constant step size.

Randomly shuffle the data $X_1, \cdots, X_n$, and let the first m observations be $\mathbf{S}_{\text{train}}$ and the remaining be $\mathbf{S}_{\text{test}}$;
Let $\texttt{error1} = \widehat{J}_{\text{SM,test}}(\hat{f}^{\texttt{old}})$, where $\hat{f}^{\texttt{old}} = \hat{f}^{(0)} = 0$;
Let $\hat{f}^{\texttt{new}} = \hat{f}^{\texttt{old}} - \tau \nabla \widehat{J}_{\text{SM,train}}(\hat{f}^{\texttt{old}})$;
Compute $\texttt{error2} = \widehat{J}_{\text{SM,test}}(\hat{f}^{\texttt{new}})$;
while $\texttt{error1} > \texttt{error2}$ **do**
    Let $\hat{f}^{\texttt{old}} = \hat{f}^{\texttt{new}}$ and $\texttt{error1} = \texttt{error2}$;
    Update $\hat{f}^{\texttt{new}} = \hat{f}^{\texttt{old}} - \tau \nabla \widehat{J}_{\text{SM,train}}(\hat{f}^{\texttt{old}})$;
    Compute $\texttt{error2} = \widehat{J}_{\text{SM,test}}(\hat{f}^{\texttt{new}})$;
return $\hat{f}^{\texttt{old}}$.

candidate, say t_0, for now. We use $K-1$ folds of data to perform the gradient descent algorithm and terminate at t_0, and evaluate the SM loss functional constructed using the remaining fold of data at this gradient descent iterate. Note that, for each number of iterations candidate, we end up with K values of the SM loss functional, one obtained from each fold. We average all these K values and regard this average as the estimated risk of the corresponding number of iterations candidate. Repeat this process for each number of iterations candidate, and obtain an estimated risk for each of them. The best number of iterations is the one with the lowest estimated risk. In the end, we perform the gradient descent algorithm on the entire data and terminate at this best number of iterations. It is obvious that the final early stopping SM density estimate utilizes all data and avoid the drawback in the hold-out method. The complete algorithm of this K-fold CV method is shown in Algorithm 2.

Algorithm 2 K-fold cross validation method to determine when to terminate

Require:

- $X_1, \cdots, X_n$, the data from p_0;
- K, the number of folds for cross validation;
- τ, the constant step size;
- $t_1, \cdots, t_m$, a list of number of iterations candidates to terminate.

Randomly partition the data $X_1, \cdots, X_n$ into K folds, denoted by $\mathbf{S}_1, \cdots, \mathbf{S}_K$;
Set `metric` to be an empty list to record the estimated risks for different folds later;
for $j = 1, \cdots, m$ **do**
    Set `metric_j` $= 0$;
    for $\ell = 1, \cdots, K$ **do**
        Let $\mathbf{S}_{\text{test}} = \mathbf{S}_\ell$ and $\mathbf{S}_{\text{train}}$ contain data excluding $\mathbf{S}_\ell$;
        Perform the gradient descent algorithm on $\widehat{J}_{\text{SM,train}}$, which is the SM loss functional constructed using $\mathbf{S}_{\text{train}}$, and terminate at t_j; denote the resulting natural parameter by $\hat{f}_\ell^{(t_j)}$;
        Update `metric_j` $+= \widehat{J}_{\text{SM,test}}(\hat{f}_\ell^{(t_j)})$, where $\widehat{J}_{\text{SM,test}}$ is the SM loss functional constructed using $\mathbf{S}_{\text{test}}$;
    Append `metric_j`$/K$ to the end of `metric`;
Let $m^* \in \{1, \cdots, m\}$ be the one with the lowest value in `metric`. Then, t_{m^*} is the best number of iterations;
Perform the gradient descent algorithm on $\widehat{J}_{\text{SM}}$, the SM loss functional constructed using *all* data, and terminate at t_{m^*}.
return $\hat{f}^{(t_{m^*})}$.

3.3 Theoretical Properties of Early Stopping SM Density Estimator

With the introduction to our early stopping SM density estimator and its computation, we discuss its theoretical properties in this section.

We will assess its theoretical properties from two aspects. First, we will look at what happens to the density estimator if we do not terminate the gradient descent algorithm and let the number of iteration approach to ∞. This will be covered in Section 3.3.1.

Second, in this early stopping approach, we use the number of iterations as a tuning parameter to perform implicit regularization, which plays exactly the same role as the penalty parameter ρ in the penalized approach. We derive a (theoretical) stopping rule and establish the convergence rates of the resulting natural parameter estimator and the density estimator using this stopping rule in Section 3.3.2.

3.3.1 Limiting SM Density Estimator as $t \to \infty$

In this section, we investigate the behavior of the early stopping SM density estimator if we keep running the gradient descent algorithm without termination.

Our main theorem is the following.

Theorem 5. *Let* $\hat{z}_2 \in \mathcal{H}$ *be the orthogonal projection of* $\hat{z}$ *onto* $\text{range}(\widehat{C})^{\perp}$*, the orthogonal complement of* $\text{range}(\widehat{C})$*. In addition to* ***(A1)*** *-* ***(A4)*** *in* ***Chapter 2*** *and* ***(B1)*** *-* ***(B5)****, we further assume*

(C1) *there exists a unique* $x^* \in \mathcal{X}$ *such that* $\hat{z}_2(x^*) > \hat{z}_2(x)$ *for all* $x \in \mathcal{X} \backslash \{x^*\}$, *and*

(C2) $\tau \in (0, 1/(d\kappa_2^2))$.

Then, $\lim_{t\to\infty} q_{\hat{f}^{(t)}}(x^*) = \infty$.

Remark 2. We can relax **(C1)** to the one that the maximizers of $\hat{z}_2$ form a set of Lebesgue measure zero. Let $\mathcal{M}$ denote the set of the maximizers of $\hat{z}_2$. Then, we can modify the proof slightly and conclude $\lim_{t\to\infty} q_{\hat{f}^{(t)}}(x^*) = \infty$ at all $x^* \in \mathcal{M}$.

The proof of Theorem 5 will be provided in Section 3.5.4, and is built upon the decomposition of $\hat{f}^{(t)}$ which we now look at.

3.3.1.1 Decomposition of $\hat{f}^{(t)}$

We decompose $\hat{f}^{(t)}$ into two parts, where one part resides in range$(\widehat{C})$ and the other resides in range$(\widehat{C})^{\perp}$, the orthogonal complement of range$(\widehat{C})$.

As we have discussed in the remark following Theorem 4, range$(\widehat{C})$ contains all functions that can be written as a linear combination of $\partial_u k(X_i, \cdot)$, for all $i = 1, \cdots, n$ and $u = 1, \cdots, d$, which is of finite dimension and forms a closed linear subspace of $\mathcal{H}$ (Corollary 6 in Section 13.3 in Royden and Fitzpatrick, 2018). Its orthogonal complement is

$$\text{range}(\widehat{C})^{\perp} := \Big\{ g \in \mathcal{H} \;\Big|\; \langle g, f \rangle_{\mathcal{H}} = 0, \text{ for all } f \in \text{range}(\widehat{C}) \Big\},$$

which also forms a closed linear subspace of $\mathcal{H}$ (Section 16.1 in Royden and Fitzpatrick, 2018).

Notice that $\text{range}(\widehat{C})^{\perp} \neq \{0\}$. To see this, suppose the opposite. Then, we have $\text{range}(\widehat{C}) = \mathcal{H}$ (Corollary 4 in Section 16.1 in Royden and Fitzpatrick, 2018). But, since $\text{range}(\widehat{C})$ is finite-dimensional, $\text{range}(\widehat{C}) = \mathcal{H}$ implies $\mathcal{H}$ is also finite-dimensional. This contradicts to **(A1)** in **Chapter 2** that $\mathcal{H}$ is infinite-dimensional.

Now, decompose $\hat{z}$ into two parts, $\hat{z}_1$ and $\hat{z}_2$, where $\hat{z}_1 := \Pi_{\text{range}(\widehat{C})}(\hat{z}) \in \text{range}(\widehat{C})$, $\hat{z}_2 := \Pi_{\text{range}(\widehat{C})^{\perp}}(\hat{z}) \in \text{range}(\widehat{C})^{\perp}$, and $\Pi_S(f)$ denotes the orthogonal projection of f onto the closed linear subspace S of $\mathcal{H}$. Note that both $\hat{z}_1$ and $\hat{z}_2$ are well-defined by the projection theorem in a Hilbert space (Theorem 2.3.1 in Brockwell and Davis, 2013).

Recall that

$$\hat{z} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\Big((\partial_u \log \mu)(X_i)\partial_u k(X_i, \cdot) + \partial_u^2 k(X_i, \cdot)\Big)$$

involves the first two partial derivatives of k. The component $-\frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\partial_u \log \mu(X_i)\partial_u k(X_i, \cdot)$ must belong to $\text{range}(\widehat{C})$, and the component $-\frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\partial_u^2 k(X_i, \cdot)$ does *not* necessarily belong to $\text{range}(\widehat{C})$, except for special choices of k. We assume $\hat{z}_2 \neq 0$ for the rest of this section.

In addition, since $\hat{z}_1 \in \text{range}(\widehat{C})$, we can find $\hat{g}_1 \in \text{range}(\widehat{C})$ satisfying $\hat{z}_1 = \widehat{C}\hat{g}_1$. We claim such a choice of $\hat{g}_1$ is unique. Suppose there exists $\tilde{g}_1 \in \text{range}(\widehat{C})$ and $\tilde{g}_1 \neq \hat{g}_1$ such that $\hat{z}_1 = \widehat{C}\tilde{g}_1$, then $0 = \widehat{C}(\hat{g}_1 - \tilde{g}_1)$, implying that $\hat{g}_1 - \tilde{g}_1 = 0$ or $\hat{g}_1 - \tilde{g}_1 \in \text{range}(\widehat{C})^{\perp}$ is nonzero. Since $\text{range}(\widehat{C})$ is a closed linear subspace of $\mathcal{H}$, the latter case cannot happen. We deduce that $\tilde{g}_1 = \hat{g}_1$, and the uniqueness follows.

With the decomposition of $\hat{z}$ we have discussed so far, the following proposition

decomposes $\hat{f}^{(t)}$ into two components, where one component belongs to range$(\widehat{C})$ and the other belongs to range$(\widehat{C})^{\perp}$.

Proposition 1. *Suppose $\hat{f}^{(0)} = 0 \in \mathcal{H}$. The $(t+1)$-st gradient descent iterate $\hat{f}^{(t+1)}$, for each $t \in \mathbb{N}_0$, can be written as the sum of the following two components*

$$\hat{f}_1^{(t+1)} := (I - (I - \tau\widehat{C})^{t+1})\hat{g}_1 \in \text{range}(\widehat{C})$$
$$\hat{f}_2^{(t+1)} := (t+1)\tau\hat{z}_2 \in \text{range}(\widehat{C})^{\perp}$$

The following proposition gives explicit formulae for $\hat{z}_1$, $\hat{z}_2$, and $\hat{g}_1$.

Proposition 2. *Let $\mathbf{G} \in \mathbb{R}^{nd \times nd}$ and $\mathbf{h} \in \mathbb{R}^{nd}$ be the quantities defined in Theorem 4. Then,*

(a) $\hat{z}_1 = \sum_{i=1}^{n}\sum_{u=1}^{d} \alpha^*_{(i-1)d+u}\partial_u k(X_i, \cdot)$, *where* $\boldsymbol{\alpha}^* := (\alpha_1^*, \cdots, \alpha_{nd}^*) \in \mathbb{R}^{nd}$ *satisfies the linear system* $\mathbf{G}\boldsymbol{\alpha}^* = \mathbf{h}$;

(b) $\hat{z}_2 = \sum_{i=1}^{n}\sum_{u=1}^{d}\big[\big(-\frac{1}{n}(\partial_u \log \mu)(X_i) - \alpha^*_{(i-1)d+u}\big)\partial_u k(X_i, \cdot) - \frac{1}{n}\partial_u^2 k(X_i, \cdot)\big]$;

(c) $\hat{g}_1 = \sum_{i=1}^{n}\sum_{u=1}^{d} \beta^*_{(i-1)d+u}\partial_u k(X_i, \cdot)$, *where* $\boldsymbol{\beta}^* := (\beta_1^*, \cdots, \beta_{nd}^*) \in \mathbb{R}^{nd}$ *satisfies the linear system* $-\frac{1}{n}\mathbf{G}\boldsymbol{\beta}^* = \boldsymbol{\alpha}^*$.

Now that we have decomposed $\hat{f}^{(t+1)}$ into two components, the following proposition established the boundedness of $\hat{f}_1^{(t+1)}$ over $\mathcal{X}$.

Proposition 3. *Under the same assumptions in Theorem 5, there exists $M > 0$ such that, for all $x \in \mathcal{X}$ and all $t \in \mathbb{N}_0$, $|\langle \hat{f}_1^{(t+1)}, k(x, \cdot)\rangle_{\mathcal{H}}| \leq M$.*

The proofs of Propositions 1 - 3 can be found in Section 3.5.4.

3.3.1.2 Numerical Illustration of Theorem 5

We use the `waiting` data to illustrate Theorem 5.

With the help of (2.11) in **Chapter 2** and Proposition 2, we plot $\hat{z}$, $\hat{z}_1$ and $\hat{z}_2$ in the three panels of Figure 3. In particular, note that, from the right panel of Figure 3, $\hat{z}_2$ achieves the maximum at 108. With Theorem 5, we expect see that, as the number of iterations t increases, the density value at 108 also increases. Figure 4, the plot of the density value at 108 against the number of iterations, confirms our expectation.

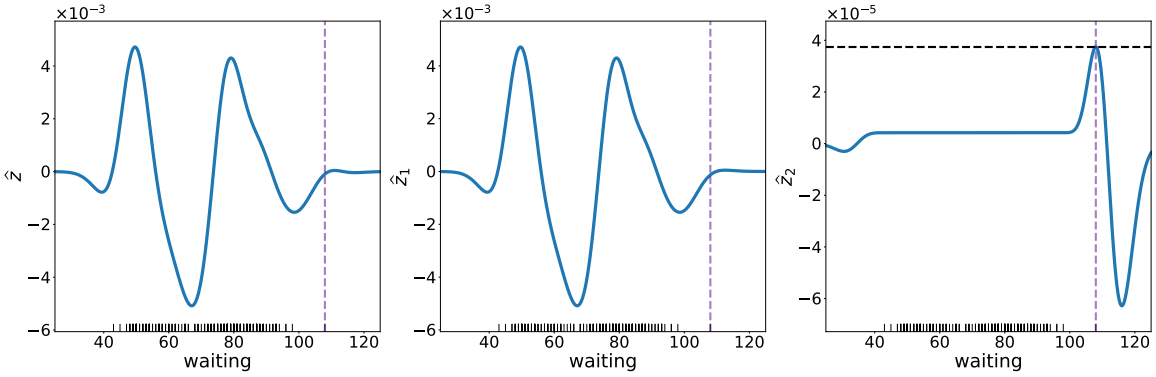

Figure 3: Plots of $\hat{z}$ (the left panel), $\hat{z}_1$ (the middle panel) and $\hat{z}_2$ (the right panel). The rug plot indicates the location of data.

3.3.2 Rate of Convergence

We study the non-asymptotic properties of the early stopping SM density estimator in this section. More specifically, we attempt to answer the following questions:

1. Assume $p_0 = q_{f_0} \in \mathcal{Q}_{\text{ker}}$ for some $f_0 \in \mathcal{H}$. What can we say about the gap between $\hat{f}^{(t)}$ and f_0, for each $t \in \mathbb{N}$?

2. What can we say about the gap between $q_{\hat{f}^{(t)}}$ and p_0, for each $t \in \mathbb{N}$?

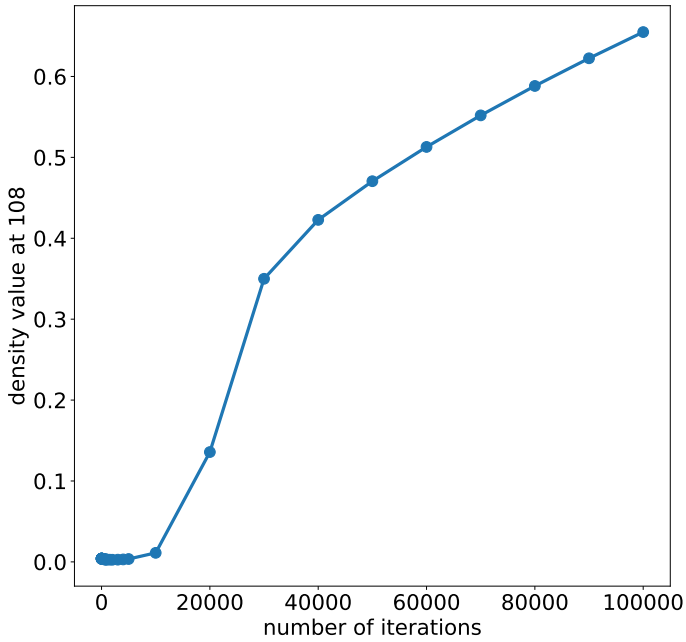

Figure 4: Density value at 108 against the number of iterations.

3. Based on the gaps above, can we derive an early stopping rule?

Throughout this section, besides **(B1)** - **(B5)**, we further assume the following:

(B6) The true density p_0 belongs to $\mathcal{Q}_{\text{ker}}$ and there exists $f_0 \in \mathcal{H}$ such that $p_0 = q_{f_0}$.

(B7) There exists $\gamma > 0$ such that $f_0 \in \text{range}(C^\gamma)$, where $C : \mathcal{H} \to \mathcal{H}$ is given by (3.1). That is, there exists $g_0 \in \mathcal{H}$ such that $C^\gamma g_0 = f_0$.

Our main result is the following theorem.

Theorem 6. *Under **(A1)** - **(A4)** in Chapter 2 and **(B1)** - **(B7)**, for each $n \in \mathbb{N}$, there exists an early stopping rule*

$$t^* : \mathbb{N} \to \mathbb{N}, \qquad n \mapsto \lceil n^{\frac{1}{2(\gamma+2)}} \rceil$$

such that, with probability at least $1-\delta$ *for* $\delta \in (0,1)$*, the following inequality holds*

$$\|\hat{f}^{(t^*(n))} - f_0\|_{\mathcal{H}} \leq (4C_1 + C_2) n^{-\frac{\gamma}{2(\gamma+2)}},$$

where $C_1 := 2d\tau t^2(\kappa_3+\kappa_4)(\tau d\kappa_2^2+1)\sqrt{2\log(2/\delta)}$ *and* $C_2 := \gamma^\gamma \|g_0\|_{\mathcal{H}}/(\tau e)^\gamma$.

The proof follows Yao, Rosasco, and Caponnetto (2007) and details will be provided in Section 3.5.5. The proof depends on the analysis of two sequences of gradient descent iterates: one is the gradient descent updates depending on the SM loss functional $\widehat{J}_{\mathrm{SM}}$ (which depends on finitely many samples) and the other one is the gradient descent updates depending on the H-divergence (or the population-version SM loss functional) $J_{\mathrm{SM}}(f)$ which can be viewed as the SM loss functional with infinitely many samples. We denote the former sequence by $\{\hat{f}^{(t)}\}_{t\in\mathbb{N}}$ as before and the latter one by $\{f^{(t)}\}_{t\in\mathbb{N}}$, and let both start from the zero function, i.e., $f^{(0)} = \hat{f}^{(0)} = 0 \in \mathcal{H}$. Using the triangle inequality, we can bound $\|\hat{f}^{(t)} - f_0\|_{\mathcal{H}}$ from above by

$$\|\hat{f}^{(t)} - f_0\|_{\mathcal{H}} \leq \|\hat{f}^{(t)} - f^{(t)}\|_{\mathcal{H}} + \|f^{(t)} - f_0\|_{\mathcal{H}}, \tag{3.9}$$

where the population-version sequence of iterates $\{f^{(t)}\}_{t\in\mathbb{N}}$ is used as an intermediate.

In (3.9), $\|f^{(t)} - f_0\|_{\mathcal{H}}$ is called the *approximation error* (or *bias*), and $\|\hat{f}^{(t)} - f^{(t)}\|_{\mathcal{H}}$, is called the *sample error* (or *variance*). We will see later that, as t increases, the derived upper bound of the sample error increases and that of the approximation error decreases. More specifically, as the gradient descent algorithm evolves, the population-version sequence of iterates $\{f^{(t)}\}_{t\in\mathbb{N}}$ converge to f_0 eventually but the

sample-version sequence of iterates $\{\hat{f}^{(t)}\}_{t\in\mathbb{N}}$ do *not*. Therefore, we have a bias-variance tradeoff we discussed earlier: if we stop the algorithm at a too early time, the resulting $\hat{f}^{(t)}$ has a small variance but a large bias; and, on the other hand, if we keep running it and terminate too late, the resulting $\hat{f}^{(t)}$ has a small bias but a large variance. By minimizing the sum of these two upper bounds, we can identify a stopping rule.

In the next two sections, we will focus on the approximation error and the sample error, respectively, and derive their upper bounds that will help us prove Theorem 6.

3.3.2.1 An Upper Bound on the Approximation Error

In this section, we focus on the problem of minimizing J_{SM} over $\mathcal{H}$, discuss the population-version sequence of iterates $\{f^{(t)}\}_{t\in\mathbb{N}}$, and assess the approximation error $\|f^{(t)} - f_0\|_{\mathcal{H}}$.

Proposition 4 (Frechét gradient of J_{SM})**.** *Under* ***(A1)*** *-* ***(A4)*** *in Chapter 2 and* ***(B1)*** *-* ***(B5)****, the Frechét gradient of J_{SM}, denoted by ∇J_{SM}, is a map from $\mathcal{H}$ to $\mathcal{H}$ given by $\nabla J_{\mathrm{SM}}(f) = Cf - z$ for all $f \in \mathcal{H}$.*

The proof of Proposition 4 is almost identical to that of Theorem 2 and is omitted here.

With Proposition 4, the population-version of gradient descent iterates are

$$f^{(t+1)} = f^{(t)} - \tau\nabla J_{\mathrm{SM}}(f^{(t)}) = (I - \tau C)f^{(t)} + \tau z, \qquad \text{for all } t \in \mathbb{N}_0,$$

where $\tau \in (0, 1/(d\kappa_2^2))$ is the constant step size and $f^{(0)} \in \mathcal{H}$ is the starting point.

Using the same argument as that of Theorem 3, we can link $f^{(t+1)}$ to $f^{(0)}$ as

$$f^{(t+1)} = (I - \tau C)^{t+1} f^{(0)} + \tau \sum_{j=0}^{t} (I - \tau C)^{t-j} z.$$

With all ingredients above, we can now derive an upper bound of the approximation error $\|f^{(t)} - f_0\|_{\mathcal{H}}$.

Theorem 7 (An upper bound on the approximation error)**.** *Choose $f^{(0)} = 0$ and the step size $\tau \in (0, 1/(d\kappa_2^2))$. Under the same assumptions in Theorem 6, we have, for all $t \in \mathbb{N}$,*

$$\|f^{(t)} - f_0\|_{\mathcal{H}} \leq \left(\frac{\gamma}{\tau e t}\right)^{\gamma} \|g_0\|_{\mathcal{H}}.$$

The proof of Theorem 7 will be provided in Section 3.5.6.

In particular, note that $\|f^{(t)} - f_0\|_{\mathcal{H}} \leq \big((\frac{\gamma}{\tau e})^{\gamma} \|g_0\|_{\mathcal{H}}\big) t^{-\gamma}$. If we let $t \to \infty$, we obtain $\|f^{(t)} - f_0\|_{\mathcal{H}} \to 0$ and deduce $f^{(t)} \to f_0$. Therefore, as the gradient descent algorithm evolves, the approximation error decreases to 0 and the population-version iterates converge to the true natural parameter f_0.

3.3.2.2 An Upper Bound on the Sample Error

In this section, we turn to the sample error $\|\hat{f}^{(t)} - f^{(t)}\|_{\mathcal{H}}$ and derive an upper bound of it. The main result is the following.

Theorem 8 (An upper bound of the sample error)**.** *For both population- and sample-version iterations, choose $f^{(0)} = \hat{f}^{(0)} = 0$ and the common constant step size $\tau \in$*

$(0, 1/(d\kappa_2^2))$. *Under the same assumptions in Theorem 6, with probability at least* $1 - \delta$ *for* $\delta \in (0, 1)$*, the following inequality holds*

$$\|\hat{f}^{(t)} - f^{(t)}\|_{\mathcal{H}} \leq 2d\tau t^2(\kappa_3 + \kappa_4)(\tau d\kappa_2^2 + 1)\sqrt{\frac{2}{n}\log\left(\frac{2}{\delta}\right)},$$

for all $t \in \mathbb{N}$.

The proof of Theorem 8 will be provided in Section 3.5.7.

3.3.2.3 Upper Bounds on the Distances between p_0 and $q_{\hat{f}^{(t^*(n))}}$

Theorem 6 establishes the stopping rule $t^* := t^*(n)$ and the convergence rate of $\hat{f}^{(t^*)}$. We can carry it over to bounding various distances between p_0 and $q_{\hat{f}^{(t^*)}}$. The main result is the following corollary.

Corollary 1 (Various distances between p^* and $q_{\hat{f}^{(t^*)}}$)**.** *Under the same assumptions as in Theorem 6, with the stopping rule* t^* *therein, the following inequalities hold with probability at least* $1 - \delta$ *for* $\delta \in (0, 1)$*,*

(a) in terms of the H-divergence,

$$\mathrm{H}(p_0 \| q_{\hat{f}^{(t^*)}}) \leq C_3 n^{-\frac{\gamma}{\gamma+2}};$$

(b) in terms of the KL-divergence,

$$\mathrm{KL}(p_0 \| q_{\hat{f}^{(t^*)}}) \leq C_4 n^{-\frac{\gamma}{\gamma+2}};$$

(c) in term of the Hellinger distance defined as $\mathrm{He}(p\|q) := \left(\frac{1}{2}\int_{\mathcal{X}}(\sqrt{p(x)}-\sqrt{q(x)})^2\mathrm{d}x\right)^{\frac{1}{2}}$ *for any two pdfs* $p, q : \mathcal{X} \to [0, \infty)$*,*

$$\mathrm{He}(p_0\|q_{\hat{f}^{(t^*)}}) \leq C_5 n^{-\frac{\gamma}{2(\gamma+2)}};$$

(d) in terms of the L^1 *distance,*

$$\|p_0 - q_{\hat{f}^{(t^*)}}\|_{L^1} \leq C_6 n^{-\frac{\gamma}{2(\gamma+2)}}.$$

All C_3, C_4, C_5 *and* C_6 *are positive constants independent of the sample size* n *and will be made explicit in the proof.*

The proof of Corollary 1 will be provided in Section 3.5.8.

3.3.2.4 Discussion on (B7)

We end this section by a discussion on **(B7)**.

It is well-known that establishing the convergence rate is possible only if certain smoothness assumption is made on the quantity of interest, which is f_0 in our case. In the literature on the nonparametric function estimation, the classic smoothness assumption has been made on the differentiability and the continuity of the derivatives of f_0, i.e., the Hölder condition (see, for example, Chapter 1 in Tsybakov, 2009). In our analysis above, the smoothness condition is **(B7)**, i.e., $f_0 \in \mathrm{range}(C^\gamma)$ for some $\gamma > 0$. This assumption has been used in the studies of various kernel-based machine learning algorithms by, for example, Caponnetto and De Vito (2006), Bauer,

Pereverzev, and Rosasco (2007), Smale and Zhou (2007), Yao, Rosasco, and Caponnetto (2007), Lo Gerfo et al. (2008), Rastogi and Sampath (2017), and Lin et al. (2018).

We now elucidate what this assumption really means. To start with, recall that $C : \mathcal{H} \to \mathcal{H}$ is self-adjoint and compact (Theorem 4(i) in Sriperumbudur et al., 2017). By the Hilbert-Schmidt theorem (Section 16.6 in Royden and Fitzpatrick, 2018), there exists an orthonormal basis for $\overline{\text{range}(C)}$, $\{\psi_\nu\}_{\nu\in\mathbb{N}}$, such that $C\psi_\nu = \xi_\nu\psi_\nu$ for all $\nu \in \mathbb{N}$, where $\xi_1 \geq \xi_2 \geq \cdots > 0$, and $\lim_{\nu\to\infty} \xi_i = 0$. For any $g \in \mathcal{H}$, we have

$$Cg = \sum_{\nu=1}^{\infty} \xi_\nu \langle g, \psi_\nu\rangle_{\mathcal{H}} \psi_\nu.$$

Lemma 14 in Sriperumbudur et al. (2017) and **(A3)** implies $\overline{\text{range}(C)} = \mathcal{H}$. Thus, $\{\psi_\nu\}_{\nu\in\mathbb{N}}$ indeed forms an orthonormal basis for $\mathcal{H}$.

Since $f_0 \in \mathcal{H}$, we have

$$f_0 = \sum_{\nu=1}^{\infty} \langle f_0, \psi_\nu\rangle_{\mathcal{H}} \psi_\nu, \tag{3.10}$$

where $\{\langle f_0, \psi_\nu\rangle_{\mathcal{H}}\}_{\nu=1}^{\infty}$ is the Fourier coefficient of f_0 relative to the basis $\{\psi_\nu\}_{\nu=1}^{\infty}$. Meanwhile, under **(B7)**, we have $f_0 \in \text{range}(C^\gamma)$ and can find $g_0 \in \mathcal{H}$ such that $C^\gamma g_0 = f_0$. We then have

$$f_0 = C^\gamma g_0 = \sum_{\nu=1}^{\infty} \xi_\nu^\gamma \langle g_0, \psi_\nu\rangle_{\mathcal{H}} \psi_\nu. \tag{3.11}$$

Comparing the coefficients in (3.11) and (3.10), it follows that, for all $\nu \in \mathbb{N}$,

$$\xi_\nu^\gamma \langle g_0, \psi_\nu \rangle_{\mathcal{H}} = \langle f_0, \psi_\nu \rangle_{\mathcal{H}}.$$

Since, in addition, we can express $g_0 = \sum_{\nu=1}^{\infty} \langle g_0, \psi_\nu \rangle_{\mathcal{H}} \psi_\nu$, we have

$$\|g_0\|_{\mathcal{H}}^2 = \sum_{\nu=1}^{\infty} \langle g_0, \psi_\nu \rangle_{\mathcal{H}}^2 = \sum_{\nu=1}^{\infty} \frac{\langle f_0, \psi_\nu \rangle_{\mathcal{H}}^2}{\xi_\nu^{2\gamma}} < \infty. \tag{3.12}$$

View $\sum_{\nu=1}^{\infty} \langle f_0, \psi_\nu \rangle^2 \xi_\nu^{-2\gamma}$ as the weighted sum of $\{\xi_\nu^{-2\gamma}\}_{\nu=1}^{\infty}$ with weights $\{\langle f_0, \psi_\nu \rangle_{\mathcal{H}}^2\}_{\nu=1}^{\infty}$. Since $\{\xi_\nu\}_{\nu \in \mathbb{N}}$ are non-increasing and $\lim_{\nu \to \infty} \xi_\nu = 0$, with a large $\gamma > 0$, in order to ensure the convergence in (3.12), smaller weights must be given to the eigenfunctions with smaller eigenvalues and larger weights must be given to the eigenfunctions with larger eigenvalues, implying f_0 is smoother. Therefore, the value of γ can be viewed as a measurement of smoothness of f_0.

3.4 Comparison to Penalized SM Density Estimator

This section aims to compare the penalized and early stopping SM density estimators, where the former was first introduced by Sriperumbudur et al. (2017) and has been reviewed in **Chapter 2**.

3.4.1 Early Stopping SM Density Estimator as the Solution of a Penalized SM Loss Functional

From Theorem 3, we have, if $\hat{f}^{(0)} = 0$, $\hat{f}^{(t)} = \tau \sum_{j=0}^{t-1} (I - \tau \widehat{C})^j \hat{z}$ for all $t \in \mathbb{N}$. Then, it can be shown that $\hat{f}^{(t)}$ is the minimizer of the penalized SM loss functional $\widehat{J}_{\mathrm{SM}}(f) + \frac{1}{2}\langle f, P_t f \rangle_{\mathcal{H}}$, where $P_t : \mathcal{H} \to \mathcal{H}$ is given by

$$P_t := \left(\tau \sum_{j=0}^{t-1} (I - \tau \widehat{C})^j \right)^{-1} - \widehat{C}.$$

In particular, note that, by our choice of $\tau \in (0, 1/(d\kappa_2^2))$, the operator $\tau \sum_{j=0}^{t-1} (I - \tau \widehat{C})^j$ is invertible and P_t is well-defined.

Therefore, we observe that the early stopping SM density estimator is equivalent to the penalized SM density with a different penalty functional: the penalty functional in the penalized approach is $\frac{\rho}{2}\|f\|_{\mathcal{H}}^2 = \frac{\rho}{2}\langle f, f \rangle_{\mathcal{H}}$, whereas the penalty functional in the early stopping approach is $\frac{1}{2}\langle f, P_t f \rangle_{\mathcal{H}}$.

3.4.2 Behavior When $\rho \to 0^+$

In Theorem 5, we have shown that, if there exists a unique $x^* \in \mathcal{X}$ such that $\hat{z}_2(x^*) > \hat{z}_2(x)$ for all $x \in \mathcal{X} \backslash \{x^*\}$, $q_{\hat{f}^{(t)}}(x^*) \to \infty$ as $t \to \infty$, where $\hat{z}_2$ is the orthogonal projection of $\hat{z}$ onto $\mathrm{range}(\widehat{C})^{\perp}$.

We show a similar result for the penalized SM density estimator in the following theorem.

Theorem 9. *Let $\hat{z}_2 \in \mathcal{H}$ be the orthogonal projection of $\hat{z}$ onto* $\mathrm{range}(\widehat{C})^{\perp}$. *Under*

(A1) - ***(A4)*** *in* ***Chapter 2***, ***(B1)*** - ***(B5)***, *and* ***(C1)*** *in Theorem 5, we have* $\lim_{\rho \to 0^+} q_{\hat{f}^{(\rho)}}(x^*) = \infty$.

The proof of Theorem 9 (given in Section 3.5.9) depends on the decomposition of $\hat{f}^{(\rho)}$ into two components, which we now state.

Proposition 5. *The following identities hold*

$$\Pi_{\text{range}(\widehat{C})}(\widehat{C} + \rho I)^{-1} = (\widehat{C} + \rho I)^{-1}\Pi_{\text{range}(\widehat{C})}, \tag{3.13}$$

$$\Pi_{(\text{range}(\widehat{C}))^{\perp}}(\widehat{C} + \rho I)^{-1} = \rho^{-1}\Pi_{\text{range}(\widehat{C})^{\perp}}. \tag{3.14}$$

Then, we can write $\hat{f}^{(\rho)}$ *as the sum of the following two components*

$$\hat{f}_1^{(\rho)} := (\widehat{C} + \rho I)^{-1}\hat{z}_1 \in \text{range}(\widehat{C}),$$

$$\hat{f}_2^{(\rho)} := -\rho^{-1}\hat{z}_2 \in \text{range}(\widehat{C})^{\perp}.$$

Furthermore, we can show the boundedness of $\hat{f}_1^{(\rho)}$ over $\mathcal{X}$ as below.

Proposition 6. *Under the same assumptions as in Theorem 9, there exists a real number* $M > 0$ *such that for all* $x \in \mathcal{X}$ *and all* $\rho > 0$, $\left|\langle \hat{f}_1^{(\rho)}, k(x, \cdot)\rangle_{\mathcal{H}}\right| \leq M$.

The proofs of Propositions 5 and 6 will be given in Section 3.5.9 as well.

We now use the `waiting` variable to illustrate Theorem 9. As we have seen from Figure 3, $\hat{z}_2$ achieves the maximum at 108. With the result of Theorem 9, we expect to see that, as the value of the penalty parameter ρ keeps decreasing, the density value at 108 keeps increasing. Figure 5, the plot of the density value at 108 against

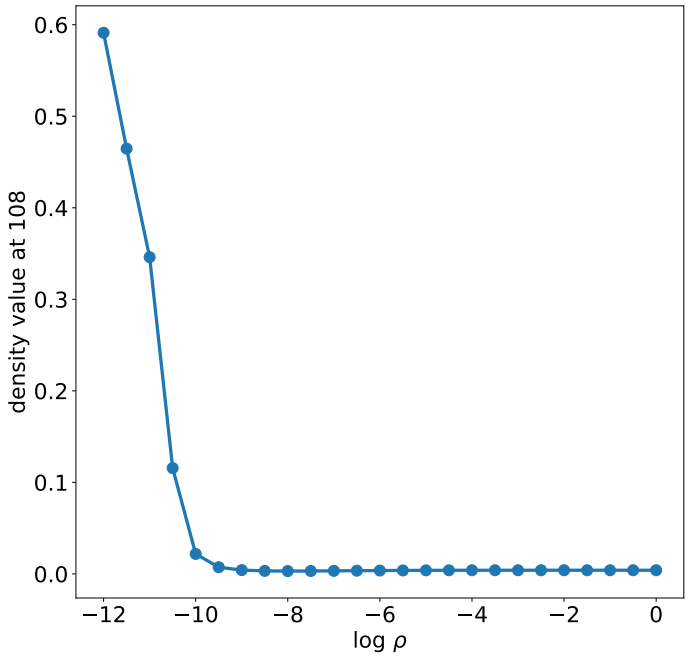

Figure 5: Density value at 108 against $\log \rho$.

$\log \rho$, confirms our expectation.

3.4.3 Comparison through Eigen-decomposition

We show the similarities of the penalized and early stopping SM density estimators through theie eigen-decomposition.

To this end, we first observe that $\widehat{C}$ has finite rank, and must be a compact operator. In addition, it is self-adjoint. By the Hilbert-Schmidt theorem, there exists a set of eigenfunctions $\{\widehat{\psi}_\nu\}_{\nu=1}^R$ that forms an orthonormal basis for range$(\widehat{C})$ and $\widehat{C}\widehat{\psi}_\nu = \hat{\xi}_\nu \widehat{\psi}_\nu$ for all $\nu = 1, \cdots, R$, where $\hat{\xi}_1 \geq \hat{\xi}_2 \geq \cdots \geq \hat{\xi}_R > 0$ are the eigenvalues of $\widehat{C}$ and $R \in \mathbb{N}$ denotes the rank of $\widehat{C}$.

By (2.13) in **Chapter 2**, we have

$$\hat{f}^{(\rho)} = (\widehat{C} + \rho I)^{-1}\hat{z} = (\widehat{C} + \rho I)^{-1}\widehat{C}\hat{g}_1 + \frac{1}{\rho}\hat{z}_2$$

$$= \sum_{\nu=1}^{R} \frac{\hat{\xi}_\nu}{\hat{\xi}_\nu + \rho} \langle \hat{g}_1, \widehat{\psi}_\nu \rangle_{\mathcal{H}} \widehat{\psi}_\nu + \frac{1}{\rho} \hat{z}_2, \tag{3.15}$$

where $\hat{g}_1 \in \mathcal{H}$ satisfying $\hat{z}_1 = \widehat{C}\hat{g}_1$. In addition, by Proposition 1, we have

$$\begin{aligned} \hat{f}^{(t)} = \tau \sum_{j=0}^{t-1} (I - \tau\widehat{C})^j \hat{z} &= (I - (I - \tau\widehat{C})^t) g_1 + t\tau\hat{z}_2 \\ &= \sum_{\nu=1}^{R} (1 - (1 - \tau\hat{\xi}_\nu)^t) \langle \hat{g}_1, \widehat{\psi}_\nu \rangle_{\mathcal{H}} \widehat{\psi}_\nu + t\tau\hat{z}_2. \end{aligned} \tag{3.16}$$

Comparing (3.15) and (3.16), we see $\hat{f}^{(\rho)}$ and $\hat{f}^{(t)}$ are very similar. The main difference lies in how they regularize the eigenvalues $\{\hat{\xi}_\nu\}_{\nu=1}^R$. In $\hat{f}^{(\rho)}$, the regularization is done through the function $x \mapsto \frac{x}{x+\rho}$, and in $\hat{f}^{(t)}$, the regularization is done through the function $x \mapsto 1 - (1 - \tau x)^t$, for $x > 0$.

For sufficiently small eigenvalues $\hat{\xi}_\nu$, $\frac{\hat{\xi}_\nu}{\hat{\xi}_\nu + \rho} \approx 0$ and $1 - (1 - \tau\hat{\xi}_\nu)^t \approx 0$. Therefore, both of them attempt to attenuate the effects of the smaller eigenvalues and let the larger eigenvalues to dominate.

3.4.4 Comparison of Convergence Rates

So far, we have been focusing more on the similarities of these two approaches. We stress a key difference of these two approaches by examining their convergence rates.

It has been shown in Sriperumbudur et al. (2017) that, under the same assumptions as in Theorem 6, we have $\|\hat{f}^{(\rho)} - f_0\|_{\mathcal{H}} = \mathcal{O}_{p_0}(n^{-\min\left\{\frac{1}{4}, \frac{\gamma}{2(\gamma+1)}\right\}})$. As we increase the value of γ (implying f_0 is smoother), the rate does *not* improve with γ and the best possible rate is $n^{-\frac{1}{4}}$, which is achieved when $\gamma = 1$. As we have discussed earlier,

the value of γ in **(B7)** indicates the degree of smoothness of f_0, and the larger the value of γ is, the smoother f_0 is. However, the rate of $\hat{f}^{(\rho)}$ does *not* really capture this smoothness, since, as soon as γ exceeds 1, the rate stabilizes at $n^{-\frac{1}{4}}$ and never improves. This unsatisfactory feature of the penalized SM density estimator is called the *saturation phenomenon* in the literature, and has been the main motivation of designing new regularized estimators that do not saturate (Engl, Hanke, and Neubauer, 1996; Bauer, Pereverzev, and Rosasco, 2007; Lo Gerfo et al., 2008).

However, Theorem 6 implies that with the stopping rule t^*, we have $\|\hat{f}^{(t)} - f_0\|_{\mathcal{H}} = \mathcal{O}_{p_0}(n^{-\frac{\gamma}{2(\gamma+2)}})$. Note, in particular, that this rate improves as γ increases. As $\gamma \to \infty$, this rate approaches to $n^{-\frac{1}{2}}$, which is faster compared to that in the penalized approach.

3.4.5 Numerical Examples

We finally compare the early stopping and penalized SM density estimators numerically using the `waiting` data in the Old Faithful Geyser dataset.

The resulting penalized and early stopping SM density estimates are shown in Figure 6. It is obvious that these two approaches yield very similar density estimates: when the regularization is large (i.e., ρ is large or t is small), density estimates are very similar to μ; as ρ decreases or t increases, density estimates become more reasonable and reveal the bimodal feature of data; if we further decrease ρ or increase t so that there is very small regularization, density estimates contain a bump or become a spike at the isolated observation 108.

Additional numerical examples confirm our observations above.

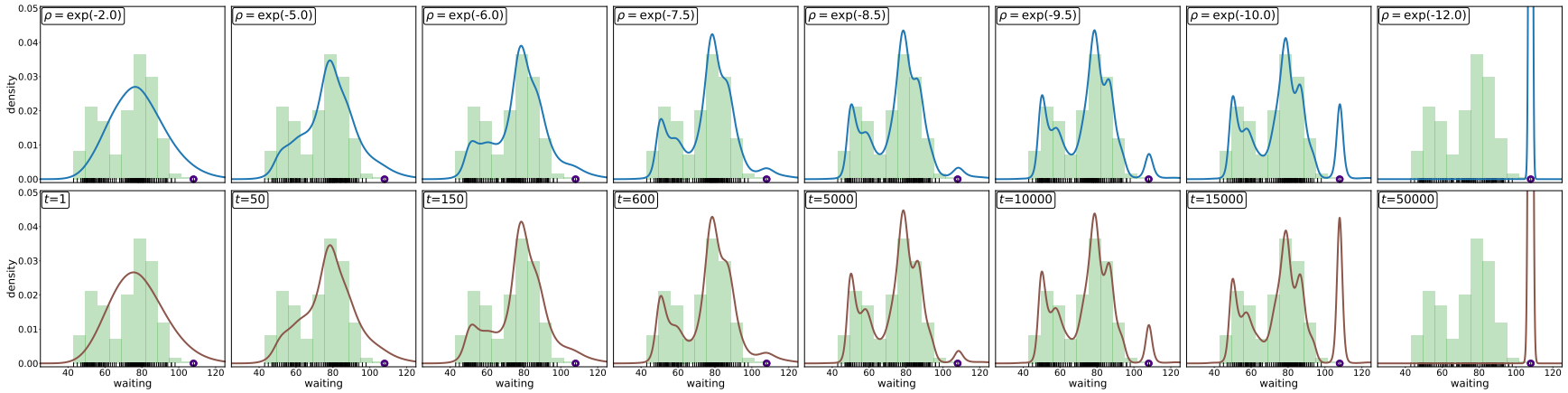

Figure 6: The penalized (first row) and early stopping (second row) SM density estimates with various choices of λ and t, respectively, shown at the upper left corner. Histogram of data using the Freedman-Diaconis rule is shown in green. The rug plot indicates the location of data and the purple circle indicates the location of the observation 108.

3.5 Auxiliary Results and Proofs

3.5.1 Proof of Theorem 2

Proof of Theorem 2. In order to establish the desired result, we first show $\widehat{J}_{\mathrm{SM}}$ is Fréchet differentiable over $\mathcal{H}$ and derive its Frechét derivative at $f \in \mathcal{H}$, and then derive the Fréchet gradient operator of $\widehat{J}_{\mathrm{SM}}$ using its definition.

Let $f \in \mathcal{H}$ be arbitrary, and $g \in \mathcal{H}$ be nonzero. It is easy to show $\widehat{C}$ is linear. Then, using its linearity, we have

$$\begin{aligned}\widehat{J}_{\mathrm{SM}}(f+g) - \widehat{J}_{\mathrm{SM}}(f) &= \left(\frac{1}{2}\langle f+g, \widehat{C}(f+g)\rangle_{\mathcal{H}} - \langle f+g, \hat{z}\rangle_{\mathcal{H}}\right) - \left(\frac{1}{2}\langle f, \widehat{C}f\rangle_{\mathcal{H}} - \langle f, \hat{z}\rangle_{\mathcal{H}}\right) \\ &= \langle g, \widehat{C}f - \hat{z}\rangle + \frac{1}{2}\langle g, \widehat{C}g\rangle_{\mathcal{H}}.\end{aligned}$$

It then follows

$$\begin{aligned}0 \leq \frac{\left|\widehat{J}(f+g) - \widehat{J}(f) - \langle g, \widehat{C}f - \hat{z}\rangle_{\mathcal{H}}\right|}{\|g\|_{\mathcal{H}}} &= \frac{|\frac{1}{2}\langle g, \widehat{C}g\rangle_{\mathcal{H}}|}{\|g\|_{\mathcal{H}}} \\ \overset{\text{(i)}}{\leq} \frac{1}{2}\left\|\widehat{C}g\right\|_{\mathcal{H}} &\overset{\text{(ii)}}{\leq} \frac{1}{2n}\sum_{i=1}^{n}\sum_{u=1}^{d}\|\partial_u k(X_i, \cdot)\|_{\mathcal{H}}^2\|g\|_{\mathcal{H}} \\ \overset{\text{(iii)}}{\leq} \frac{1}{2}d\kappa_2^2\|g\|_{\mathcal{H}} \to 0, &\qquad \text{as } \|g\|_{\mathcal{H}} \to 0,\end{aligned}$$

where we use the Cauchy-Schwartz inequality in (i), the triangle inequality and the Cauchy-Schwartz inequality in (ii), and **(B5)** in (iii). In addition, note the map $\mathrm{D}\widehat{J}_{\mathrm{SM}}(f) : \mathcal{H} \to \mathbb{R}$ defined by $\mathrm{D}\widehat{J}_{\mathrm{SM}}(f)(g) = \langle g, \widehat{C}f - \hat{z}\rangle_{\mathcal{H}}$, for all $g \in \mathcal{H}$, is linear and bounded, by **(B5)**. We conclude that $\widehat{J}_{\mathrm{SM}}$ is Frechét differentiable at $f \in \mathcal{H}$ with the

Fréchet derivative at $f \in \mathcal{H}$ being

$$\mathrm{D}\widehat{J}_{\mathrm{SM}}(f)(g) = \langle g, \widehat{C}f - \hat{z}\rangle_{\mathcal{H}}, \qquad \text{for all } g \in \mathcal{H}.$$

Since our choice of $f \in \mathcal{H}$ here is arbitrary, we conclude that $\widehat{J}_{\mathrm{SM}}$ is Frechét differentiable over $\mathcal{H}$.

Finally, since the Fréchet gradient of $\widehat{J}_{\mathrm{SM}}$ at $f \in \mathcal{H}$ is the unique element $\nabla\widehat{J}_{\mathrm{SM}}(f) \in \mathcal{H}$ satisfying $\langle \nabla\widehat{J}_{\mathrm{SM}}(f), g\rangle_{\mathcal{H}} = \mathrm{D}\widehat{J}(f)(g)$, for all $g \in \mathcal{H}$. We conclude that the Fréchet gradient of $\widehat{J}_{\mathrm{SM}}$ is the map from $\mathcal{H}$ to $\mathcal{H}$ and is given by $\nabla\widehat{J}_{\mathrm{SM}}(f) = \widehat{C}f - \hat{z} \in \mathcal{H}$ for all $f \in \mathcal{H}$. ■

3.5.2 Proof of Theorem 3

Proof of Theorem 3. We prove by induction. In the base case $t = 0$, it is straightforward from (3.3) that

$$\hat{f}^{(1)} = \hat{f}^{(0)} - \tau_0(\widehat{C}\hat{f}^{(0)} - \hat{z}) = (I - \tau_0\widehat{C})\hat{f}_0 + \tau_0\hat{z}. \tag{3.17}$$

Setting $t = 0$ in (3.4) yields

$$\hat{f}^{(1)} = \prod_{i=0}^{0}(I - \tau_i\widehat{C})\hat{f}^{(0)} + \sum_{j=0}^{0}\left[\prod_{i=j+1}^{0}(I - \tau_i\widehat{C})\right]\tau_j\hat{z} = (I - \tau_0\widehat{C})\hat{f}^{(0)} + \tau_0\hat{z},$$

which matches (3.17).

Now, we assume (3.4) holds for $t = s$ for some $s \in \mathbb{N}$, and we wish to show that it also holds for $t = s + 1$. By (3.3), we know that $\hat{f}^{(s+1)} = (I - \tau_s\widehat{C})\hat{f}^{(s)} + \tau_s\hat{z}$, and by

the inductive hypothesis, we know that $\hat{f}^{(s)} = \prod_{i=0}^{s-1}(I - \tau_i\widehat{C})\hat{f}^{(0)} + \sum_{j=0}^{s-1}\big[\prod_{i=j+1}^{s-1}(I - \tau_i\widehat{C})\big]\tau_j\hat{z}$. Hence,

$$\begin{aligned}
\hat{f}^{(s+1)} &= \big(I - \tau_s\widehat{C}\big)\left[\prod_{i=0}^{s-1}(I - \tau_i\widehat{C})\hat{f}^{(0)} + \sum_{j=0}^{s-1}\left(\prod_{i=j+1}^{s-1}(I - \tau_i\widehat{C})\right)\tau_j\hat{z}\right] + \tau_s\hat{z} \\
&= \prod_{i=0}^{s}(I - \tau_i\widehat{C})\hat{f}^{(0)} + \sum_{j=0}^{s-1}\prod_{i=j+1}^{s}(I - \tau_i\widehat{C})\tau_j\hat{z} + \prod_{i=s+1}^{s}(I - \tau_i\widehat{C})\tau_s\hat{z} \\
&= \prod_{i=0}^{s}(I - \tau_i\widehat{C})\hat{f}^{(0)} + \sum_{j=0}^{s}\left(\prod_{i=j+1}^{s}(I - \tau_i\widehat{C})\right)\tau_j\hat{z},
\end{aligned}$$

which is the desired result.

The case of the constant step size is straightforward. ■

3.5.3 Proof of Theorem 4

In order to prove Theorem 4, we need the following lemma, which is essentially a generalized version of the binomial theorem with elements in a Hilbert space.

Lemma 1. *For all $j \in \mathbb{N}_0$, we have*

$$(I - \tau\widehat{C})^j = \sum_{\ell=0}^{j}\binom{j}{\ell}(-\tau)^\ell\widehat{C}^\ell. \tag{3.18}$$

Proof of Lemma 1. We prove by induction. When $j = 0$, the left-hand side of (3.18) is simply $(I - \tau\widehat{C})^0 = I$, and the right-hand side is

$$\sum_{\ell=0}^{0}\binom{j}{\ell}(-\tau)^\ell\widehat{C}^\ell = \binom{0}{0}(-\tau)^0\widehat{C}^0 = I.$$

Hence, (3.18) holds for $j = 0$.

Now, we assume (3.18) holds for $j = s$, and we show it also holds for $j = s+1$. With $j = s+1$, the left-hand side of (3.18) becomes

$$
\begin{aligned}
(I - \tau\widehat{C})^{s+1} &= (I - \tau\widehat{C})(I - \tau\widehat{C})^{s} \\
&= (I - \tau\widehat{C})\left[\sum_{\ell=0}^{s}\binom{s}{\ell}(-\tau)^{\ell}\widehat{C}^{\ell}\right] \\
&= I\left[\sum_{\ell=0}^{s}\binom{s}{\ell}(-\tau)^{\ell}\widehat{C}^{\ell}\right] - \tau\widehat{C}\left[\sum_{\ell=0}^{s}\binom{s}{\ell}(-\tau)^{\ell}\widehat{C}^{\ell}\right] \\
&= \sum_{\ell=0}^{s}\binom{s}{\ell}(-\tau)^{\ell}\widehat{C}^{\ell} + \sum_{\ell=0}^{s}\binom{s}{\ell}(-\tau)^{\ell+1}\widehat{C}^{\ell+1} \\
&= I + \sum_{\ell=1}^{s}\binom{s}{\ell}(-\tau)^{\ell}\widehat{C}^{\ell} + \sum_{\ell=0}^{s-1}\binom{s}{\ell}(-\tau)^{\ell+1}\widehat{C}^{\ell+1} + \binom{s}{s}(-\tau)^{s+1}\widehat{C}^{s+1} \\
&= I + \sum_{\ell=1}^{s}\binom{s}{\ell}(-\tau)^{\ell}\widehat{C}^{\ell} + \sum_{\ell=1}^{s}\binom{s}{\ell-1}(-\tau)^{\ell}\widehat{C}^{\ell} + (-\tau)^{s+1}\widehat{C}^{s+1} \\
&= I + \sum_{\ell=1}^{s}\left[\binom{s}{\ell} + \binom{s}{\ell-1}\right](-\tau)^{\ell}\widehat{C}^{\ell} + (-\tau)^{s+1}\widehat{C}^{s+1} \\
&= \binom{s+1}{0}I + \sum_{\ell=1}^{s}\binom{s+1}{\ell}(-\tau)^{\ell}\widehat{C}^{\ell} + \binom{s+1}{s+1}(-\tau)^{s+1}\widehat{C}^{s+1} \\
&= \sum_{\ell=0}^{s+1}\binom{s+1}{\ell}(-\tau)^{\ell}\widehat{C}^{\ell},
\end{aligned}
$$

where we use the inductive hypothesis in the second equality. ∎

We also need the following lemma which gives an explicit characterization of $\widehat{C}^{\ell}\hat{z}$.

Lemma 2. *For all $\ell \in \mathbb{N}$, we have*

$$
\widehat{C}^{\ell}\hat{z} = \frac{1}{n^{\ell}}\sum_{i=1}^{n}\sum_{u=1}^{d}\left[\mathbf{G}^{\ell-1}\mathbf{h}\right]_{(i-1)d+u}\partial_{u}k(X_i, \cdot), \tag{3.19}
$$

where $\mathbf{G} \in \mathbb{R}^{nd\times nd}$ and $\mathbf{h} \in \mathbb{R}^{nd}$ are the same as those defined in Theorem 4.

Proof of Lemma 2. First note that, for any $\ell \geq 1$, $\widehat{C}^{\ell}\hat{z}$ resides in range$(\widehat{C})$ that contains all functions of the form

$$\widehat{C}g = \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\langle \partial_u k(X_i, \cdot), g\rangle_{\mathcal{H}}\partial_u k(X_i, \cdot), \qquad \text{for some } g \in \mathcal{H}.$$

In other words, all functions in range$(\widehat{C})$ can be written as a linear combination of $\partial_u k(X_i, \cdot)$, for all $i = 1, \cdots, n$ and $u = 1, \cdots, d$.

We prove (3.19) by induction. Note when $\ell = 1$, the left-hand side of (3.19) is

$$\begin{aligned}\widehat{C}\hat{z} =& \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\langle \partial_u k(X_i, \cdot), \hat{z}\rangle_{\mathcal{H}}\partial_u k(X_i, \cdot) = \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}[\mathbf{h}]_{(i-1)d+u}\partial_u k(X_i, \cdot)\\ =& \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}[\mathbf{G}^{1-1}\mathbf{h}]_{(i-1)d+u}\partial_u k(X_i, \cdot).\end{aligned}$$

Hence, (3.19) holds for $\ell = 1$.

Now, suppose that (3.19) holds for $\ell = s$. We want to show it also holds for $\ell = s + 1$. Note the following

$$\begin{aligned}\widehat{C}^{s+1}\hat{z} =& \widehat{C}(\widehat{C}^{s}\hat{z})\\ =& \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\partial_u k(X_i, \cdot) \otimes \partial_u k(X_i, \cdot)\right)\left(\frac{1}{n^s}\sum_{j=1}^{n}\sum_{v=1}^{d}[\mathbf{G}^{s-1}\mathbf{h}]_{(j-1)d+v}\partial_v k(X_j, \cdot)\right)\\ =& \frac{1}{n^{s+1}}\sum_{i=1}^{n}\sum_{u=1}^{d}\left(\sum_{j=1}^{n}\sum_{v=1}^{d}[\mathbf{G}^{s-1}\mathbf{h}]_{(j-1)d+v}\langle \partial_u k(X_i, \cdot), \partial_v k(X_j, \cdot)\rangle_{\mathcal{H}}\right)\partial_u k(X_i, \cdot)\\ =& \frac{1}{n^{s+1}}\sum_{i=1}^{n}\sum_{u=1}^{d}\left(\sum_{j=1}^{n}\sum_{v=1}^{d}[\mathbf{G}^{s-1}\mathbf{h}]_{(j-1)d+v}\partial_u\partial_v k(X_i, X_j)\right)\partial_u k(X_i, \cdot)\\ =& \frac{1}{n^{s+1}}\sum_{i=1}^{n}\sum_{u=1}^{d}[\mathbf{G}^{s}\mathbf{h}]_{(i-1)d+u}\partial_u k(X_i, \cdot),\end{aligned}$$

which is the desired result. ■

We now prove Theorem 4.

Proof of Theorem 4. Since, for $t \in \mathbb{N}_0$, we have $\hat{f}^{(t+1)} = \tau \sum_{j=0}^{t} (I - \tau \widehat{C})^j \hat{z}$, using the results of Lemma 1, we can rewrite it as

$$\begin{aligned}
\hat{f}^{(t+1)} &= \tau \sum_{j=0}^{t} \sum_{\ell=0}^{j} \binom{j}{\ell} (-\tau)^\ell \widehat{C}^\ell \hat{z} \\
&= \tau \hat{z} + \tau \sum_{j=1}^{t} \sum_{\ell=0}^{j} \binom{j}{\ell} (-\tau)^\ell \widehat{C}^\ell \hat{z} \\
&= \tau \hat{z} + \tau \sum_{j=1}^{t} \left[\hat{z} + \sum_{\ell=1}^{j} \binom{j}{\ell} (-\tau)^\ell \widehat{C}^\ell \hat{z} \right] \\
&= \tau(t+1)\hat{z} + \tau \sum_{j=1}^{t} \sum_{\ell=1}^{j} \binom{j}{\ell} (-\tau)^\ell \widehat{C}^\ell \hat{z}.
\end{aligned}$$

Using Lemma 2, we have

$$\begin{aligned}
\hat{f}^{(t+1)} &= \tau(t+1)\hat{z} + \tau \sum_{j=1}^{t} \sum_{\ell=1}^{j} \binom{j}{\ell} (-\tau)^\ell \left[\frac{1}{n^\ell} \sum_{i=1}^{n} \sum_{u=1}^{d} [\mathbf{G}^{\ell-1}\mathbf{h}]_{(i-1)d+u} \partial_u k(X_i, \cdot) \right] \\
&= \tau(t+1)\hat{z} + \tau \sum_{i=1}^{n} \sum_{u=1}^{d} \left[\sum_{j=1}^{t} \sum_{\ell=1}^{j} \binom{j}{\ell} \left(\frac{(-\tau)^\ell}{n^\ell} \mathbf{G}^{\ell-1}\mathbf{h} \right) \right]_{(i-1)d+u} \partial_u k(X_i, \cdot).
\end{aligned}$$

This proves (3.5).

To prove (3.6), first note the matrix $\mathbf{G}$ is the Gram matrix of the set of vectors $\{\partial_u k(X_i, \cdot)$ for $i = 1, \cdots, n$ and $u = 1, \cdots, d\}$, implying that $\mathbf{G}$ is positive semi-definite. Hence, we can write $\mathbf{G} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, where again $\mathbf{Q} \in \mathbb{R}^{nd \times nd}$ is an orthogonal matrix and $\mathbf{\Lambda} \in \mathbb{R}^{nd \times nd}$ is a diagonal matrix with the eigenvalues of $\mathbf{G}$, $\lambda_1 \geq \lambda_2 \geq$

$\cdots \geq \lambda_{nd} \geq 0$, on the diagonal. Hence,

$$\begin{aligned}\sum_{j=1}^{t}\sum_{\ell=1}^{j}\binom{j}{\ell}\left(\frac{(-\tau)^{\ell}}{n^{\ell}}\mathbf{G}^{\ell-1}\mathbf{h}\right) &= -\frac{\tau}{n}\sum_{j=1}^{t}\sum_{\ell=1}^{j}\binom{j}{\ell}\left(\left(-\frac{\tau}{n}\right)^{\ell-1}(\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\top})^{\ell-1}\mathbf{h}\right)\\ &= -\frac{\tau}{n}\sum_{j=1}^{t}\sum_{\ell=1}^{j}\binom{j}{\ell}\left(\left(-\frac{\tau}{n}\right)^{\ell-1}\mathbf{Q}\mathbf{\Lambda}^{\ell-1}\mathbf{Q}^{\top}\mathbf{h}\right)\\ &= -\frac{\tau}{n}\mathbf{Q}\left[\sum_{j=1}^{t}\sum_{\ell=1}^{j}\binom{j}{\ell}\left(-\frac{\tau}{n}\mathbf{\Lambda}\right)^{\ell-1}\right]\mathbf{Q}^{\top}\mathbf{h}. \qquad (3.20)\end{aligned}$$

Now, since $\mathbf{\Lambda}$ is a diagonal matrix, the sum and the power in (3.20) can be performed only on the diagonal elements. Since, by the binomial theorem,

$$(1+x)^{j} = 1 + \sum_{\ell=1}^{j}\binom{j}{\ell}x^{\ell}, \qquad \text{for all } x \in \mathbb{R},$$

we have, for all $w = 1, \cdots, nd$, if $\lambda_w \neq 0$,

$$\begin{aligned}\sum_{j=1}^{t}\sum_{\ell=1}^{j}\binom{j}{\ell}\left(-\frac{\tau}{n}\lambda_w\right)^{\ell-1} &= \left(-\frac{\tau}{n}\lambda_w\right)^{-1}\sum_{j=1}^{t}\left[\sum_{\ell=1}^{j}\binom{j}{\ell}\left(-\frac{\tau}{n}\lambda_w\right)^{\ell}\right]\\ &= \left(-\frac{\tau}{n}\lambda_w\right)^{-1}\sum_{j=1}^{t}\left[\left(1-\frac{\tau}{n}\lambda_w\right)^{j}-1\right]\\ &= \left(-\frac{\tau}{n}\lambda_w\right)^{-1}\left(1-\frac{\tau}{n}\lambda_w\right)\frac{1-\left(1-\frac{\tau}{n}\lambda_w\right)^{t}}{1-\left(1-\frac{\tau}{n}\lambda_w\right)} - t\left(-\frac{\tau}{n}\lambda_w\right)^{-1}\\ &= -\left(\frac{n}{\tau\lambda_w}\right)^{2}\left(1-\frac{\tau}{n}\lambda_w\right)\left(1-\left(1-\frac{\tau}{n}\lambda_w\right)^{t}\right)+\frac{tn}{\tau\lambda_w},\end{aligned}$$

and, if $\lambda_w = 0$, we have

$$\sum_{j=1}^{t}\sum_{\ell=1}^{j}\binom{j}{\ell}\left(-\frac{\tau}{n}\lambda_w\right)^{\ell-1} = \sum_{j=1}^{t} j = \frac{(t+1)t}{2},$$

which are exactly the diagonal elements of $\widetilde{\mathbf{\Lambda}}$ defined in the theorem. Therefore, we have

$$\sum_{j=1}^{t}\sum_{\ell=1}^{j}\binom{j}{\ell}\left(\frac{(-\tau)^{\ell}}{n^{\ell}}\mathbf{G}^{\ell-1}\mathbf{h}\right)=-\frac{\tau}{n}\mathbf{Q}\widetilde{\mathbf{\Lambda}}\mathbf{Q}^{\top}\mathbf{h},$$

In addition, we have

$$\begin{aligned}\hat{f}^{(t+1)}=&\tau(t+1)\hat{z}+\tau\sum_{i=1}^{n}\sum_{u=1}^{d}\left[-\frac{\tau}{n}\mathbf{Q}\widetilde{\mathbf{\Lambda}}\mathbf{Q}^{\top}\mathbf{h}\right]_{(i-1)d+u}\partial_u k(X_i,\,\cdot\,)\\=&\tau(t+1)\hat{z}-\frac{\tau^2}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\left[\mathbf{Q}\widetilde{\mathbf{\Lambda}}\mathbf{Q}^{\top}\mathbf{h}\right]_{(i-1)d+u}\partial_u k(X_i,\,\cdot\,),\end{aligned}$$

which completes the proof. ■

3.5.4 Proofs of Results in Section 3.3.1

Proof of Proposition 1. Since we can write $\hat{z}=\hat{z}_1+\hat{z}_2$ with $\hat{z}_1=\widehat{C}\hat{g}_1\in\mathrm{range}(\widehat{C})$ and $\hat{z}_2\in\mathrm{range}(\widehat{C})^{\perp}$, we can write the t-th gradient descent iterate as

$$\hat{f}^{(t+1)}=(I-\tau\widehat{C})\hat{f}^{(t)}+\tau(\hat{z}_1+\hat{z}_2)=(I-\tau\widehat{C})\hat{f}^{(t)}+\tau\widehat{C}\hat{g}_1+\tau\hat{z}_2.$$

Subtracting both sides of the preceding equation by $\hat{g}_1$ yields

$$\hat{f}^{(t+1)}-\hat{g}_1=(I-\tau\widehat{C})(\hat{f}^{(t)}-g_1)+\tau\hat{z}_2. \tag{3.21}$$

Projecting both sides of (3.21) onto range$(\widehat{C})$ and range$(\widehat{C})^{\perp}$, respectively, we obtain

$$\begin{aligned}\Pi_{\text{range}(\widehat{C})}(\hat{f}^{(t+1)}-\hat{g}_1) &= \Pi_{\text{range}(\widehat{C})}\big((I-\tau\widehat{C})(\hat{f}^{(t)}-\hat{g}_1)+\tau\hat{z}_2\big)\\ &= (I-\tau\widehat{C})\Pi_{\text{range}(\widehat{C})}(\hat{f}^{(t)}-\hat{g}_1),\end{aligned}$$

and

$$\Pi_{\text{range}(\widehat{C})^{\perp}}(\hat{f}^{(t+1)}-\hat{g}_1) = \Pi_{\text{range}(\widehat{C})^{\perp}}(\hat{f}^{(t)}-\hat{g}_1)+\tau\hat{z}_2.$$

We then have

$$\begin{aligned}\Pi_{\text{range}(\widehat{C})}(\hat{f}^{(t+1)}-\hat{g}_1) &= (I-\tau\widehat{C})^{t+1}\Pi_{\text{range}(\widehat{C})}(\hat{f}^{(0)}-\hat{g}_1)\\ &= (I-\tau\widehat{C})^{t+1}\Pi_{\text{range}(\widehat{C})}(-\hat{g}_1)\\ &= -(I-\tau\widehat{C})^{t+1}\hat{g}_1,\end{aligned}$$

where we use $\hat{f}^{(0)}=0$ to obtain the second equality and $\hat{g}_1 \in \text{range}(\widehat{C})$ to obtain the last equality, and

$$\Pi_{\text{range}(\widehat{C})^{\perp}}(\hat{f}^{(t+1)}-\hat{g}_1) = \Pi_{\text{range}(\widehat{C})^{\perp}}(\hat{f}^{(0)}-\hat{g}_1)+(t+1)\tau\hat{z}_2 = (t+1)\tau\hat{z}_2,$$

where we again use $\hat{f}^{(0)}=0$ and $\hat{g}_1 \in \text{range}(\widehat{C})$.

Finally, we have

$$\hat{f}^{(t+1)} = \hat{g}_1 + \Pi_{\text{range}(\widehat{C})}(\hat{f}^{(t+1)}-\hat{g}_1) + \Pi_{\text{range}(\widehat{C})^{\perp}}(\hat{f}^{(t+1)}-\hat{g}_1)$$

$$= \hat{g}_1 - (I - \tau\widehat{C})^{t+1}\hat{g}_1 + \tau(t+1)\hat{z}_2$$

$$= (I - (I - \tau\widehat{C})^{t+1})\hat{g}_1 + \tau(t+1)\hat{z}_2,$$

which is the desired result. ■

Proof of Proposition 2. (a) Since $\hat{z}_1$ is the orthogonal projection of $\hat{z}$ onto range$(\widehat{C})$, by the definition of projection, we have

$$\hat{z}_1 = \underset{w \in \text{range}(\widehat{C})}{\arg\min} \|\hat{z} - w\|_{\mathcal{H}}^2.$$

Since $w \in \text{range}(\widehat{C})$, we can write $w = \sum_{i=1}^{n}\sum_{u=1}^{d} \alpha_{(i-1)d+u}\partial_u k(X_i, \cdot)$ for some $\boldsymbol{\alpha} := (\alpha_1, \cdots, \alpha_{nd})^\top \in \mathbb{R}^{nd}$. Then,

$$\begin{aligned}
\|\hat{z} - w\|_{\mathcal{H}}^2 &= \left\|\hat{z} - \sum_{i=1}^{n}\sum_{u=1}^{d} \alpha_{(i-1)d+u}\partial_u k(X_i, \cdot)\right\|_{\mathcal{H}}^2 \\
&= \|\hat{z}\|_{\mathcal{H}}^2 - 2\sum_{i=1}^{n}\sum_{u=1}^{d} \alpha_{(i-1)d+u}\langle \hat{z}, \partial_u k(X_i, \cdot)\rangle_{\mathcal{H}} + \left\|\sum_{i=1}^{n}\sum_{u=1}^{d} \alpha_{(i-1)d+u}\partial_u k(X_i, \cdot)\right\|_{\mathcal{H}}^2 \\
&= \|\hat{z}\|_{\mathcal{H}}^2 - 2\boldsymbol{\alpha}^\top \mathbf{h} + \boldsymbol{\alpha}^\top \mathbf{G}\boldsymbol{\alpha}.
\end{aligned}$$

Since $\mathbf{G}$ is a Gram matrix and is positive semi-definite, the function $\boldsymbol{\alpha} \mapsto \|\hat{z}\|_{\mathcal{H}}^2 - 2\boldsymbol{\alpha}^\top \mathbf{h} + \boldsymbol{\alpha}^\top \mathbf{G}\boldsymbol{\alpha}$ is convex and differentiable in $\boldsymbol{\alpha}$. Then, $\boldsymbol{\alpha}^* := \arg\min_{\boldsymbol{\alpha}}\{\|\hat{z}\|_{\mathcal{H}}^2 - 2\boldsymbol{\alpha}^\top \mathbf{h} + \boldsymbol{\alpha}^\top \mathbf{G}\boldsymbol{\alpha}\}$ must satisfy the first-order optimality condition $\mathbf{0} = -\mathbf{h} + \mathbf{G}\boldsymbol{\alpha}^*$, which is the desired linear system.

(b) The result is straightforward using the relationship $\hat{z} = \hat{z}_1 + \hat{z}_2$, the definition of $\hat{z}$, the result from (a).

(c) Since $\hat{g}_1$ belongs to range$(\widehat{C})$ and satisfies the relationship $\hat{z}_1 = \widehat{C}\hat{g}_1$, we let $\hat{g}_1 = \sum_{j=1}^{n}\sum_{v=1}^{d}\beta^*_{(j-1)d+v}\partial_j k(X_b, \cdot)$ and must have

$$\begin{aligned}\hat{z}_1 = \widehat{C}g_1 &= \left[\frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\partial_u k(X_i, \cdot) \otimes \partial_u k(X_i, \cdot)\right]\left(\sum_{j=1}^{n}\sum_{v=1}^{d}\beta^*_{(j-1)d+v}\partial_v k(X_j, \cdot)\right) \\ &= \sum_{i=1}^{n}\sum_{u=1}^{d}\left[\frac{1}{n}\sum_{j=1}^{n}\sum_{v=1}^{d}\beta^*_{(j-1)d+v}\partial_u\partial_{v+d}k(X_i, X_j)\right]\partial_u k(X_i, \cdot).\end{aligned}$$

It follows that, for all $i = 1, \cdots, n$ and $u = 1, \cdots, d$,

$$\alpha^*_{(i-1)d+u} = \frac{1}{n}\sum_{j=1}^{n}\sum_{v=1}^{d}\beta^*_{(j-1)d+v}\partial_u\partial_{v+d}k(X_i, X_j).$$

In other words, $\boldsymbol{\beta}^* := (\beta^*_1, \cdots, \beta^*_{nd})^\top \in \mathbb{R}^{nd}$ must satisfy the linear system $\frac{1}{n}\mathbf{G}\boldsymbol{\beta}^* = \boldsymbol{\alpha}^*$. ■

Proof of Proposition 3. Recall from Proposition 1 that $\hat{f}_1^{(t+1)} = (I - (I - \tau\widehat{C})^t)\hat{g}_1$. Then, note the following

$$\begin{aligned}|\langle \hat{f}_1^{(t+1)}, k(x, \cdot)\rangle_{\mathcal{H}}| &= \left|\langle (I - (I - \tau\widehat{C})^t)\hat{g}_1, k(x, \cdot)\rangle_{\mathcal{H}}\right| \\ &\overset{\text{(i)}}{\leq} \|I - (I - \tau\widehat{C})^t\|\|\hat{g}_1\|_{\mathcal{H}}\|k(x, \cdot)\|_{\mathcal{H}} \\ &\overset{\text{(ii)}}{\leq} \|\hat{g}_1\|_{\mathcal{H}}\sup_{x\in\mathcal{X}}\|k(x, \cdot)\|_{\mathcal{H}} \\ &\overset{\text{(iii)}}{\leq} \kappa_1\|\hat{g}_1\| =: M.\end{aligned}$$

In the development above, (i) follows from the Cauchy-Schwartz inequality and the sub-multiplicative property and the definition of the operator norm. Due to **(C2)**,

$\|I - (I - \tau\widehat{C})^t\| \leq 1$ for all t and (ii) holds. Finally, the inequality (iii) follows from **(A2)**. ∎

We are now ready to prove Theorem 5.

Proof of Theorem 5. Notice that

$$q_{\hat{f}^{(t)}}(x^*) = \frac{\mu(x^*)\exp\big(\langle (I - (I - \tau\widehat{C})^t)g_1, k(x^*, \cdot)\rangle_{\mathcal{H}} + \tau t\hat{z}_2(x^*)\big)}{\int_{\mathcal{X}} \mu(x)\exp\big(\langle (I - (I - \tau\widehat{C})^t)g_1, k(x, \cdot)\rangle_{\mathcal{H}} + \tau t\hat{z}_2(x)\big)\mathrm{d}x}.$$

Using Proposition 3, we can bound $q_{\hat{f}^{(t)}}(x^*)$ from below by

$$q_{\hat{f}^{(t)}}(x^*) \geq \frac{\mu(x^*)\exp\big(-M + t\tau\hat{z}_2(x^*)\big)}{\int_{\mathcal{X}} \mu(x)\exp\big(M + t\tau\hat{z}_2(x)\big)\mathrm{d}x} = \frac{\exp(-2M)\mu(x^*)}{\int_{\mathcal{X}} \mu(x)\frac{\exp(t\tau\hat{z}_2(x))}{\exp(t\tau\hat{z}_2(x^*))}\mathrm{d}x}, \tag{3.22}$$

where the last equality follows by dividing both the numerator and the denominator by $\exp(t\tau\hat{z}_2(x^*))$.

Since $\hat{z}_2(x^*) \geq \hat{z}_2(x)$ for all $x \in \mathcal{X}$, it follows that

$$\frac{\exp(t\tau\hat{z}_2(x))}{\exp(t\tau\hat{z}_2(x^*))} = \left(\frac{\exp(\tau\hat{z}_2(x))}{\exp(\tau\hat{z}_2(x^*))}\right)^t \leq 1,$$

and the equality holds if and only if $x = x^*$. Then, for all $x \in \mathcal{X}$ and all $t \in \mathbb{N}$,

$$\left|\mu(x)\left(\frac{\exp(\tau\hat{z}_2(x))}{\exp(\tau\hat{z}_2(x^*))}\right)^t\right| \leq \mu(x).$$

In addition, since μ is a pdf over over $\mathcal{X}$ by **(A4)**, an application of Lebesgue's

dominated convergence theorem yields

$$\lim_{t\to\infty}\int_{\mathcal{X}}\mu(x)\left(\frac{\exp(\tau\hat{z}_2(x))}{\exp(\tau\hat{z}_2(x^*))}\right)^t\mathrm{d}x = \int_{\mathcal{X}}\mu(x)\lim_{t\to\infty}\left(\frac{\exp(\tau\hat{z}_2(x))}{\exp(\tau\hat{z}_2(x^*))}\right)^t\mathrm{d}x$$
$$= \int_{\mathcal{X}}\mu(x)\mathbb{1}_{\{x^*\}}(x)\,\mathrm{d}x = 0,$$

where $\mathbb{1}_S$ is the indicator function for the set S.

Taking the limit of $t \to \infty$ of both sides of (3.22), we have

$$\lim_{t\to\infty} q_{\hat{f}^{(t)}}(x^*) \geq \frac{\exp(-2M)\mu(x^*)}{\lim_{t\to\infty}\int_{\mathcal{X}}\mu(x)\left(\frac{\exp(\tau\hat{z}_2(x))}{\exp(\tau\hat{z}_2(x^*))}\right)^t\mathrm{d}x} = \infty,$$

since the numerator is strictly positive by **(A4)** and the denominator approaches to 0 as $t \to \infty$. ∎

3.5.5 Proof of Theorem 6

Proof of Theorem 6. Based on the inequality (3.9) and Theorems 7 and 8, we know that with probability at least $1-\delta$ for $\delta \in (0,1)$, the following inequality holds

$$\|\hat{f}^{(t)} - f_0\|_{\mathcal{H}} \leq C_1\frac{t^2}{\sqrt{n}} + C_2 t^{-\gamma}, \tag{3.23}$$

where C_1 and C_2 are constants defined in the statement of the theorem and come from the upper bounds of the sample error and the approximation error, respectively. Let the stopping rule be $t^*(n) = \lceil n^\beta \rceil$, the smallest integer greater than or equal to n^β for some $\beta > 0$. Plugging this $t^*(n)$ into the RHS of (3.23), we essentially obtain

the following function of β,

$$h(\beta) := C_1 n^{2\beta - 1/2} + C_2 n^{-\gamma\beta}.$$

By elementary calculus, h achieves the minimum at $\beta^* := \frac{1}{2(\gamma+2)}$.

As a consequence, assuming $n^{\beta^*} \leq t^*(n) = \lceil n^{\beta^*} \rceil = \eta n^{\beta^*} \leq n^{\beta^*} + 1$ for some $\eta \in [1, 2]$, we have

$$\begin{aligned}
\|\hat{f}^{(t)} - f_0\|_{\mathcal{H}} \leq & C_1 \frac{(\eta n^{\beta^*})^2}{\sqrt{n}} + C_2 (\eta n^{\beta^*})^{-\gamma} \\
\leq & C_1 \eta^2 n^{2\beta^* - \frac{1}{2}} + C_2 \eta^{-\gamma} n^{-\gamma\beta^*} \\
\leq & (4C_1 + C_2) n^{-\frac{\gamma}{2(\gamma+2)}}.
\end{aligned}$$

This completes the proof. ■

3.5.6 Proof of Theorem 7

In order to prove Theorem 7, we first need the following two lemmas: one describes the relationship between f_0 and z, which is one part of Theorem 4(ii) by Sriperumbudur et al. (2017), and the other one gives an expression of $f^{(t)} - f_0$ that is easier to work with.

Lemma 3 (Sriperumbudur et al., 2017, Theorem 4(ii)). *Under* ***(A1)*** - ***(A4)*** *in* ***Chapter 2*** *and* ***(B1)*** - ***(B7)****, we have* $Cf_0 = z$.

Lemma 4. *Assume* $f^{(0)} = 0$. *For all* $t \in \mathbb{N}_0$, $f^{(t)} - f_0 = -(I - \tau C)^t f_0$.

Proof of Lemma 4. Note, when $t = 0$, the desired result holds obviously.

Now, let $t \in \mathbb{N}$. By the relationship $f^{(t)} = f^{(t-1)} - \tau(Cf^{(t-1)} - z)$ and Lemma 3, we have

$$f^{(t)} = f^{(t-1)} - \tau(Cf^{(t-1)} - Cf_0) = f^{(t-1)} - \tau C(f^{(t-1)} - f_0).$$

Subtracting both sides by f_0 yields

$$\begin{aligned} f^{(t)} - f_0 = f^{(t-1)} - f_0 - \tau C(f^{(t-1)} - f_0) &= (I - \tau C)(f^{(t-1)} - f_0) \\ &= (I - \tau C)^t(f^{(0)} - f_0). \end{aligned}$$

The desired result follows since $f^{(0)} = 0$. ■

We now prove Theorem 7.

Proof of Theorem 7. By Lemma 4, we have $\|f^{(t)} - f_0\|_{\mathcal{H}} = \|(I - \tau C)^t f_0\|_{\mathcal{H}}$. By the relationship $C^\gamma g_0 = f_0$ in **(B7)**, we have

$$\|f^{(t)} - f_0\|_{\mathcal{H}} = \|(I - \tau C)^t C^\gamma g_0\|_{\mathcal{H}} \leq \|(I - \tau C)^t C^\gamma\| \|g_0\|_{\mathcal{H}}.$$

Next, we need to find an upper bound for the operator norm of $(I - \tau C)^t C^\gamma$. Since C is a compact self-adjoint operator (Theorem 4(i) in Sriperumbudur et al., 2017), with an application of the Hilbert-Schmidt Theorem, we have,

$$Cf = \sum_{\nu=1}^{\infty} \xi_\nu \langle f, \psi_\nu \rangle_{\mathcal{H}} \psi_\nu, \qquad \text{for all } f \in \mathcal{H},$$

where $\{\psi_\nu\}_{\nu\in\mathbb{N}}$ are the eigenvectors of C that form an orthonormal basis of $\overline{\text{range}(C)}$ and $\{\xi_\nu\}_{\nu\in\mathbb{N}}$ are corresponding eigenvalues satisfying $\lim_{\nu\to\infty}\xi_\nu = 0$ and $C\psi_\nu = \xi_\nu\psi_\nu$ for all $\nu \in \mathbb{N}$. It follows that

$$\begin{aligned}\|(I-\tau C)^t C^\gamma\| &\le \sup_\nu\big\{(1-\tau\xi_\nu)^t\xi_\nu^\gamma\big\} = \sup_\nu \exp\big(t\log(1-\tau\xi_\nu)+\gamma\log\xi_\nu\big)\\ &\le \sup_\nu \exp\big(-t\tau\xi_\nu+\gamma\log\xi_\nu\big),\end{aligned} \tag{3.24}$$

where the last inequality follows from the basic inequality $\log(1+x) \le x$ for all $x > -1$. Also, note that $\log(1-\tau\xi_\nu)$ is well-defined since $\tau < 1/(d\kappa_2^2) \le 1/\|C\|$ so that $\tau\xi_\nu < 1$ for all $\nu \in \mathbb{N}$.

Define the function $h(x) = -\tau t x + \gamma\log x$ for all $x > 0$ and we maximize h. By elementary calculus, h achieves the maximum at $x^* = \frac{\gamma}{\tau t} > 0$ and the maximum value is $h(x^*) = -\gamma + \gamma\log(\frac{\gamma}{\tau t})$. Plugging $h(x^*)$ back to (3.24), we have

$$\|(I-\tau C)^t C^\gamma\| \le \left(\frac{\gamma}{\tau e t}\right)^\gamma, \tag{3.25}$$

and the desired result follows. ∎

3.5.7 Proof of Theorem 8

In order to prove Theorem 8, we first need the following proposition that gives an equivalent expression of $\hat{f}^{(t)}$ that links operators $\widehat{C}$ and C.

Proposition 7. *The sample-version gradient descent updates can be written as*

$$\hat{f}^{(t+1)} = (I - \tau C)^{t+1}\hat{f}^{(0)} + \tau \sum_{j=0}^{t} (I - \tau C)^{t-j}\big((C - \widehat{C})\hat{f}_j + \hat{z}\big).$$

Proof of Proposition 7. We start from (3.3) and note

$$\begin{aligned}
\hat{f}^{(t+1)} &= (I - \tau\widehat{C})\hat{f}^{(t)} + \tau\hat{z} \\
&= (I - \tau C + \tau C - \tau\widehat{C})\hat{f}^{(t)} + \tau\hat{z} \\
&= (I - \tau C)\hat{f}^{(t)} + \tau((C - \widehat{C})\hat{f}^{(t)} + \hat{z}).
\end{aligned}$$

Now, we obtain a non-homogeneous linear first-order difference equation. The rest can be proved by induction and is similar to the proof of Theorem 3, which we omit. ∎

We also need the following lemma, which is an extension of the Hoeffding inequality to random variables in a Hilbert space. It will help us to obtain upper bounds for $\|\widehat{C} - C\|$ and $\|\hat{z} - z\|_{\mathcal{H}}$.

Lemma 5 (Hoeffding-Pinelis Inequality)**.** *Let $\{W_i\}_{i=1}^n$ be an independent random sequence of mean zero in a separable Hilbert space $\mathcal{H}$ with the norm $\|\cdot\|_{\mathcal{H}}$ such that, for all $i = 1, \cdots, n$, $\|W_i\|_{\mathcal{H}} \le c_i < \infty$ almost surely. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} W_i\right\|_{\mathcal{H}} \ge \epsilon\right) \le 2\exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^{n} c_i^2}\right).$$

Equivalently, with probability at least $1-\delta$ *where* $\delta \in (0,1)$*, we have*

$$\left\| \sum_{i=1}^{n} W_i \right\|_{\mathcal{H}} \leq \sqrt{2\left(\sum_{i=1}^{n} c_i^2\right) \log\left(\frac{2}{\delta}\right)}.$$

If, in particular, $c_i \equiv c > 0$ *for all* i*, then with probability at least* $1-\delta$ *for* $\delta \in (0,1)$*, we have*

$$\frac{1}{n}\left\| \sum_{i=1}^{n} W_i \right\|_{\mathcal{H}} \leq c\sqrt{\frac{2}{n} \log\left(\frac{2}{\delta}\right)}.$$

In order to use Lemma 5 to bound $\|\widehat{C} - C\|$ and $\|\hat{z} - z\|_{\mathcal{H}}$, we need to show the (almost surely) boundedness of several quantities, which we now state and prove.

Lemma 6. *Under* ***(B1)*** *-* ***(B5)****, the operator* $C : \mathcal{H} \to \mathcal{H}$ *defined in* (3.1) *is a Hilbert-Schmidt operator with* $\|C\|_{\text{HS}} \leq d\kappa_2^2$*, where* $\|\cdot\|_{\text{HS}}$ *denotes the Hilbert-Schmidt norm, and the function* z *defined in* (3.2) *satisfies* $\|z\|_{\mathcal{H}} \leq d(\kappa_3 + \kappa_4)$*.*

A brief introduction to the Hilbert-Schmidt operator can be found in **Section A1.4 in Appendix 1**.

Proof of Lemma 6. In the proof of Theorem 4 in Sriperumbudur et al. (2017), they have shown that the operator $C : \mathcal{H} \to \mathcal{H}$ is of trace class and

$$\text{trace}(C) = \int_{\mathcal{X}} p_0(x) \sum_{u=1}^{d} \partial_u \partial_{u+d} k(x,x) \mathrm{d}x \overset{\text{(i)}}{\leq} d\kappa_2^2 < \infty,$$

where (i) follows from **(B5)**. By the relationship between the Hilbert-Schmidt norm and the trace (see **Proposition 10(c) in Appendix 1**), we conclude that $\|C\|_{\text{HS}} \leq$

$\mathrm{trace}(C) \leq d\kappa_2^2$.

As for the result on z, using **Proposition 4(b) in Appendix 1**, we have

$$\|z\|_{\mathcal{H}} \leq \sum_{u=1}^{d} \int_{\mathcal{X}} p_0(x) \Big(|\partial_u \log \mu(x)| \|\partial_u k(x, \cdot)\|_{\mathcal{H}} + \|\partial_u^2 k(x, \cdot)\|_{\mathcal{H}} \Big) \mathrm{d}x \overset{\text{(ii)}}{\leq} d(\kappa_3 + \kappa_4),$$

where the inequality (ii) is again due to **(B5)**. ■

Lemma 7. *Under* ***(B1)*** *-* ***(B5)****, for all* $i = 1, \cdots, n$*, the operator* $\widehat{C}_i : \mathcal{H} \to \mathcal{H}$ *defined by*

$$\widehat{C}_i := \sum_{u=1}^{d} \partial_u k(X_i, \cdot) \otimes \partial_u k(X_i, \cdot) \tag{3.26}$$

is a Hilbert-Schmidt operator with $\|\widehat{C}_i\|_{\mathrm{HS}} \leq d\kappa_2^2$*, and the function* $\hat{z}_i$ *defined by*

$$\hat{z}_i := -\sum_{u=1}^{d} \big(\partial_u \log \mu(X_i) \partial_u k(X_i, \cdot) + \partial_u^2 k(X_i, \cdot) \big) \in \mathcal{H} \tag{3.27}$$

satisfies $\|\hat{z}_i\|_{\mathcal{H}} \leq d(\kappa_3 + \kappa_4)$*. In addition, we have* $\|\widehat{C}\|_{\mathrm{HS}} \leq d\kappa_2^2$ *and* $\|\hat{z}\|_{\mathcal{H}} \leq d(\kappa_3 + \kappa_4)$*.*

Proof of Lemma 7. We first show $\widehat{C}_i$ is a Hilbert-Schmidt operator. Pick an arbitrary orthonormal basis $\{e_\ell\}_{\ell \in \mathbb{N}}$ of $\mathcal{H}$ and an arbitrary $i \in \{1, \cdots, n\}$. Then, note the following

$$\begin{aligned} \|\widehat{C}_i\|_{\mathrm{HS}}^2 = \sum_{\ell=1}^{\infty} \|\widehat{C}_i e_\ell\|_{\mathcal{H}}^2 &= \sum_{\ell=1}^{\infty} \left\| \sum_{u=1}^{d} \partial_u e_\ell(X_i) \partial_u k(X_i, \cdot) \right\|_{\mathcal{H}}^2 \\ &\overset{\text{(i)}}{\leq} d \sum_{\ell=1}^{\infty} \sum_{u=1}^{d} |\partial_u e_\ell(X_i)|^2 \|\partial_u k(X_i, \cdot)\|_{\mathcal{H}}^2 \end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(ii)}}{=} d \sum_{u=1}^{d} \left[\sum_{\ell=1}^{\infty} \left| \langle \partial_u k(X_i, \cdot), e_\ell \rangle_{\mathcal{H}} \right|^2 \right] \|\partial_u k(X_i, \cdot)\|_{\mathcal{H}}^2 \\
&\overset{\text{(iii)}}{=} d \sum_{u=1}^{d} \|\partial_u k(X_i, \cdot)\|_{\mathcal{H}}^2 \|\partial_u k(X_i, \cdot)\|_{\mathcal{H}}^2 \\
&\overset{\text{(iv)}}{=} d \sum_{u=1}^{d} (\partial_u \partial_{u+d} k(X_i, X_i))^2 \\
&\leq d^2 \kappa_2^4 < \infty,
\end{aligned}
$$

and hence, $\|\widehat{C}_i\|_{\text{HS}} \leq d\kappa_2^2$. In the derivation above, (i) is due to the Cauchy-Schwartz inequality. We interchange the sums in (ii) due to the Tonelli-Fubini Theorem. An application of the Parseval's identity yields the equality (iii). We use the reproducing property in (iv). We use **(B5)** in the last inequality.

As for the result on $\hat{z}_i$, for any $i \in \{1, \cdots, n\}$, we use **(B6)** and have

$$
\|\hat{z}_i\|_{\mathcal{H}} \leq \sum_{u=1}^{d} \Big(|\partial_u \log \mu(X_i)| \|\partial_u k(X_i, \cdot)\|_{\mathcal{H}} + \|\partial_u^2 k(X_i, \cdot)\|_{\mathcal{H}} \Big) \leq d(\kappa_3 + \kappa_4).
$$

The bounds on $\|\widehat{C}\|_{\text{HS}}$ and $\|\hat{z}\|_{\mathcal{H}}$ easily follows by the triangle inequality. ∎

We now use Lemmas 5 - 7 to derive upper bounds of $\|\widehat{C} - C\|$ in Proposition 8 and $\|\hat{z} - z\|_{\mathcal{H}}$ in Proposition 9.

Proposition 8. *Under the same assumptions in Theorem 8, with probability at least* $1 - \delta$ *for* $\delta \in (0, 1)$,

$$
\|\widehat{C} - C\| \leq 2d\kappa_2^2 \sqrt{\frac{2}{n} \log\left(\frac{2}{\delta}\right)}.
$$

Proof. Let $\widehat{C}_i : \mathcal{H} \to \mathcal{H}$ be the operator defined in (3.26). We have $\widehat{C} = \frac{1}{n}\sum_{i=1}^n \widehat{C}_i$ and $\mathbb{E}[\widehat{C}_i] = \mathbb{E}[\widehat{C}] = C$ for all $i = 1, \cdots, n$. Define $W_i := \widehat{C}_i - C$. It follows that $\{W_i\}_{i=1}^n$ is a sequence of independent random variables in $\mathcal{H}$ with zero mean and $\widehat{C} - C = \frac{1}{n}\sum_{i=1}^n W_i$.

Notice that, for all $i = 1, \cdots, n$, by Lemma 10(a) in **Appendix 1**,

$$\|W_i\| = \|\widehat{C}_i - C\| \leq \|\widehat{C}_i - C\|_{\text{HS}} \leq \|\widehat{C}_i\|_{\text{HS}} + \|C\|_{\text{HS}} \leq 2d\kappa_2^2.$$

Then, the desired result follows from Lemma 10(b) in **Appendix 1** and Lemma 5. ∎

Proposition 9. *Under the same assumptions in Theorem 8, with probability at least* $1 - \delta$ *for* $\delta \in (0, 1)$,

$$\|\hat{z} - z\|_{\mathcal{H}} \leq 2d(\kappa_3 + \kappa_4)\sqrt{\frac{2}{n}\log\left(\frac{2}{\delta}\right)}.$$

Proof. If we let $z_i \in \mathcal{H}$ be (3.27), we have $\hat{z} = \frac{1}{n}\sum_{i=1}^n z_i$ and $\mathbb{E}[z_i] = \mathbb{E}[\hat{z}] = z$. Define the random variable $W_i = \hat{z}_i - z$ for all $i = 1, \cdots, n$. It follows that $\{W_i\}_{i=1}^n$ is an sequence of independent random variable in $\mathcal{H}$ with zero mean. Also, we have $\hat{z} - z = \frac{1}{n}\sum_{i=1}^n W_i$.

We now verify the (almost surely) boundedness of W_i for all $i = 1, \cdots, n$. Applying the triangle inequality, we have

$$\|W_i\|_{\mathcal{H}} = \|\hat{z}_i - z\|_{\mathcal{H}} \leq \|\hat{z}_i\|_{\mathcal{H}} + \|z\|_{\mathcal{H}} \leq 2d(\kappa_3 + \kappa_4) < \infty.$$

The desired result follows directly from Lemma 5. ■

In order to prove Theorem 8, we also need an upper bound of $\hat{f}^{(t)}$ for all $t \in \mathbb{N}_0$, which is established in the following proposition.

Proposition 10. *Let $\hat{f}^{(0)} = 0$. Under the same assumptions in Theorem 8, for any $t \in \mathbb{N}_0$, we have $\|\hat{f}^{(t)}\|_{\mathcal{H}} \leq \tau t d(\kappa_3 + \kappa_4)$.*

Proof. By Theorem 3 and $\hat{f}^{(0)} = 0$, we have $\hat{f}^{(t)} = \tau \sum_{j=0}^{t-1} (I - \tau \widehat{C})^{t-j-1} \hat{z}$. Then, by the triangle inequality, the sub-multiplicative property of norm, the assumptions in Theorem 8, and Lemma 7, we have

$$\|\hat{f}^{(t)}\|_{\mathcal{H}} \leq \tau \sum_{j=0}^{t-1} \|(I - \tau \widehat{C})^{t-j-1} \hat{z}\|_{\mathcal{H}} \leq \tau t \|\hat{z}\|_{\mathcal{H}} \leq \tau t d(\kappa_3 + \kappa_4).$$

■

Now, we prove Theorem 8.

Proof of Theorem 8. Let $t \geq 1$. With $f^{(0)} = \hat{f}^{(0)} = 0$, we have

$$\begin{aligned}
\hat{f}^{(t)} - f^{(t)} &= \tau \sum_{j=0}^{t-1} (I - \tau C)^{t-j-1} ((C - \widehat{C}) f^{(j)} + \hat{z}) - \tau \sum_{j=0}^{t-1} (I - \tau C)^{t-j-1} z \\
&= \tau \sum_{j=0}^{t-1} (I - \tau C)^{t-j-1} \big((C - \widehat{C}) \hat{f}^{(j)} + \hat{z} - z\big).
\end{aligned}$$

Then, we can bound the norm of the preceding difference as follows

$$\begin{aligned}\|\hat{f}^{(t)} - f^{(t)}\|_{\mathcal{H}} &\overset{\text{(i)}}{\leq} \tau \sum_{j=0}^{t-1} \|I - \tau C\|^{t-j-1} \Big(\|C - \widehat{C}\| \|\hat{f}^{(j)}\|_{\mathcal{H}} + \|\hat{z} - z\|_{\mathcal{H}} \Big) \\ &\overset{\text{(ii)}}{\leq} \tau \bigg(\|\widehat{C} - C\| \cdot \sum_{j=0}^{t-1} \|\hat{f}^{(j)}\|_{\mathcal{H}} + t \|\hat{z} - z\|_{\mathcal{H}} \bigg),\end{aligned}$$

where (i) follows from the triangle inequality, the definition and the sub-mulplicative property of the operator norm, and (ii) follows from $\|I - \tau C\| \leq 1$.

By Proposition 10, we have, for all $t \in \mathbb{N}_0$, $\|\hat{f}^{(t)}\|_{\mathcal{H}} \leq \tau t d(\kappa_3 + \kappa_4)$, and thus,

$$\begin{aligned}\|\hat{f}^{(t)} - f^{(t)}\|_{\mathcal{H}} \leq & \tau \bigg(\|\widehat{C} - C\| \sum_{j=0}^{t-1} \tau j d(\kappa_3 + \kappa_4) + t \|\hat{z} - z\|_{\mathcal{H}} \bigg) \\ = & \tau \bigg(\|\widehat{C} - C\| \frac{t(t-1)}{2} \tau d(\kappa_3 + \kappa_4) + t \|\hat{z} - z\|_{\mathcal{H}} \bigg) \\ \leq & \tau t^2 \big(\|\widehat{C} - C\| \tau d(\kappa_3 + \kappa_4) + \|\hat{z} - z\|_{\mathcal{H}} \big).\end{aligned}$$

Using Propositions 8 and 9, we have, with probability at least $1 - \delta$ for $\delta \in (0, 1)$,

$$\|\widehat{C} - C\| \leq 2 d \kappa_2^2 \sqrt{\frac{2}{n} \log\Big(\frac{2}{\delta}\Big)}, \quad \text{and} \quad \|\hat{z} - z\|_{\mathcal{H}} \leq 2 d (\kappa_3 + \kappa_4) \sqrt{\frac{2}{n} \log\Big(\frac{2}{\delta}\Big)}.$$

It follows that, with probability at least $1 - \delta$ for $\delta \in (0, 1)$,

$$\|\hat{f}^{(t)} - f^{(t)}\|_{\mathcal{H}} \leq 2 d \tau t^2 (\kappa_3 + \kappa_4)(\tau d \kappa_2^2 + 1) \sqrt{\frac{2}{n} \log\Big(\frac{2}{\delta}\Big)}.$$

■

3.5.8 Proof of Corollary 1

In order to prove Corollary 1, we need the following lemma (Lemma A.1 in Sriperumbudur et al., 2017).

Lemma 8. *Let $L^\infty(\mathcal{X})$ denote the class of bounded measurable functions $f : \mathcal{X} \to \mathbb{R}$ endowed with the uniform norm $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$, and $\mathcal{P}_\infty$ contain all pdfs over $\mathcal{X}$ of the form*

$$q_f(x) := \mu(x) \exp\big(f(x) - A(f)\big) \text{ for all } x \in \mathcal{X}, \qquad f \in L^\infty(\mathcal{X}).$$

Then, for any $q_f, q_g \in \mathcal{P}_\infty$, we have the following

(a) in terms of the KL-divergence, there exists a universal constant $c > 0$ such that

$$\mathrm{KL}(q_f \| q_g) \le c e^{\|f-g\|_\infty} \|f - g\|_\infty^2 (1 + \|f - g\|_\infty);$$

(b) in terms of the Hellinger distance, we have

$$\mathrm{He}(q_f \| q_g) \le e^{\frac{1}{2}\|f-g\|_\infty} \|f - g\|_\infty;$$

(c) in terms of the L^1 distance,

$$\|q_f - q_g\|_{L^1} \le 2e^{2\|f-g\|_\infty} \|f - g\|_\infty.$$

Proof of Corollary 1. (a) By Theorem 4(i) in Sriperumbudur et al. (2017), we know

that $\mathrm{H}(p^*\|q_f) = \frac{1}{2}\langle f - f_0, C(f-f_0)\rangle_{\mathcal{H}}$. In addition, by Lemma 6 and Lemma 10 in **Appendix 1**, we know $\|C\| \le \|C\|_{\mathrm{HS}} \le d\kappa_2^2$. Thus, we have

$$\mathrm{H}(p_0\|q_{\hat{f}^{(t^*)}}) \le \frac{1}{2}\|C\|\|\hat{f}^{(t^*)} - f_0\|_{\mathcal{H}}^2 \le \underbrace{\frac{1}{2}d\kappa_2^2(4C_1 + C_2)^2}_{=:C_3} n^{-\frac{\gamma}{\gamma+2}}.$$

(b) First note that, for any $f, g \in \mathcal{H}$,

$$\|f-g\|_\infty = \sup_{x\in\mathcal{X}}|f(x) - g(x)| \le \sup_{x\in\mathcal{X}}\sqrt{k(x,x)}\|f-g\|_{\mathcal{H}} \le \kappa_1\|f-g\|_{\mathcal{H}} < \infty.$$

By Lemma 8(a), there exists a universal constant $c > 0$ such that

$$\mathrm{KL}(p_0\|q_{\hat{f}^{(t^*)}}) \le c\kappa_1^2 e^{\kappa_1\|\hat{f}^{(t^*)} - f_0\|_{\mathcal{H}}}\|\hat{f}^{(t^*)} - f_0\|_{\mathcal{H}}^2(1 + \kappa_1\|\hat{f}^{(t^*)} - f_0\|_{\mathcal{H}}).$$

Since

$$\exp(\kappa_1\|\hat{f}^{(t^*)} - f_0\|_{\mathcal{H}}) \le \exp(\kappa_1(4C_1 + C_2)n^{-\frac{\gamma}{2(\gamma+2)}}) \le \exp(\kappa_1(4C_1 + C_2)),$$

and $1 + \kappa_1\|\hat{f}^{(t^*)} - f_0\|_{\mathcal{H}} \le 1 + \kappa_1(4C_1 + C_2)$, we have

$$\mathrm{KL}(p_0\|q_{\hat{f}^{(t^*)}}) \le \underbrace{c\kappa_1^2\exp(\kappa_1(4C_1 + C_2))(1 + \kappa_1(4C_1 + C_2))(4C_1 + C_2)^2}_{=:C_4} n^{-\frac{\gamma}{\gamma+2}}.$$

(c) By Lemma 8(b), we have

$$\mathrm{He}(p_0\|q_{\hat{f}^{(t^*)}}) \le \exp\left(\frac{1}{2}\kappa_1(4C_1 + C_2)n^{-\frac{\gamma}{2(\gamma+2)}}\right)\kappa_1(4C_1 + C_2)n^{-\frac{\gamma}{2(\gamma+2)}}$$

$$\leq \underbrace{\exp\left(\frac{1}{2}\kappa_1(4C_1+C_2)\right)\kappa_1(4C_1+C_2)}_{=:C_5} n^{-\frac{\gamma}{2(\gamma+2)}}.$$

(d) By Lemma 8(c), we have

$$\begin{aligned}
\|p_0 - q_{\hat{f}^{(t^*)}}\|_{L^1} &\leq 2\kappa_1 \exp\big(2\kappa_1 \|\hat{f}^{(t^*)} - f_0\|_{\mathcal{H}}\big)\|\hat{f}^{(t^*)} - f_0\|_{\mathcal{H}} \\
&\leq 2\kappa_1 \exp\big(2\kappa_1(4C_1+C_2)n^{-\frac{\gamma}{2(\gamma+2)}}\big)(4C_1+C_2)n^{-\frac{\gamma}{2(\gamma+2)}} \\
&\leq \underbrace{2\kappa_1(4C_1+C_2)\exp(2\kappa_1(4C_1+C_2))}_{=:C_6} n^{-\frac{\gamma}{2(\gamma+2)}}.
\end{aligned}$$

■

3.5.9 Proof of Results in Section 3.4

Proof of Proposition 5. In order to show (3.13) and (3.14), it is sufficient to show $\Pi_{\text{range}(\widehat{C})}(\widehat{C}+\rho I) = (\widehat{C}+\rho I)\Pi_{\text{range}(\widehat{C})}$ and $(\widehat{C}+\rho I)\Pi_{\text{range}(\widehat{C})^\perp} = \rho\Pi_{\text{range}(\widehat{C})^\perp}$. Note the following

$$\begin{aligned}
\Pi_{\text{range}(\widehat{C})}(\widehat{C}+\rho I) &= \Pi_{\text{range}(\widehat{C})}(\widehat{C}) + \Pi_{\text{range}(\widehat{C})}(\rho I) \\
&\overset{(\star)}{=} \widehat{C}\Pi_{\text{range}(\widehat{C})} + \rho\Pi_{\text{range}(\widehat{C})} \\
&= (\widehat{C}+\rho I)\Pi_{\text{range}(\widehat{C})},
\end{aligned}$$

where $(\star)$ follows from the definition of $\Pi_{\text{range}(\widehat{C})}$.

As for the other, notice $\Pi_{\text{range}(\widehat{C})^\perp}(\widehat{C}) = (I - \Pi_{\text{range}(\widehat{C})})\widehat{C} = \widehat{C} - \widehat{C} = 0$, and, hence,

$$\Pi_{\text{range}(\widehat{C})^\perp}(\widehat{C}+\rho I) = \Pi_{\text{range}(\widehat{C})^\perp}\widehat{C} + \rho\Pi_{\text{range}(\widehat{C})^\perp} = \rho\Pi_{\text{range}(\widehat{C})^\perp}.$$

Finally, to show the decomposition of $\hat{f}^{(\rho)}$, we have

$$\begin{aligned}
\hat{f}^{(\rho)} &= \big(\Pi_{\text{range}(\widehat{C})} + \Pi_{\text{range}(\widehat{C})^{\perp}}\big)(\widehat{C} + \rho I)^{-1}(\hat{z}_1 + \hat{z}_2) \\
&= \Pi_{\text{range}(\widehat{C})}(\widehat{C} + \rho I)^{-1}(\hat{z}_1 + \hat{z}_2) + \Pi_{\text{range}(\widehat{C})^{\perp}}(\widehat{C} + \rho I)^{-1}(\hat{z}_1 + \hat{z}_2) \\
&\overset{\text{(i)}}{=} (\widehat{C} + \rho I)^{-1}\Pi_{\text{range}(\widehat{C})}(\hat{z}_1 + \hat{z}_2) + \rho^{-1}\Pi_{\text{range}(\widehat{C})^{\perp}}(\hat{z}_1 + \hat{z}_2) \\
&\overset{\text{(ii)}}{=} (\widehat{C} + \rho I)^{-1}\hat{z}_1 + \rho^{-1}\hat{z}_2,
\end{aligned}$$

where we use (3.13) and (3.14) in (i) and the definitions of $\hat{z}_1$ and $\hat{z}_2$ in (ii). ∎

Proof of Proposition 6. First, note by Proposition 5 and $\hat{z}_1 = \widehat{C}\hat{g}_1$, we have $\hat{f}_1^{(\rho)} = (\widehat{C} + \rho I)^{-1}\widehat{C}\hat{g}_1$.

Since $\widehat{C}$ has finite rank, it must be a compact operator. In addition, it is self-adjoint. The Hilbert-Schmidt theorem guarantees that, for any $f \in \mathcal{H}$,

$$\widehat{C}f = \sum_{\nu=1}^{R} \hat{\xi}_\nu \langle f, \widehat{\psi}_\nu \rangle_{\mathcal{H}} \widehat{\psi}_\nu,$$

where $\hat{\xi}_1 \geq \hat{\xi}_2 \geq \cdots \geq \hat{\xi}_R > 0$ are the eigenvalues of $\widehat{C}$, $\{\psi_\nu\}_{\nu=1}^{R}$ are the corresponding eigenfunctions that form an orthonormal basis for range$(\widehat{C})$, and R denotes the rank of $\widehat{C}$. Then, we have

$$\hat{f}_1^{(\rho)} = (\widehat{C} + \rho I)^{-1}\widehat{C}\hat{g}_1 = \sum_{\nu=1}^{R} \frac{\hat{\xi}_\nu}{\hat{\xi}_\nu + \rho} \langle \hat{g}_1, \widehat{\psi}_\nu \rangle_{\mathcal{H}} \widehat{\psi}_\nu.$$

Hence,

$$\begin{aligned}\left|\langle \hat{f}_1^{(\rho)}, k(x,\,\cdot\,)\rangle_{\mathcal{H}}\right| &= \left|\left\langle \sum_{\nu=1}^{R} \frac{\hat{\xi}_\nu}{\hat{\xi}_\nu+\rho}\langle \hat{g}_1, \widehat{\psi}_\nu\rangle_{\mathcal{H}}\widehat{\psi}_\nu, k(x,\,\cdot\,)\right\rangle_{\mathcal{H}}\right| \\ &\overset{\text{(i)}}{\leq} \sum_{\nu=1}^{R} \frac{\hat{\xi}_\nu}{\hat{\xi}_\nu+\rho}\|\hat{g}_1\|_{\mathcal{H}}\|\widehat{\psi}_\nu\|_{\mathcal{H}}^2\|k(x,\,\cdot\,)\|_{\mathcal{H}} \\ &\overset{\text{(ii)}}{\leq} R\kappa_1\|\hat{g}_1\| =: M < \infty.\end{aligned}$$

We apply the triangle inequality and the Cauchy-Schwartz inequality in (i). Since $\frac{\hat{\xi}_\nu}{\hat{\xi}_\nu+\rho} \in (0,1)$ for all $\nu = 1, \cdots, R$ and $\{\varphi_\ell\}_{\ell=1}^R$ are orthonormal, we obtain the inequality (ii). Finally, the boundedness follows from **(A2)**. ∎

Proof of Theorem 9. By Proposition 5, we know, for any $x \in \mathcal{X}$,

$$\langle \hat{f}^{(\rho)}, k(x,\,\cdot\,)\rangle_{\mathcal{H}} = \langle (\widehat{C}+\rho I)^{-1}\hat{z}_1, k(x,\,\cdot\,)\rangle_{\mathcal{H}} + \rho^{-1}\langle \hat{z}_2, k(x,\,\cdot\,)\rangle_{\mathcal{H}},$$

and, hence,

$$q_{\hat{f}^{(\rho)}}(x^*) = \frac{\mu(x^*)\exp\big(\langle (\widehat{C}+\rho I)^{-1}\hat{z}_1, k(x^*,\,\cdot\,)\rangle_{\mathcal{H}} + \rho^{-1}\langle \hat{z}_2, k(x^*,\,\cdot\,)\rangle_{\mathcal{H}}\big)}{\int_{\mathcal{X}} \mu(x)\exp\big(\langle (\widehat{C}+\rho I)^{-1}\hat{z}_1, k(x,\,\cdot\,)\rangle_{\mathcal{H}} + \rho^{-1}\langle \hat{z}_2, k(x,\,\cdot\,)\rangle_{\mathcal{H}}\big)\mathrm{d}x}.$$

By Proposition 6, we can bound $q_{\hat{f}^{(\rho)}}(x^*)$ from below by

$$\begin{aligned} q_{\hat{f}^{(\rho)}}(x^*) &\geq \frac{\mu(x^*)\exp(-M + \rho^{-1}\langle \hat{z}_2, k(x^*,\,\cdot\,)\rangle_{\mathcal{H}})}{\int_{\mathcal{X}} \mu(x)\exp(M + \rho^{-1}\langle \hat{z}_2, k(x,\,\cdot\,)\rangle_{\mathcal{H}})\,\mathrm{d}x} \\ &= \frac{\mu(x^*)\exp(-2M)}{\int_{\mathcal{X}} \mu(x)\frac{\exp(\rho^{-1}\langle \hat{z}_2, k(x,\,\cdot\,)\rangle_{\mathcal{H}})}{\exp(\rho^{-1}\langle \hat{z}_2, k(x^*,\,\cdot\,)\rangle_{\mathcal{H}})}\,\mathrm{d}x} \end{aligned} \tag{3.28}$$

where the last equality follows by dividing both the numerator and the denominator

by $\exp(M + \rho^{-1}\langle \hat{z}_2, k(x^*, \cdot)\rangle_{\mathcal{H}})$. Define

$$R_\rho(x) := \frac{\exp(\rho^{-1}\langle \hat{z}_2, k(x, \cdot)\rangle_{\mathcal{H}})}{\exp(\rho^{-1}\langle \hat{z}_2, k(x^*, \cdot)\rangle_{\mathcal{H}})} = \left(\frac{\exp(\langle \hat{z}_2, k(x, \cdot)\rangle_{\mathcal{H}})}{\exp(\langle \hat{z}_2, k(x^*, \cdot)\rangle_{\mathcal{H}})}\right)^{\frac{1}{\rho}},$$

and $R(x) := (R_\rho(x))^\rho$. Since $\hat{z}_2(x^*) \geq \hat{z}_2(x)$ for all $x \in \mathcal{X}$, it follows that $R(x) \leq 1$ for all $x \in \mathcal{X}$ with the equality being held if and only if $x = x^*$.

We examine the limit of $R_\rho(x)$ as $\rho \to 0^+$ and consider the following two cases:

Case 1: if $x = x^*$, $R(x^*) = 1$ for all $\rho > 0$ and it follows that $\lim_{\rho\to 0^+} R_\rho(x^*) = 1$;

Case 2: if $x \neq x^*$, $R(x^*) < 1$ and it follows that $\lim_{\rho\to 0^+} R_\rho(x) = 0$.

Therefore, $\lim_{\rho\to 0^+} R_\rho(x) = \mathbb{1}_{\{x^*\}}(x)$, where $\mathbb{1}_S$ is the indicator function for the set S.

Since, for all $x \in \mathcal{X}$ and any $\rho > 0$, $|\mu(x)R_\rho(x)| \leq \mu(x)$, and that μ is a pdf over $\mathcal{X}$, we can swap the limit and the integral and obtain

$$\lim_{\rho\to 0^+} \int_{\mathcal{X}} \mu(x)R_\rho(x)\,\mathrm{d}x = \int_{\mathcal{X}} \mu(x) \lim_{\rho\to 0^+} R_\rho(x)\,\mathrm{d}x = 0.$$

To obtain the desired result, taking the limit of $\rho \to 0^+$ of both sides in (3.28), we have

$$\lim_{\rho\to 0^+} q_{\hat{f}^{(\rho)}}(x^*) \geq \frac{\mu(x^*)\exp(-2M)}{\lim_{\rho\to 0^+}\int_{\mathcal{X}} \mu(x)R_\rho(x)\,\mathrm{d}x} = \infty,$$

since the numerator is strictly positive by **(A4)** and the denominator approaches to 0 as $\rho \to 0^+$. The desired result follows. ■

References

Bauer, Frank, Sergei Pereverzev, and Lorenzo Rosasco (Feb. 2007). "On regularization algorithms in learning theory". In: *J. Complex.* 23.1, pp. 52–72.

Boyd, Stephen and Lieven Vandenberghe (Mar. 2004). *Convex Optimization*. en. Cambridge University Press.

Brockwell, Peter J and Richard A Davis (Nov. 2013). *Time Series: Theory and Methods*. en. Springer Science & Business Media.

Bühlmann, Peter and Bin Yu (June 2003). "Boosting with the L^2 Loss". In: *J. Am. Stat. Assoc.* 98.462, pp. 324–339.

Caponnetto, A and E De Vito (Aug. 2006). "Optimal Rates for the Regularized Least-Squares Algorithm". en. In: *Found. Comut. Math.* 7.3, pp. 331–368.

Engl, Heinz Werner, Martin Hanke, and A Neubauer (July 1996). *Regularization of Inverse Problems*. en. Springer Science & Business Media.

Lin, Junhong et al. (Oct. 2018). "Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces". In: *Appl. Comput. Harmon. Anal.*

Lo Gerfo, L et al. (July 2008). "Spectral algorithms for supervised learning". en. In: *Neural Comput.* 20.7, pp. 1873–1897.

Raskutti, Garvesh, Martin J. Wainwright, and Bin Yu (2014). "Early Stopping and Non-parametric Regression: An Optimal Data-dependent Stopping Rule". In: *Journal of Machine Learning Research* 15.11, pp. 335–366.

Rastogi, Abhishake and Sivananthan Sampath (2017). "Optimal Rates for the Regularized Learning Algorithms under General Source Condition". In: *Frontiers in Applied Mathematics and Statistics* 3, p. 3.

Royden, H. L. and P.M. Fitzpatrick (2018). *Real Analysis*. eng. Fourth edition [2018 reissue]. Pearson modern classic. New York, NY: Pearson. ISBN: 9780134689494.

Smale, Steve and Ding-Xuan Zhou (Mar. 2007). "Learning Theory Estimates via Integral Operators and Their Approximations". en. In: *Constr. Approx.* 26.2, pp. 153–172.

Sriperumbudur, Bharath et al. (2017). "Density Estimation in Infinite Dimensional Exponential Families". In: *Journal of Machine Learning Research* 18.57, pp. 1–59. URL: http://jmlr.org/papers/v18/16-011.html.

Tsybakov, Alexandre B (2009). *Introduction to Nonparametric Estimation*. Springer series in statistics. Dordrecht: Springer.

Yao, Yuan, Lorenzo Rosasco, and Andrea Caponnetto (Aug. 2007). "On Early Stopping in Gradient Descent Learning". In: *Constr. Approx.* 26.2, pp. 289–315.

Zhang, Tong and Bin Yu (Aug. 2005). "Boosting with early stopping: Convergence and consistency". en. In: *aos* 33.4, pp. 1538–1579.