On Nonparametric Density Estimation in Kernel Exponential Families and the Sensitivity of Density Estimators

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Chenxi Zhou, B.S.

Graduate Program in Department of Statistics

The Ohio State University

2022

Dissertation Committee:

Vincent Q. Vu, Advisor

Yoonkyung Lee

Sebastian A. Kurtek

Abstract

This dissertation is concerned with the nonparametric density estimation problem in a kernel exponential family, which is an exponential family induced by a reproducing kernel Hilbert space (RKHS). The corresponding density estimation problem can be formulated as a convex minimization problem over a RKHS or a subset of it. The loss functionals we focus on are the negative log-likelihood (NLL) loss functional and the score matching (SM) loss functional.

We propose a new density estimator called the early stopping SM density estimator, which is obtained by applying the gradient descent algorithm to minimizing the SM loss functional and terminating the algorithm early. We investigate various statistical properties of this density estimator. We also compare this early stopping SM density estimator with the penalized SM density estimator that has been studied in the literature and address their similarities and differences.

In addition, we propose an algorithm to compute the penalized maximum likelihood (ML) density estimator that is obtained by minimizing the penalized NLL loss functional. We empirically compare the penalized and early stopping SM density estimators with the penalized ML density estimator and find out that when there is a small amount of regularization (corresponding to small values of the penalty parameter or large values of the number of iterations), the regularized SM density estimates contain a bump or become a spike at the isolated observation, but the penalized ML

density estimates do not. Moreover, if we remove the isolated observation, the resulting regularized SM density estimates do not contain a bump or a spike when the regularization is small. We attempt to explain why this happens.

Observations above motivate us to study the sensitivities of different density estimators to the presence of an additional observation. We extend the definition of the influence function by allowing its input to be function-valued statistical functionals. We study various properties of this extended influence functions of ML and SM (log-)density projections in finite-dimensional and kernel exponential families, and empirically demonstrate that regularized SM density estimators in a kernel exponential family are more sensitive to the presence of an additional observation than the penalized ML density estimator when the amount of regularization is small.

To my family

Vita

May 2015	B.S. Mathematics and Economics, The Ohio State University
August 2015 — May 2019	Graduate Teaching and Research Associate, Department of Statistics, The Ohio State University
August 2019 — present	Graduate Research Associate, Nationwide Center for Advanced Customer Insights, Fisher College of Business, The Ohio State University

Fields of Study

Major Field: Statistics

Table of Contents

Page

Abstract . . . ii

Dedication . . . iv

Vita . . . v

List of Figures . . . x

1. Introduction . . . 1

    1.1 A Review of Density Estimation Methods . . . 2
        1.1.1 Parametric Approach . . . 2
        1.1.2 Nonparametric Approach . . . 3
            1.1.2.1 Nonparametric Maximum Likelihood Density Estimation . . . 4
                1.1.2.1.1 Penalized Maximum Likelihood Density Estimation . . . 5
                1.1.2.1.2 Shape-constrained Maximum Likelihood Density Estimation . . . 8
            1.1.2.2 Nonparametric Score Matching Density Estimation . . . 12
    1.2 Nonparametric Density Estimation in Kernel Exponential Families . . . 14
    1.3 Organization of the Remaining Dissertation . . . 16

2. Kernel Exponential Family and Density Estimation Problem in It . . . 18

    2.1 Kernel Exponential Families . . . 18
        2.1.1 A Review of Finite-dimensional Exponential Family . . . 18
        2.1.2 Kernel Exponential Family . . . 19
        2.1.3 Properties of $\mathcal{Q}_{\text{ker}}$ . . . 20
            2.1.3.1 Characterization of $\mathcal{F}$ for Bounded Kernels . . . 20
            2.1.3.2 Convexity of A . . . 21
            2.1.3.3 Differential Properties of A . . . 22
        2.1.4 Connection to Finite-dimensional Exponential Families . . . 25

2.1.5 Assumptions on $\mathcal{H}$ and k and Their Implications 31
2.2 Nonparametric Density Estimation in $\mathcal{Q}_{\text{ker}}$ 32
2.2.1 Density Estimation in $\mathcal{Q}_{\text{ker}}$ using $\widehat{L}_{\text{NLL}}$ 32
2.2.2 Density estimation in $\mathcal{Q}_{\text{ker}}$ using $\widehat{L}_{\text{SM}}$ 36
2.3 Proofs 40
2.3.1 Proof of Proposition 2.1 40
2.3.2 Proof of Proposition 2.2 40
2.3.3 Proof of Lemma 2.1 41
2.3.4 Proof of Proposition 2.3 42
2.3.5 Proof of Proposition 2.5 44

3. Early Stopping Score Matching Density Estimator 46

3.1 An Overview 46
3.2 Early Stopping SM Density Estimator 48
3.2.1 Computation of $\hat{f}^{(t)}$ 52
3.2.2 Numerical Examples of Early Stopping SM Density Estimators 54
3.2.3 When to Terminate the Algorithm 56
3.3 Theoretical Properties of Early Stopping SM Density Estimator . . 58
3.3.1 Limiting SM Density Estimator as $t \to \infty$ 60
3.3.1.1 Decomposition of $\hat{f}^{(t)}$ 60
3.3.1.2 Numerical Illustration of Theorem 3.5 63
3.3.2 Rate of Convergence 63
3.3.2.1 An Upper Bound on the Approximation Error 66
3.3.2.2 An Upper Bound on the Sample Error 67
3.3.2.3 Upper Bounds on the Distances between p_0 and $q_{\hat{f}^{(t^*(n))}}$ 67
3.3.2.4 Discussion on **(B7)** 68
3.4 Comparison to Penalized SM Density Estimator 70
3.4.1 Early Stopping SM Density Estimator as the Solution of a Penalized SM Loss Functional 70
3.4.2 Behavior When $\rho \to 0^+$ 70
3.4.3 Comparison through Eigen-decomposition 72
3.4.4 Comparison of Convergence Rates 73
3.4.5 Numerical Examples 74
3.5 Auxiliary Results and Proofs 76
3.5.1 Proof of Theorem 3.2 76
3.5.2 Proof of Theorem 3.3 77
3.5.3 Proof of Theorem 3.4 77
3.5.4 Proofs of Results in Section 3.3.1 82
3.5.5 Proof of Theorem 3.6 85
3.5.6 Proof of Theorem 3.7 86
3.5.7 Proof of Theorem 3.8 88
3.5.8 Proof of Corollary 3.1 93

3.5.9 Proof of Results in Section 3.4 . . . 95

4. Comparison of Regularized ML and SM Density Estimators in $\mathcal{Q}_{\text{ker}}$ . . . 99

4.1 Penalized ML Density Estimator . . . 100
4.1.1 Failure of the Representer Theorem . . . 100
4.1.2 Construction of a Finite-dimensional Approximating Space . 102
4.1.3 Computation of the Minimizer of the Penalized NLL Loss Functional . . . 105
4.1.3.1 Batch Monte Carlo Approximation of $\nabla \widetilde{A}(\boldsymbol{\beta})$ . . . 106
4.1.3.2 Gradient Descent Algorithm to Minimize $\widetilde{J}_{\text{NLL},\lambda}$ . . . 109
4.1.4 Numerical Illustration . . . 109
4.2 Regularized SM Density Estimators with $f \in \widetilde{\mathcal{H}}$ . . . 111
4.2.1 Penalized SM Density Estimator with $f \in \widetilde{\mathcal{H}}$ . . . 112
4.2.2 Early Stopping SM Density Estimator with $f \in \widetilde{\mathcal{H}}$ . . . 113
4.3 Comparison of Regularized ML and SM Density Estimators . . . 114
4.4 Discussion on the Presence of a Spike in SM Density Estimates . . . 117
4.5 Proofs . . . 119
4.5.1 Proof of Proposition 4.1 . . . 119
4.5.2 Details about Example 4.1 . . . 119
4.5.3 Proof of Proposition 4.2 . . . 123
4.5.4 Proof of Proposition 4.3 . . . 124
4.5.5 Proof of Proposition 4.4 . . . 125

5. Influence Function of a (Log-)Density Function and Its Properties . . . 127

5.1 Influence Function and Its Applications in Statistics . . . 127
5.2 Extension of the Influence Function in Density Estimation Problem 132
5.3 Influence Function of (Log-)Density Projection in a Finite-dimensional Exponential Family . . . 136
5.4 Influence Function of (Log-)Density Projection in a Kernel Exponential Family . . . 142
5.5 Proofs . . . 145
5.5.1 Proof of Proposition 5.1 . . . 145
5.5.2 Proof of Proposition 5.2 . . . 145
5.5.3 Proof of Theorem 5.1 . . . 146
5.5.4 Proof of Theorem 5.2 . . . 147

6. Numerical Studies of the Sensitivities of Penalized ML and SM (Log-)Density Estimators in $\mathcal{Q}_{\text{ker}}$ . . . 151

6.1 Comparison of the Sensitivities of Penalized ML and SM Density Estimators . . . 151
6.1.1 Computation of the Sample Influence Function . . . 152

6.1.2 Comparison of the Sample Influence Functions of Log-density and Density Estimators . . . 154
6.1.3 Comparison of the Sensitivities . . . 160
6.2 The Sensitivity of K-fold Cross-validated Penalized SM Density Estimator . . . 164
6.3 Which One to Use: Penalized ML or Regularized SM Density Estimators? . . . 166

7. Summary and Future Directions . . . 168

7.1 Summary . . . 168
7.2 Future Directions . . . 169

Appendices 172

A. Math Background . . . 172

A.1 Fréchet Differentiability and Derivative . . . 172
A.2 Bochner Integral . . . 175
A.3 Partial Derivative of a Kernel Function . . . 177
A.4 Some Theories on Bounded Linear Operators . . . 178

List of Figures

Figure **Page**

1.1 Penalized SM density estimates of the `waiting` variable with (first row) and without (second row) the isolated observation 108 (indicated by the purple circle). Histogram of the `waiting` variable with the bin width selected by the Freedman-Diaconis rule (Freedman and Diaconis, 1981) is shown in green. 15

3.1 Left panel shows μ and right panel shows $\log \mu$. The rug plot indicates the location of data. 55

3.2 Early stopping SM density estimates for different values of number of iterations labeled at the upper left corner. Histogram of data with the bin width chosen by 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. 56

3.3 Plots of $\hat{z}$ (left panel), $\hat{z}_1$ (middle panel) and $\hat{z}_2$ (right panel). The rug plot indicates the location of data. 63

3.4 Density value at 108 against the number of iterations. 64

3.5 Density value at 108 against $\log \rho$. 72

3.6 The penalized (first row) and early stopping (second row) SM density estimates with various choices of ρ and t, respectively, shown at the upper left corner. Histogram of data with the bin width chosen by 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. 75

4.1 Left panel: the minimum of $\widetilde{J}_{\mathrm{NLL},\lambda}$ against the gap between two adjacent points at which kernel functions are centered in different choices of finite-dimensional approximating subspace. Different opacity indicates different values of λ, and the more opaque line indicates the smaller λ value. Right panel: the minimum of $\widetilde{J}_{\mathrm{NLL},\lambda}$ against different values of $\log \lambda$. 110

4.2 Penalized ML (first row), penalized SM (second row), and early stopping SM (third row) density estimates of the `waiting` variable. Histogram of data with the bin width chosen by 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 isolated observation 108. 115

4.3 Penalized ML (first row), penalized SM (second row), and early stopping SM (third row) density estimates of the `waiting` variable with the isolated observation 108 removed. Histogram of data with the bin width chosen by the Freedman-Diaconis rule is shown in green. The rug plot indicates the location of data. 116

5.1 $\mathrm{IF}_x(T, F, y)$ (left panel) and $\mathrm{IF}_x(\widetilde{T}, F, y)$ (right panel) evaluated at different $x \in \mathcal{X}$ with $\mathbb{E}_F[X] = 0$ and $y = 2$. The black dashed vertical line indicates the location of the contaminant y. 135

6.1 Fix $\rho = e^{-11}$. Panels [A] and [B] show the penalized SM log-density estimates with and without the additional observation $y = 120$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 120$. Panel [F] shows the sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 120$. 156

6.2 Fix $\rho = e^{-11}$. Panels [A] and [B] show the penalized SM log-density estimates with and without the additional observation $y = 180$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 180$. Panel [F] shows the resulting sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 180$. 157

6.3 Fix $\lambda = e^{-15}$. Panels [A] and [B] show the penalized ML log-density estimates with and without the additional observation $y = 120$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 120$. Panel [F] shows the resulting sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 120$. 158

6.4 Fix $\lambda = e^{-15}$. Panels [A] and [B] show the penalized ML log-density estimates with and without the additional observation $y = 180$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 180$. Panel [F] shows the resulting sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 180$. 159

6.5 Left panel shows the overall influence versus different choices of y, where we fix $\rho = e^{-11}$ and the rugs indicate the location of the `waiting` data. Right panel shows the overall influence against different choices of ρ (shown in log scale), where we fix $y = 120$. 161

6.6 Heat map of the overall influence on the penalized SM log-density estimates against y and ρ (shown in log scale). White rugs indicate locations of the `waiting` data. 162

6.7 Heat maps of the overall influence on penalized ML (left) and SM (right) log-density estimates against y and the RKHS norm of the natural parameter under F_n (shown in log scale). Red vertical line in left panel indicates the case $\lambda = 0$. White rugs indicate locations of `waiting` data. 163

6.8 Overall influence of y on the K-fold cross-validated penalized SM density estimates against the values of y. We choose $K = 3$ (left panel), 5 (middle panel), and 10 (right panel). 166

7.1 ML log-concave density estimate with 100 random samples from the standard normal distribution, where the density estimate is computed using the `R` package `logcondens` (Dümbgen and Rufibach, 2010). Histogram with the bin width chosen by the Freedman-Diaconis rule is shown in green. 171

Chapter 1: Introduction

Density estimation is a classical and fundamental problem in statistics. With independent and identically distributed (i.i.d) data drawn from an unknown probability density function (pdf) p_0 over a domain $\mathcal{X} \subseteq \mathbb{R}^d$, the *density estimation problem* seeks to reconstruct p_0 using these data (Silverman, 1986).

Density estimation has found a wide range of statistical applications. On one hand, density estimates can be used in exploratory data analysis. More specifically, they can be utilized as an informal investigation of the properties of a given dataset and provide descriptive features such as multimodality, skewness, and tail behavior (Silverman, 1986; Izenman, 2009). On the other hand, density estimates can be used as an intermediate step to perform further statistical analysis, such as classification (Ripley, 1996) and clustering (Fukunaga and Hostetler, 1975). Due to the wide applications, there have been abundant works contributing to this problem.

We will provide a review of different approaches to the density estimation problem and discuss their advantages and disadvantages in Section 1.1. We then turn to the focus of this dissertation, the density estimation problem in an exponential family induced by a reproducing kernel Hilbert space (RKHS), and introduce the problems we focus on in this dissertation in Section 1.2. The organization of the rest of this dissertation will be given in Section 1.3.

1.1 A Review of Density Estimation Methods

Existing density estimation methods can be broadly categorized into the parametric approach (Subsection 1.1.1) and the nonparametric approach (Subsection 1.1.2). The parametric approach imposes a strong assumption that p_0 belongs to a parametric family known up to a few parameters, whereas the nonparametric approach abandons such a restrictive constraint and makes much fewer assumptions about p_0, which allows data to speak for themselves and offers more flexibility.

In the remaining part of this section, we provide details of these two different approaches and discuss their advantages and disadvantages.

1.1.1 Parametric Approach

For the parametric approach, we assume p_0 belongs to a parametric family

$$\mathcal{Q}_{\text{para}} := \Big\{ q_\theta : \mathcal{X} \to [0, \infty) \;\Big|\; \theta \in \Theta \Big\},$$

where $\theta := (\theta_1, \cdots, \theta_m)^\top$ is an m-dimensional parameter and $\Theta \subseteq \mathbb{R}^m$ is the parameter space; in other words, we assume there exists $\theta_0 \in \Theta$ such that $p_0 = q_{\theta_0}$. The density estimation problem then reduces to a parameter estimation problem. Once we obtain an estimator of θ_0, say $\hat{\theta}$, the resulting density estimator is $q_{\hat{\theta}}$.

Many methods of estimating θ_0 are available, for example, the method of moments and the method of maximum likelihood. The latter method, first proposed by Fisher (1922), considers to maximize the likelihood function

$$\prod_{i=1}^{n} q_\theta(X_i), \qquad \text{subject to } \theta \in \Theta, \tag{1.1}$$

or, equivalently, maximize the (averaged) log-likelihood function

$$\frac{1}{n} \sum_{i=1}^{n} \log q_\theta(X_i), \qquad \text{subject to } \theta \in \Theta. \tag{1.2}$$

The maximum likelihood estimator of θ_0, denoted by $\hat{\theta}_{\text{ML}}$, is any point in Θ maximizing (1.1) or (1.2).

Since the functional form of pdfs in $\mathcal{Q}_{\text{para}}$ is known, $\hat{\theta}_{\text{ML}}$ is generally very easy to compute, where $\hat{\theta}_{\text{ML}}$ either has an analytic form (e.g., when $\mathcal{Q}_{\text{para}}$ is the family of Gaussian pdfs with unknown mean and covariance matrix) or can be obtained by solving an optimization problem of dimensionality at most m.

Statistical properties of these estimators have been well understood. For example, under certain regularity conditions, $\hat{\theta}_{\text{ML}}$ can be shown to be consistent, asymptotically efficient, and asymptotically normally distributed (Casella and Berger, 2002). Some of these favorable statistical properties can be extended to density estimators under additional assumptions.

Although density estimators from this parametric approach have computational advantages and possess many nice statistical properties, this approach relies on the rigid assumption $p_0 \in \mathcal{Q}_{\text{para}}$, which is hard or even impossible to verify in practice and is "entirely a matter for the practical statistician" (Fisher, 1922). If $p_0 \notin \mathcal{Q}_{\text{para}}$, the resulting density estimator from $\mathcal{Q}_{\text{para}}$ can be misleading and can lead to serious misspecification issues, which has been discussed by Huber (1967) and White (1982). Therefore, Fisher (1922) suggested to perform *a posteriori* test to examine the adequacy of the parametric assumption and the potential existence of misspecification.

1.1.2 Nonparametric Approach

If a parametric model is not postulated, we pursue the nonparametric approach and make as few assumptions as possible about p_0. This is the focus of this dissertation. In particular, we focus on nonparametric methods via minimizing a loss

functional plus a penalty functional over a class of pdfs,

$$\underset{q\in\mathcal{Q}}{\text{minimize}}\left\{\widehat{L}(q)+\lambda P(q)\right\}, \tag{1.3}$$

where $\mathcal{Q}$ is a pre-specified class of pdfs over $\mathcal{X}$, $\widehat{L}:\mathcal{Q}\to\mathbb{R}$ is a loss functional, $P:\mathcal{Q}\to[0,\infty)$ is a penalty functional, and $\lambda>0$ is the penalty parameter. In (1.3), $\widehat{L}$ depends on data and measures the goodness-of-fit of q to data, and a smaller value of $\widehat{L}(q)$ means the better the fit of q to data; and P is typically independent of data and measures the smoothness or size of q, and a larger value of $P(q)$ implies q is less smooth or more complex. Hence, the objective functional in (1.3) represents two conflicting goals: we demand q to have a good fit to data, but we also require it to contain less variations and be not too complex. The penalty parameter λ controls the tradeoff between these two conflicting goals.

We will focus on two different choices of $\widehat{L}$ in this dissertation, the negative log-likelihood loss functional (Subsection 1.1.2.1) and the score matching loss functional (Subsection 1.1.2.2).

1.1.2.1 Nonparametric Maximum Likelihood Density Estimation

Throughout this subsection, we let $\widehat{L}$ in (1.3) be the (averaged) negative log-likelihood (NLL) loss functional

$$\widehat{L}_{\text{NLL}}(q):=-\frac{1}{n}\sum_{i=1}^{n}\log q(X_i). \tag{1.4}$$

We call the density estimator via minimizing $\widehat{L}_{\text{NLL}}$ the maximum likelihood (ML) density estimator, as minimizing the NLL loss functional is equivalent to maximizing the log-likelihood functional.

Minimizing $\widehat{L}_{\mathrm{NLL}}$ can be viewed as minimizing a sample version of the *Kullback-Leibler divergence* (KL-divergence)

$$\mathrm{KL}(p\|q) := \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x, \tag{1.5}$$

where p and q are two pdfs over $\mathcal{X}$ and satisfy $\mathrm{KL}(p\|q) < \infty$. To see this, assuming $\mathrm{KL}(p_0\|q) < \infty$ for all $q \in \mathcal{Q}$, with $X_1, \cdots, X_n \overset{\text{i.i.d}}{\sim} p_0$, we can approximate $\mathrm{KL}(p_0\|q)$ by

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\big(\log p_0(X_i) - \log q(X_i)\big) &= \frac{1}{n}\sum_{i=1}^{n}\log p_0(X_i) - \frac{1}{n}\sum_{i=1}^{n}\log q(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log p_0(X_i) + \widehat{L}_{\mathrm{NLL}}(q).\end{aligned}$$

Notice that $\frac{1}{n}\sum_{i=1}^{n}\log p_0(X_i)$ is independent of q. The desired conclusion follows.

If we let $\mathcal{Q}$ be the class of all pdfs over $\mathcal{X}$ and $P(q) = 0$ for all $q \in \mathcal{Q}$, $\widehat{L}_{\mathrm{NLL}}$ is unbounded from below. To see this, suppose $\mathcal{X} = \mathbb{R}$ and let

$$q_{\sigma^2}(x) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x - X_j)^2\right), \qquad \text{for all } x \in \mathbb{R},$$

where $\sigma^2 > 0$. It is easy to verify that q_{σ^2} is a valid pdf over $\mathbb{R}$. Then, by shrinking $\sigma^2 \to 0$, we have $\widehat{L}_{\mathrm{NLL}}(q_{\sigma^2}) \to -\infty$.

Therefore, in order to obtain a sensible density estimator using $\widehat{L}_{\mathrm{NLL}}$, we need to impose certain constraints on $\mathcal{Q}$ and/or choose a nonzero P. Several proposals have been made in the literature.

1.1.2.1.1 Penalized Maximum Likelihood Density Estimation

In their seminal paper, Good and Gaskins (1971) proposed to use a nonzero P and minimize

$$\begin{aligned}-\frac{1}{n}\sum_{i=1}^{n}\log\gamma^2(X_i) + \left(\lambda_1\int_{\mathcal{X}}(\gamma'(x))^2\mathrm{d}x + \lambda_2\int_{\mathcal{X}}(\gamma''(x))^2\mathrm{d}x\right), \\ \text{subject to } \int_{\mathcal{X}}\gamma^2(x)\mathrm{d}x = 1,\end{aligned} \tag{1.6}$$

where $\gamma^2 = q$, $\lambda_1 > 0$, and $\lambda_2 > 0$. The specific form of the penalty functional in (1.6) is motivated from penalizing both the slope and the curvature to ensure the resulting density estimates are very smooth. If $\hat{\gamma}$ is a solution to (1.6), the resulting density estimator is $\hat{q} = \hat{\gamma}^2$.

There are at least two advantages of working with γ rather than the pdf q in (1.6). First, the density estimator by this approach is automatically nonnegative over $\mathcal{X}$, since the density estimator is $\hat{q} = \hat{\gamma}^2$.

Furthermore, due to the constraint $\int_{\mathcal{X}} \gamma^2(x)\mathrm{d}x = 1$, γ must belong to $L^2(\mathcal{X})$, the class of square-integrable functions over $\mathcal{X}$, which is a Hilbert space. This provides computational convenience that one can exploit to compute the minimizer of (1.6). Supposing $\{\varphi_j\}_{j=1}^{\infty}$ is an orthonormal basis of $L^2(\mathcal{X})$, we can then approximate any $\gamma \in L^2(\mathcal{X})$ by $\gamma \approx \sum_{j=1}^{m} c_j \varphi_j$, for a large m. As a result, the minimization problem (1.6) over $L^2(\mathcal{X})$, an infinite-dimensional class of functions, reduces to a computationally tractable problem over $\mathbb{R}^m$, as one only needs to determine the values of $c_1, \cdots, c_m$. The value of m can be determined via cross-validation or through an iterative fashion as Good and Gaskins (1971) did.

When solving (1.6), one has to deal with the constraint $\int_{\mathcal{X}} \gamma^2(x)\mathrm{d}x = 1$, which can be difficult in practice. In order to remedy this, Leonard (1978) introduced the logistic transformation of the density function, proposed to parametrize q as $q(x) = \exp(f(x))/\int_{\mathcal{X}} \exp(f(t))\mathrm{d}t$ for all $x \in \mathcal{X}$, and considered to minimize

$$-\frac{1}{n}\sum_{i=1}^{n} f(X_i) + \log\left(\int_{\mathcal{X}} \exp(f(x))\mathrm{d}x\right) + \frac{\lambda}{2}\widetilde{P}(f), \qquad \text{subject to } f \in \mathcal{H}, \tag{1.7}$$

where $\mathcal{H}$ is a pre-specified class of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ and $\widetilde{P}(f) := P(q)$. If the solution to (1.7) is $\hat{f}$, the resulting density estimator is $\exp(\hat{f}(x))/\int_{\mathcal{X}} \exp(\hat{f}(t))\mathrm{d}t$, for all $x \in \mathcal{X}$, which is nonnegative over $\mathcal{X}$ and integrates to 1.

The main disadvantage of the formulation (1.7) is that it may *not* have a unique solution, since, if $\hat{f}$ is a solution, then $\hat{f} + c$ is a solution as well, for any $c \in \mathbb{R}$. To remedy this, Silverman (1982) proposed to minimize

$$-\frac{1}{n}\sum_{i=1}^{n} f(X_i) + \int_{\mathcal{X}} \exp(f(x))\mathrm{d}x + \frac{\lambda}{2}\widetilde{P}(f), \qquad \text{subject to } f \in \mathcal{H}. \tag{1.8}$$

Assuming $\widetilde{P}$ depends only on the square of derivatives of f, Silverman (1982) proved the minimizer of (1.8) exists and is unique under very mild conditions, and showed, if $\hat{f}$ is the solution to (1.8), $\exp(\hat{f})$ is automatically the density estimator. Moreover, Silverman (1982) established the consistency and the asymptotic convergence rate of his estimator under various function norms. However, no practical algorithms were proposed to compute his density estimator.

Penalty functionals we have reviewed so far all depend on the squared L^2 norm of the derivatives of the root-density or a transformation of the log-density. This is computationally convenient as the corresponding objective functional is differentiable so that first- and second-order iterative optimization algorithms can be applied to compute the respective minimizers and the resulting density estimators. Such penalty functionals, nonetheless, allow no jumps or piecewise linear bends in density estimates and often lead to over-smoothed density estimates (Sardy and Tseng, 2010). Motivated by these observations, Koenker and Mizera (2007) considered to minimize (1.8) and chose $\widetilde{P}$ to be the total variation of f',

$$\widetilde{P}(f) = \sup \sum_{i=1}^{m} |f'(u_i) - f'(u_{i-1})|, \tag{1.9}$$

where the supremum is taken over all partitions of $u_1 < u_2 < \cdots < u_m$ in $\mathcal{X} \subset \mathbb{R}$. To facilitate the computation, Koenker and Mizera (2007) assumed f is a piecewise linear function supported on $[X_{(1)}, X_{(n)}]$ with knots at the order statistics $X_{(1)}, \cdots, X_{(n)}$ and proposed to use the interior point method to compute the minimizer. From their

simulation studies, Koenker and Mizera (2007) found that the penalty functional (1.9) works particularly well when p_0 is not smooth or contains sharp peaks. However, theoretical properties of their density estimator, such as consistency and the rate of convergence, remain unexplored.

1.1.2.1.2 Shape-constrained Maximum Likelihood Density Estimation

Even though density estimators by minimizing $\widehat{L}_{\mathrm{NLL}}$ plus a penalty functional have very nice statistical properties, the quality of density estimates depends heavily on the choice of the penalty parameter $\lambda > 0$, which is a nontrivial task in general.

A different direction of estimating p_0 via minimizing $\widehat{L}_{\mathrm{NLL}}$ is to impose certain qualitative properties on p_0, such as monotonicity, unimodality, or log-concavity. This shape-constrained approach is attractive as it requires no choice of the penalty parameter and is fully automatic (Cule, Samworth, and Stewart, 2010).

Shape-constrained density estimation originated from Grenander (1956), who studied the problem of estimating a non-increasing pdf over $[0, \infty)$ via minimizing L_{NLL}. It turns out that the solution, called the *Grenander estimator*, exists and is the left derivative of the *least concave majorant* of the empirical distribution function $\widehat{F}_n$, where the *least concave majorant* of a function F on $[0, +\infty)$ is

$$\inf\left\{ G \;\middle|\; G \text{ is concave over } [0,\infty), \text{ and } G(x) \geq F(x) \text{ for all } x \geq 0 \right\}.$$

Various statistical properties of the Grenander estimator, such as pointwise consistency and pointwise asymptotic distribution, have been investigated by Rao (1969), Groeneboom (1984), and Birge (1989).

The Grenander estimator can be extended to the case where p_0 is unimodal with the known mode. Suppose p_0 is unimodal with the known mode m_0, and is non-decreasing on $(-\infty, m_0]$ and is non-increasing on $[m_0, +\infty)$. With the aid of the

Grenander estimator, a natural estimator of p_0 is the derivative of the empirical distribution function obtained by the union of the greatest convex minorant of $\widehat{F}_n$ over $(-\infty, m_0]$ and the least concave majorant over $[m_0, +\infty)$ (Birgé, 1997). Theoretical properties of the Grenander estimator can be carried over to the unimodal density estimator.

When the mode m_0 is unknown, which is typically the case, however, the minimizer of $\widehat{L}_{\mathrm{NLL}}$ over the class of all unimodal pdfs over $\mathbb{R}$ does *not* exist, since one can put the infinite density value at one of the observations (Birgé, 1997). Wegman (1970a), Wegman (1970b) and Birgé (1997) have proposed different unimodal density estimators when the mode is unknown and studied statistical properties of their respective estimators.

Despite their easiness of implementation and nice statistical properties, the monotone and unimodal density estimators discussed so far are hard to be generalized to the multivariate setting. A different shape constraint that has drawn a lot of attention in the past two decades and is easy to be generalized to the multivariate setting is the log-concavity. A function $p : \mathcal{X} \to [0, +\infty)$ is said to be *log-concave* if $\log p$ is a concave function on $\mathcal{X}$ with $\log 0 = -\infty$.

Indeed, many standard families of parametric density functions are log-concave, including all Gaussian pdfs with positive definite covariance matrix, all gamma pdfs $\Gamma(\alpha, \beta)$ with shape parameter $\alpha \geq 1$, all beta pdfs $\mathrm{Beta}(\alpha, \beta)$ with $\alpha \geq 1$ and $\beta \geq 1$, and so on. A comprehensive list of log-concave parametric families together with their applications in economics can be found in Bagnoli and Bergstrom (2005).

Log-concave density functions have many useful properties. The log-concavity implies that the density is automatically unimodal and has convex level sets. The convolution of a log-concave density function with any unimodal density function is

again unimodal (Ibragimov, 1956). The convolution of two log-concave pdfs is again log-concave, implying that if random variables X and Y have log-concave densities and are independent, then their sum $X + Y$ also has a log-concave density. Furthermore, random vectors with a log-concave density function have moment generating functions that are finite in a neighborhood of the origin and, thus, have moments of all orders (Samworth, 2018). In addition, if $X = (X_1^\top, X_2^\top)^\top \in \mathbb{R}^d$ has a log-concave density, the marginal densities of X_1 and X_2 are log-concave and the conditional density of X_1 given $X_2 = x_2$ is also log-concave for each x_2. Last but not the least, if X is a d-dimensional random vector with a log-concave pdf and A is a fixed $m \times d$ matrix of rank m, then the random vector AX has a log-concave density on $\mathbb{R}^m$. From these properties, the class of log-concave density functions share many similarities with the class of Gaussian density functions, and can be viewed as an infinite-dimensional generalization of the latter (Cule, Samworth, and Stewart, 2010; Samworth, 2018).

Log-concave density estimation over $\mathbb{R}^d$ via minimizing $\widehat{L}_{\mathrm{NLL}}$ amounts to choosing $\mathcal{Q}$ to be the class of all log-concave pdfs over $\mathbb{R}^d$, which is equivalent to minimizing

$$-\frac{1}{n}\sum_{i=1}^{n} f(X_i) + \int_{\mathcal{X}} \exp(f(x))\mathrm{d}x, \qquad \text{subject to } f : \mathcal{X} \to [-\infty, \infty) \text{ is concave.} \tag{1.10}$$

It turns out that the solution to (1.10) exists and is unique, and admits a finite-dimensional representation. If $d = 1$, the solution is a piecewise linear function with knots at the order statistics $X_{(1)}, \cdots, X_{(n)}$, and is $-\infty$ over $\mathbb{R}\backslash[X_{(1)}, X_{(n)}]$ (Walther, 2002; Pal, Woodroofe, and Meyer, 2007; Dümbgen and Rufibach, 2009); if $d > 1$, the solution is characterized as a "tent function" supported on the convex hull of the data (Cule, Samworth, and Stewart, 2010). Here, for a fixed $y = (y_1, \cdots, y_n) \in \mathbb{R}^n$, a *tent function* is a function $\bar{h}_y : \mathbb{R}^d \to \mathbb{R}$ with the property that $\bar{h}_y$ is the pointwise least

concave function satisfying $\bar{h}_y(X_i) \geq y_i$ for all $i = 1, \cdots, n$. If $\hat{f}_{\text{lc}}$ is the minimizer of (1.10), the ML log-concave density estimator is $\hat{q}_{\text{lc}}(x) := \exp(\hat{f}_{\text{lc}}(x))$ for all $x \in \mathcal{X}$.

Various algorithms have been proposed to compute $\hat{f}_{\text{lc}}$. In the case of $d = 1$, Walther (2002), Pal, Woodroofe, and Meyer (2007), and Rufibach (2007) proposed to use the iterative convex minorant algorithm or a variant of it, and Dümbgen, Huesler, and Rufibach (2007) proposed an active set algorithm that turns out to be very efficient and was implemented in the `R` package `logcondens` (Dümbgen and Rufibach, 2010). For the cases of $d > 1$, two main difficulties exist. With the solution characterized as the "tent function", the corresponding objective function is non-differentiable. Cule, Samworth, and Stewart (2010) adopted Shor's r-algorithm, a subgradient method for minimizing non-differentiable convex functions over the Euclidean spaces, to handle the non-differentiability issue. The other difficulty is the computation of the integral appearing in (1.10) and its subgradient. Cule, Samworth, and Stewart (2010) proposed to triangulate the convex hull of data and compute the integral over each simplex in the triangulation. The time of computing $\hat{f}_{\text{lc}}$ unfortunately increases quickly with the sample size n and the dimensionality d and can be intolerably long for large n and d. As is reported in Table 1 by Cule, Samworth, and Stewart (2010), it takes about 224 minutes to compute a 4-dimensional log-concave density estimate using a sample of size 2000. In recent years, more efficient algorithms to compute $\hat{f}_{\text{lc}}$ for $d > 1$ cases have been proposed by Axelrod et al. (2019), Rathke and Schnörr (2019), and Chen, Mazumder, and Samworth (2021).

Theoretical properties of log-concave density estimators have also been investigated recently. When $d = 1$, under the assumption that p_0 is log-concave, Pal, Woodroofe, and Meyer (2007) proved $\hat{q}_{\text{lc}}$ is consistent for p_0 under the Hellinger distance, and Doss and Wellner (2016) established the corresponding convergence rate.

Dümbgen and Rufibach (2009) established the uniform consistency of $\hat{q}_{\mathrm{lc}}$ and the corresponding convergence rate. Furthermore, Balabdaoui, Rufibach, and Wellner (2009) derived the pointwise limiting distribution of $\hat{q}_{\mathrm{lc}}$. For the multivariate case, if p_0 is log-concave over $\mathbb{R}^d$, then $\hat{q}_{\mathrm{lc}}$ has been shown to be consistent under the exponentially weighted total variation distance and the Hellinger distance (Cule, Samworth, and Stewart, 2010; Dümbgen, Samworth, and Schuhmacher, 2011; Kim and Samworth, 2016). Additional details about log-concave density estimation via minimizing $\widehat{L}_{\mathrm{NLL}}$ can be found in Samworth (2018), a comprehensive survey of the recent progress in this topic.

Based on our discussion so far, the main advantage of the log-concave density estimation via minimizing $\widehat{L}_{\mathrm{NLL}}$ over the penalized approach is that one does *not* need to select the penalty parameter. It does suffer from some disadvantages such as the heavy computational burden discussed earlier. In addition, density estimates have some unsatisfactory qualitative features: they typically contain kinks and are only supported over the convex hull of data, and all boundary points of the convex hull of data are discontinuous points of $\hat{q}_{\mathrm{lc}}$. These unsatisfactory features may lead to serious problems in statistical applications such as classification. Smooth log-concave density estimator has been proposed by Chen and Samworth (2013).

1.1.2.2 Nonparametric Score Matching Density Estimation

Hyvärinen (2005) proposed a different loss functional, called the *score matching (SM) loss functional*, that can be used in the density estimation problem and is given by

$$\widehat{L}_{\mathrm{SM}}(q) := \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\left(\frac{1}{2}\big(\partial_u \log q(X_i)\big)^2 + \partial_u^2 \log q(X_i)\right), \tag{1.11}$$

where $q : \mathcal{X} \to (0, \infty)$ is assumed to be a twice continuously differentiable pdf,

$$\partial_u \log q(x) := \frac{\partial}{\partial w_u} \log q(w)\Big|_{w=x}, \qquad \text{and} \qquad \partial_u^2 \log q(x) := \frac{\partial^2}{\partial w_u^2} \log q(w)\Big|_{w=x},$$

for all $u = 1, \cdots, d$, where $w := (w_1, \cdots, w_d)^\top \in \mathcal{X}$.

The SM loss functional originates from the Hyvärinen divergence (H-diveregnce) (Hyvärinen, 2005)

$$\mathrm{H}(p_0 \| q) := \frac{1}{2} \int_{\mathcal{X}} p_0(x) \|\nabla \log p_0(x) - \nabla \log q(x)\|_2^2 \mathrm{d}x, \tag{1.12}$$

where we assume p_0 is continuously differentiable over $\mathcal{X}$ and satisfies

$$\int_{\mathcal{X}} p_0(x) \|\nabla \log p_0(x)\|_2^2 \mathrm{d}x < \infty, \qquad \text{and} \qquad \int_{\mathcal{X}} p_0(x) \|\nabla \log q(x)\|_2^2 \mathrm{d}x < \infty.$$

Using the integration by parts and assuming $p_0(x)\partial_u \log q(x) \to 0$ as x approaches to the boundary of $\mathcal{X}$ for all $u = 1, \cdots, d$, we have

$$\begin{aligned} \mathrm{H}(p_0 \| q) = & \int_{\mathcal{X}} p_0(x) \sum_{u=1}^{d} \left(\frac{1}{2} (\partial_u \log q(x))^2 + \partial_u^2 \log q(x) \right) \mathrm{d}x \\ & + \frac{1}{2} \int_{\mathcal{X}} p_0(x) \sum_{u=1}^{d} (\partial_u \log p_0(x))^2 \mathrm{d}x. \end{aligned} \tag{1.13}$$

Note that the last term depends on p_0 only and is independent of q. With $X_1, \cdots, X_n \overset{\text{i.i.d}}{\sim} p_0$, the SM loss functional (2.12) is simply the sample version of (1.13) with the last term omitted.

The main motivation of using $\widehat{L}_{\mathrm{SM}}$ is that one can avoid working with the normalizing constant. Let us return to the logistic transformation of the density function discussed earlier and let $q(x) = \exp(f(x))/\int_{\mathcal{X}} \exp(f(t))\mathrm{d}t$ for all $x \in \mathcal{X}$, and suppose we know the functional form of $f : \mathcal{X} \to \mathbb{R}$. It is typical that the normalizing constant $\int_{\mathcal{X}} \exp(f(t))\mathrm{d}t$ is unknown and is analytically and computationally intractable. Then, since $\partial_u \log q(x) = \partial_u f(x)$ and $\partial_u^2 \log q(x) = \partial_u^2 f(x)$ do *not* depend on the normalizing constant $\int_{\mathcal{X}} \exp(f(t))\mathrm{d}t$, neither does $\widehat{L}_{\mathrm{SM}}$.

Minimizing $\widehat{L}_{\mathrm{SM}}$ over the class of all pdfs over $\mathcal{X}$ is unbounded below. To see this, assume $\mathcal{X} = \mathbb{R}$ and again consider

$$q_{\sigma^2}(x) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - X_j)^2\right), \qquad \text{for all } x \in \mathbb{R},$$

which is a valid pdf over $\mathbb{R}$. In addition, note that q_{σ^2} is twice continuously differentiable over $\mathbb{R}$. With some algebra, it can be shown that $\widehat{L}_{\mathrm{SM}}(q_{\sigma^2}) \to -\infty$ as $\sigma^2 \to 0^+$. This suggests that, in order to obtain a sensible density function, one has to put certain constraints on $\mathcal{Q}$ or use a nonzero P.

1.2 Nonparametric Density Estimation in Kernel Exponential Families

So far, we have discussed various approaches to density estimation. In the rest of this dissertation, we will focus on the nonparametric approach via minimizing $\widehat{L}_{\mathrm{NLL}}$ and $\widehat{L}_{\mathrm{SM}}$, and restrict $\mathcal{Q}$ to be an exponential family induced by a RKHS, which we call the *kernel exponential family* and will introduce formally in Chapter 2, and discuss various density estimators in it.

More specifically, for $\widehat{L}_{\mathrm{NLL}}$, we choose a nonzero penalty functional P and consider the penalized ML density estimator. For $\widehat{L}_{\mathrm{SM}}$, we consider two kinds of regularized density estimators: the penalized SM density estimator obtained by minimizing the SM loss functional plus a nonzero penalty functional P (Sriperumbudur et al., 2017), and the early stopping SM density estimator obtained by applying the gradient descent algorithm to minimizing the SM loss functional (with a zero P) and terminating the algorithm early to regularize (see Chapter 3).

Let us focus on the penalized SM density estimator for now. The first row in Figure 1.1 shows penalized SM density estimates of the `waiting` variable in the Old Faithful Geyser dataset (Azzalini and Bowman, 1990), with the corresponding penalty

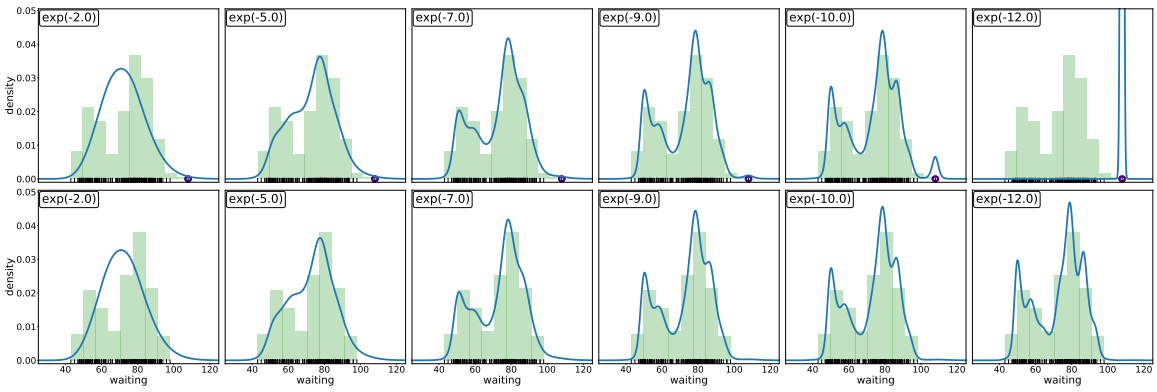

Figure 1.1: Penalized SM density estimates of the `waiting` variable with (first row) and without (second row) the isolated observation 108 (indicated by the purple circle). Histogram of the `waiting` variable with the bin width selected by the Freedman-Diaconis rule (Freedman and Diaconis, 1981) is shown in green.

parameter values listed in the upper left corner. The `waiting` variable records 299 time intervals (measured in minutes) between the starts of successive eruptions of the Old Faithful Geyser in Yellowstone National Park collected continuously from August 1st to August 15th, 1985. Notice, as the value of the penalty parameter becomes smaller, the penalized SM density estimates contain a bump or become a spike at the isolated observation 108. If we remove this isolated observation, as is shown in the second row of Figure 1.1, the resulting density estimates do *not* contain a bump or a spike when the value of the penalty parameter is small. As we will see in Chapter 3, the early stopping SM density estimator is qualitatively very similar to the penalized SM density estimator. When we compare these two kinds of regularized SM density estimators with the penalized ML density estimator in Chapter 4, however, the latter does not contain a bump or a spike, even when there is no penalty.

The observations above motivate us to study the sensitivity of these different density estimators to the presence of an isolated observation. The tool we choose is the influence function (Hampel, 1968), a classic concept from the robust statistics.

Traditionally, the influence function was defined for real- and vector-valued statistical functionals and was used to study the robustness properties of various real- and vector-valued estimators. But the object of primary interest in the density estimation problem is a function. The classic notion of the influence function is not directly applicable. We extend its definition to allow function-valued statistical functionals and to facilitate our understanding of the sensitivity of various density estimators.

1.3 Organization of the Remaining Dissertation

The rest of this dissertation is organized as follows.

In Chapter 2, we will formally introduce the kernel exponential family, discuss its properties, show its connection with the classic finite-dimensional exponential family, and discuss the density estimation problem in it using $\widehat{L}_{\mathrm{NLL}}$ and $\widehat{L}_{\mathrm{SM}}$ found in the literature.

In Chapter 3, we will focus on the early stopping SM density estimator and discuss its theoretical properties. We also compare it with the penalized SM density estimator and address their similarities and differences.

In Chapter 4, we will numerically compare two kinds of regularized SM density estimators with the penalized ML density estimator. We will demonstrate that the representer theorem, a classic theorem that characterizes the minimizer of a penalized convex loss functional over a possibly infinite-dimensional RKHS, cannot be used to characterize the minimizer of the penalized NLL loss functional. Instead, we discuss how to find a finite-dimensional subspace in $\mathcal{H}$ to approximate the minimizer of the penalize NLL loss functional, and propose an algorithm to compute the minimizer in such a subspace. In order to ensure the comparability, we also minimize the (penalized) SM loss functional in this finite-dimensional subspace and discuss how to

achieve this. Furthermore, we will explain why the regularized SM density estimates contain a bump or a spike at the isolated observation when there is very small amount of regularization.

In Chapter 5, we will discuss our approach of extending the classic notion of the influence function to the studies of the sensitivity of density estimators. We will derive the influence functions of ML and SM (log-)density projections (to be defined) in both finite-dimensional and kernel exponential families. In Chapter 6, we compare the sensitivities of penalized ML and SM log-density estimators in the kernel exponential family through numerical examples and show that the penalized SM log-density estimator is more sensitive to the presence of an isolated observation than the penalized ML log-density estimator. Since we can use regularized SM or penalized ML density estimators, we will discuss which density estimator should be used and how it should be used in practice.

Finally, in Chapter 7, we will summarize this dissertation and discuss possible future directions based on the current work.

Chapter 2: Kernel Exponential Family and Density Estimation Problem in It

In this chapter, we formally introduce the kernel exponential family and discuss the density estimation problem in it.

2.1 Kernel Exponential Families

2.1.1 A Review of Finite-dimensional Exponential Family

In order to introduce the kernel exponential family, we first review the classic finite-dimensional exponential family.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the sample space and $\varphi : \mathcal{X} \to \mathbb{R}^m$ be a measurable vector-valued function such that $\varphi(x) = (\varphi_1(x), \cdots, \varphi_m(x))^\top \in \mathbb{R}^m$ for all $x \in \mathcal{X}$. An *m-dimensional exponential family* (Brown, 1986; Barndorff-Nielsen, 2014), denoted by $\mathcal{Q}_{\text{fin}}$, in its natural parametrization form, contains all pdfs of the form

$$\tilde{q}_\theta(x) := \mu(x) \exp\big(\langle \varphi(x), \theta \rangle - B(\theta)\big) \text{ for all } x \in \mathcal{X}, \qquad \theta \in \Theta, \tag{2.1}$$

where $\mu : \mathcal{X} \to [0, \infty)$ is a pdf referred to as the *base density*, $\theta \in \mathbb{R}^m$ is the *natural parameter*, φ is referred to as the *canonical statistic*, $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^m$,

$$B(\theta) := \log\left(\int_{\mathcal{X}} \mu(x) \exp\big(\langle \theta, \varphi(x) \rangle\big) \mathrm{d}x\right) \tag{2.2}$$

is the *log-partition function*, and $\Theta := \left\{\theta \in \mathbb{R}^m \;\middle|\; B(\theta) < \infty\right\}$ is the *natural parameter space*.

The finite-dimensional exponential family was first discovered by Darmois (1935), Koopman (1936), and Pitman (1936) in studying the family of distributions whose sufficient statistics have fixed dimensionality as the sample size increases. It can also be motivated via the principle of maximum entropy: given n i.i.d samples $X_1, \cdots, X_n \in \mathcal{X}$ and m measurable functions $\varphi_j : \mathcal{X} \to \mathbb{R}$, for all $j = 1, \cdots, m$, the unique pdf p over $\mathcal{X}$ that maximizes Shannon's entropy

$$-\int_{\mathcal{X}} p(x) \log p(x) \mathrm{d}x$$

subject to the linear constraints

$$\int_{\mathcal{X}} p(x)\varphi_j(x)\mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n} \varphi_j(X_i), \qquad \text{for all } j = 1, \cdots, m,$$

takes on the form (2.1).

Exponential families are ubiquitous in statistics. Many parametric families are special cases of exponential families, such as the family of Gaussian pdfs and the family of probability mass functions (pmfs) of the binomial distribution with the known number of trials. In addition, they form the basis for the generalized linear models (McCullagh and Nelder, 1989). In recent years, they are also used extensively in the studies of graphical models (Wainwright and Jordan, 2008).

2.1.2 Kernel Exponential Family

Note that in (2.1), the function φ maps to an m-dimensional Euclidean space, which can be limited in some applications. In addition, $\tilde{q}_\theta$ depends on φ only through its inner product with $\theta \in \Theta$. Motivated by these observations, Canu and Smola (2006) proposed to replace the inner product in the Euclidean space in (2.1) by the

one in a RKHS, and introduced the *kernel exponential family*, denoted by $\mathcal{Q}_{\mathrm{ker}}$, that contains all pdfs over $\mathcal{X}$ of the form

$$q_f(x) := \mu(x) \exp\bigl(f(x) - A(f)\bigr) \text{ for all } x \in \mathcal{X}, \qquad f \in \mathcal{F}, \tag{2.3}$$

where

$$A(f) := \log\left(\int_{\mathcal{X}} \mu(x) \exp(f(x)) \mathrm{d}x\right) \tag{2.4}$$

is the *log-partition functional*, and $\mathcal{F} := \left\{ f \in \mathcal{H} \;\middle|\; A(f) < \infty \right\}$ is referred to as the *natural parameter space*.

Due to the reproducing property of k, we have $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the kernel function associated with the underlying RKHS $\mathcal{H}$, and $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ denotes the inner product in $\mathcal{H}$. Thus, comparing with (2.1), we see that $f \in \mathcal{H}$ plays the role of the natural parameter and $k(x, \cdot) \in \mathcal{H}$ plays the role of the canonical statistic.

The kernel exponential family $\mathcal{Q}_{\mathrm{ker}}$ has been used in various statistical applications, such as the anomaly detection (Canu and Smola, 2006), and the estimation of the conditional independence structure of graphical models (Sun, Kolar, and Xu, 2015).

2.1.3 Properties of $\mathcal{Q}_{\mathrm{ker}}$

The kernel exponential family $\mathcal{Q}_{\mathrm{ker}}$ has many nice properties, some of which are in common with those of $\mathcal{Q}_{\mathrm{fin}}$. In the following subsections, we discuss these properties.

2.1.3.1 Characterization of $\mathcal{F}$ for Bounded Kernels

The first property we discuss here is related to the characterization of $\mathcal{F}$ when a bounded kernel function k is used, where k is said to be *bounded* if $\kappa_1 := \sup_{x \in \mathcal{X}} \sqrt{k(x, x)} < \infty$.

Proposition 2.1. *If the kernel function k is bounded, then we have $\mathcal{F} = \mathcal{H}$.*

The proof of Proposition 2.1 can be found in Subsection 2.3.1.

As a consequence of Proposition 2.1, if $\mathcal{X} \subset \mathbb{R}^d$ is compact and k is continuous over $\mathcal{X}$, we must have $\kappa_1 < \infty$ and $\mathcal{F} = \mathcal{H}$. As another example, if $\mathcal{X} = \mathbb{R}^d$ or an unbounded proper subset of $\mathbb{R}^d$, and k is the Gaussian kernel function, $k(x, y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right)$ for all $x, y \in \mathcal{X}$, or the rational quadratic kernel function, $k(x, y) = \left(1 + \frac{\|x-y\|_2^2}{\sigma^2}\right)^{-1}$ for all $x, y \in \mathcal{X}$, where $\sigma > 0$ is the bandwidth parameter associated with each kernel function, we also have $\kappa_1 < \infty$ and $\mathcal{F} = \mathcal{H}$.

The boundedness of k is *not* a necessary condition for $\mathcal{F} = \mathcal{H}$, though. For instance, if we let $\mathcal{X} = \mathbb{R}$ and $k(x, y) = xy$ for all $x, y \in \mathbb{R}$, the corresponding $\mathcal{H}$ contains all functions f of the form

$$f(x) = \sum_{i=1}^{\infty} y_i x, \qquad \text{for all } x \in \mathcal{X},$$

that satisfies $\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} y_i y_j < \infty$. It is easy to observe $\sup_{x \in \mathcal{X}} \sqrt{k(x, x)} = \sup_{\mathcal{X}} |x| = \infty$. In addition, choose μ to be $\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ for all $x \in \mathcal{X}$. It follows that $A(f) < \infty$ for all $f \in \mathcal{H}$ and $\mathcal{F} = \mathcal{H}$.

2.1.3.2 Convexity of A

It is a classic result that the log-partition function B defined in (2.2) is a convex function and the natural parameter space Θ is a convex set (Theorem 7.1 in Barndorff-Nielsen, 2014). As the following proposition demonstrates, similar results hold for $\mathcal{Q}_{\text{ker}}$.

Proposition 2.2. *The log-partition functional A defined in (2.4) is convex over $\mathcal{F}$, and is strictly convex over $\mathcal{F}$ if $\mathcal{H}$ does* not *contain constant functions. In addition, $\mathcal{F}$ is convex.*

The proof of Proposition 2.2 can be found in Subsection 2.3.2.

As we will see in Subsection 2.2.1 of this chapter and Chapter 4, the convexity of A plays an important role in density estimation using the (penalized) NLL loss functional, which ensures the convexity of the (penalized) NLL loss functional and directly relates to the existence and the uniqueness of its minimizer.

2.1.3.3 Differential Properties of A

It is well-known that the log-partition function B in $\mathcal{Q}_{\text{fin}}$ is infinitely differentiable and has a close relationship with the cumulant and moment generating functions of the canonical statistic (Theorem 2.2 in Brown, 1986).

In this section, we study differential properties of the log-partition functional A in $\mathcal{Q}_{\text{ker}}$ defined in (2.4). We will show that A is also infinitely differentiable and link its derivatives (suitably defined) to the moments of $k(X, \cdot)$.

To start with, notice that the domain of A is a collection of functions over $\mathcal{X}$, or more precisely, a subset of a Hilbert space. We need a version of differentiability defined for functionals whose domain is a normed space or a subset of it. We choose the *Frechét differentiability*, whose definition, together with those of Fréchet derivative and gradient, is given in Section A.1 in Appendix A.

The following lemma illustrates these definitions and serves as a preparation for the derivation of the Fréchet derivative and gradient of A in Proposition 2.3.

Lemma 2.1. *Suppose* $\sup_{x\in\mathcal{X}} \sqrt{k(x,x)} < \infty$ *so that* $\mathcal{F} = \mathcal{H}$. *Let* $x \in \mathcal{X}$ *be fixed and* $J_x : \mathcal{H} \to \mathbb{R}$ *be the evaluation functional* $J_x(f) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}} = f(x)$ *for all* $f \in \mathcal{H}$. *Then,* J_x *is Frechét differentiable over* $\mathcal{H}$ *with the Frechét derivative at* $f \in \mathcal{H}$ *being*

$$\mathrm{D}J_x(f)(g) = \langle g, k(x, \cdot)\rangle_{\mathcal{H}}, \qquad \text{for all } g \in \mathcal{H}.$$

In addition, the Fréchet gradient is $\nabla J_x : \mathcal{H} \to \mathcal{H}$ *with* $\nabla J_x(f) = k(x, \cdot)$ *for all* $f \in \mathcal{H}$.

The proof of Lemma 2.1 can be found in Section 2.3.3.

With this lemma, we now establish the Fréchet differentiability of A over $\mathcal{H}$ and derive its Fréchet derivative and gradient at $f \in \mathcal{H}$.

Proposition 2.3 (Fréchet differentiability, derivative and gradient of A)**.** *Suppose* $\sup_{x \in \mathcal{X}} \sqrt{k(x,x)} < \infty$ *so that* $\mathcal{F} = \mathcal{H}$. *Then,* $A : \mathcal{H} \to \mathbb{R}$ *is Fréchet differentiable over* $\mathcal{H}$ *and its Fréchet derivative at* $f \in \mathcal{H}$ *is a bounded linear operator from* $\mathcal{H}$ *to* $\mathbb{R}$ *given by*

$$\mathrm{D}A(f)(g) = \left\langle g, \int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x \right\rangle_{\mathcal{H}} = \int_{\mathcal{X}} g(x) q_f(x) \mathrm{d}x, \tag{2.5}$$

for all $g \in \mathcal{H}$. *In addition, the Fréchet gradient of* A *is*

$$\nabla A : \mathcal{H} \to \mathcal{H}, \qquad f \longmapsto \int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x.$$

The proof of Proposition 2.3 can be found in Section 2.3.4.

Remark 2.1. Note that the integrand of $\int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x$ is an element in $\mathcal{H}$ and this integral is a Bochner integral (Diestel and Uhl, 1977; Denkowski, Migórski, and Papageorgiou, 2013); see Section A.2 in Appendix A for its definition and properties. In particular, the second equality in (2.5) follows from Proposition A.4(c) there.

In addition, we can view $\int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x$ as the output of the *kernel mean embedding* (Bertinet and Agnan, 2004; Smola et al., 2007), i.e., we map a distribution F, whose density function is q_f, into $\mathcal{H}$ via the map

$$F \longmapsto \int_{\mathcal{X}} k(x, \cdot) \mathrm{d}F(x) = \int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x.$$

Kernel mean embedding has drawn a lot of attention in recent years. Its statistical properties have been extensively studied by, to name a few, Le (2008), Sriperumbudur et al. (2010), Sriperumbudur, Fukumizu, and others (2011), and Muandet et al. (2016). It can be used in statistical hypothesis testing for independence (Gretton et al., 2005) and for the equality of two sets of random samples (Gretton et al., 2012), statistical clustering (Jegelka et al., 2009), the estimation of graphical models (Song, Gretton, and Guestrin, 2010; Song, Fukumizu, and Gretton, 2013; Song et al., 2014). More details about the kernel mean embedding can be found in the recent comprehensive survey by Muandet et al. (2017). ▶

Proposition 2.3 only considers the first-order Fréchet differentiability of A. By an inductive argument, we can show that A is r-times Fréchet differentiable for all $r \in \mathbb{N}$. In particular, the following proposition shows the result when $r = 2$.

Proposition 2.4 (Second-order Fréchet differentiability, derivative and gradient of A)**.** *Suppose* $\sup_{x \in \mathcal{X}} \sqrt{k(x,x)} < \infty$ *so that* $\mathcal{F} = \mathcal{H}$*. Then, the log-partition functional* A *is twice Fréchet differentiable over* $\mathcal{H}$ *and its second-order Frechét derivative at* $f \in \mathcal{H}$*, denoted by* $\mathrm{D}^2 A(f)$*, is a map from* $\mathcal{H}$ *to the collection of bounded linear operators from* $\mathcal{H}$ *to* $\mathbb{R}$ *given by*

$$\begin{aligned}\mathrm{D}^2 A(f)(g) = &\left[\left(\int_{\mathcal{X}} k(x,\cdot) \otimes k(x,\cdot) q_f(x)\mathrm{d}x\right)\right. \\ &\left. - \left(\int_{\mathcal{X}} k(x,\cdot) q_f(x)\mathrm{d}x\right) \otimes \left(\int_{\mathcal{X}} k(x,\cdot) q_f(x)\mathrm{d}x\right)\right] g \\ = &\left(\int_{\mathcal{X}} k(x,\cdot) g(x) q_f(x)\mathrm{d}x\right) - \left(\int_{\mathcal{X}} q_f(x) g(x)\mathrm{d}x\right)\left(\int_{\mathcal{X}} k(x,\cdot) q_f(x)\mathrm{d}x\right),\end{aligned}$$

for all $g \in \mathcal{H}$*.*

In addition, the second-order Fréchet gradient of A *at* $f \in \mathcal{F}$*, denoted by* $\nabla^2 A(f)$*, is a bounded linear operator from* $\mathcal{H}$ *to itself given by*

$$\nabla^2 A(f) = \int_{\mathcal{X}} k(x,\cdot) \otimes k(x,\cdot) q_f(x)\mathrm{d}x$$

$$- \left(\int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x \right) \otimes \left(\int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x \right)$$

The proof of Proposition 2.4 is similar to that of Proposition 2.3 and is omitted here.

The bounded linear operator $\nabla^2 A(f)$, or more generally, the operator

$$\Sigma_F := \left(\int_{\mathcal{X}} k(x, \cdot) \otimes k(x, \cdot) \mathrm{d}F(x) \right) - \left(\int_{\mathcal{X}} k(x, \cdot) \mathrm{d}F(x) \right) \otimes \left(\int_{\mathcal{X}} k(x, \cdot) \mathrm{d}F(x) \right)$$

where F is a distribution over $\mathcal{X}$, maps from $\mathcal{H}$ to itself, and is referred to as the *covariance operator* in the literature, since, for any $f, g \in \mathcal{H}$, we have

$$\langle f, \Sigma_F g \rangle_{\mathcal{H}} = \mathbb{E}_F[f(X)g(X)] - \mathbb{E}_F[f(X)]\,\mathbb{E}_F[g(X)],$$

which is the covariance between $f(X)$ and $g(X)$ with $X \sim F$. The operator Σ_F is known to be linear, bounded, self-adjoint, and of trace-class (Baker, 1973). The covariance operator has been used in such statistical and machine learning applications as dimensionality reduction (Fukumizu, Bach, and Jordan, 2004; Fukumizu, Bach, and Jordan, 2009), kernel principal component analysis (Schölkopf, Smola, and Müller, 1998), kernel canonical correlation analysis (Fukumizu, Bach, and Gretton, 2007), and independence and conditional independence measures (Gretton et al., 2005; Fukumizu et al., 2007). We will also see the covariance operator appears in the influence function of the ML log-density projection in $\mathcal{Q}_{\text{ker}}$ in Chapter 5. More details on the covariance operator and, more generally, the cross-covariance operator can be found in Section 3.2 and 4.3 in Muandet et al. (2017).

2.1.4 Connection to Finite-dimensional Exponential Families

In this section, we show the connection between $\mathcal{Q}_{\text{fin}}$ and $\mathcal{Q}_{\text{ker}}$. In particular, we show that density functions in $\mathcal{Q}_{\text{fin}}$ can be written in the form of those in $\mathcal{Q}_{\text{ker}}$ by

choosing $\mathcal{H}$ to be a finite-dimensional RKHS. Therefore, $\mathcal{Q}_{\text{fin}}$ is a special case of $\mathcal{Q}_{\text{ker}}$. If we choose $\mathcal{H}$ to be an infinite-dimensional RKHS, $\mathcal{Q}_{\text{ker}}$ is a strict generalization of $\mathcal{Q}_{\text{fin}}$.

We start with $\mathcal{Q}_{\text{fin}}$ that contains all pdfs of the form (2.1). Consider the following collection of functions

$$\mathcal{H}_0 := \left\{ \sum_{j=1}^{m} \alpha_j \varphi_j \;\middle|\; \alpha_1, \cdots, \alpha_m \in \mathbb{R} \right\},$$

which is a vector space with addition and scalar multiplication defined as

$$(f+g)(x) = f(x) + g(x), \qquad \text{and} \qquad (cf)(x) = cf(x), \qquad \text{for all } x \in \mathcal{X},$$

for all $f, g \in \mathcal{H}_0$ and all $c \in \mathbb{R}$.

Now, for any $f = \sum_{j=1}^{m} \alpha_j \varphi_j \in \mathcal{H}_0$ and $g = \sum_{j=1}^{m} \beta_j \varphi_j \in \mathcal{H}_0$, where $\alpha_j, \beta_j \in \mathbb{R}$ for all $j = 1, \cdots, m$, define the inner product between them to be

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{j=1}^{m} \alpha_j \beta_j. \tag{2.6}$$

Then, we have the following proposition.

Proposition 2.5. *The vector space $\mathcal{H}_0$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ defined in (2.6), denoted by $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$, is a RKHS with the kernel function*

$$k(x, y) = \sum_{j=1}^{m} \varphi_j(x) \varphi_j(y), \qquad \text{for all } x, y \in \mathcal{X}. \tag{2.7}$$

The proof of Proposition 2.5 can be found in Subsection 2.3.5.

With the construction above, we have

$$\begin{aligned} \tilde{q}_\theta(x) &= \mu(x) \exp\big(\langle \theta, \varphi(x) \rangle - B(\theta)\big) \\ &= \mu(x) \exp\left(\sum_{j=1}^{m} \theta_j \varphi_j(x) - B(\theta)\right) \\ &= \mu(x) \exp\big(\langle f, k(x, \cdot) \rangle_{\mathcal{H}_0} - A(f)\big), \end{aligned}$$

where $\theta := (\theta_1, \cdots, \theta_m)^\top \in \Theta \subseteq \mathbb{R}^m$, $f := \sum_{j=1}^m \theta_j \varphi_j \in \mathcal{H}_0$, $k(x, \cdot) := \sum_{j=1}^m \varphi_j(x)\varphi_j \in \mathcal{H}_0$, and

$$\begin{aligned} A(f) &= \log\left(\int_{\mathcal{X}} \mu(x) \exp\big(\langle f, k(x, \cdot)\rangle_{\mathcal{H}_0}\big) \mathrm{d}x\right) \\ &= \log\left(\int_{\mathcal{X}} \mu(x) \exp\left(\sum_{j=1}^m \theta_j \varphi_j(x)\right) \mathrm{d}x\right) = B(\theta). \end{aligned}$$

Thus, every $\tilde{q}_\theta \in \mathcal{Q}_{\text{fin}}$ can be written in the form of pdfs in $\mathcal{Q}_{\text{ker}}$, and, with the RKHS being $\mathcal{H}_0$, $\mathcal{Q}_{\text{fin}}$ is a special case of $\mathcal{Q}_{\text{ker}}$.

In the rest of this section, we provide several examples of frequently used finite-dimensional exponential families to illustrate Proposition 2.5. In these examples, we temporarily ignore the fact that μ is a pdf over $\mathcal{X}$, which is assumed in the definitions of $\mathcal{Q}_{\text{fin}}$ and $\mathcal{Q}_{\text{ker}}$.

Example 2.1 (Binomial distribution). Consider the family of pmfs of the binomial distribution with the known number of trials n and the success probability $\eta \in (0, 1)$. The general form of the pmf is

$$p_\eta(x) := \binom{n}{x} \eta^x (1-\eta)^{n-x}, \qquad \text{for all } x \in \mathcal{X} := \big\{0, 1, \cdots, n\big\}.$$

In the natural parametrization, we can rewrite p_η as

$$\tilde{q}_\theta(x) := \binom{n}{x} \exp\big(x\theta - n\log(1 + e^\theta)\big), \qquad \text{for all } x \in \mathcal{X},$$

where the natural parameter is $\theta := \log\big(\frac{\eta}{1-\eta}\big)$ and the natural parameter space is $\mathbb{R}$. We recognize

$$\mu(x) = \binom{n}{x} \quad \text{and} \quad \varphi(x) = x \text{ for all } x \in \mathcal{X}, \quad \text{and} \quad B(\theta) = n\log(1 + e^\theta).$$

Here, the canonical statistic is $\varphi = \mathrm{Id}$, the identity map, with $\varphi(x) = x$ for all $x \in \mathcal{X}$. Then, $\mathcal{H}_0$ in this case contains all functions of the form $f(x) = \theta x$ for all $x \in \mathcal{X}$, where $\theta \in \mathbb{R}$. With $\theta_1, \theta_2 \in \mathbb{R}$, the inner product between $f = \theta_1 \cdot \mathrm{Id} \in \mathcal{H}_0$ and $g = \theta_2 \cdot \mathrm{Id} \in \mathcal{H}_0$

is $\langle f, g\rangle_{\mathcal{H}_0} = \theta_1\theta_2$. The reproducing kernel is $k(x, y) = \langle x \cdot \mathrm{Id}, y \cdot \mathrm{Id}\rangle_{\mathcal{H}_0} = xy$ for all $x, y \in \mathcal{X}$.

Furthermore, with $f = \theta \cdot \mathrm{Id}$ for some $\theta \in \mathbb{R}$, we have

$$A(f) = \log\left(\sum_{x=0}^{n} \binom{n}{x} e^{\theta x}\right) = \log((1 + e^{\theta})^n) = n \log(1 + e^{\eta}) = B(\theta),$$

which is the desired result. Since $A(f) = A(\theta \cdot \mathrm{Id}) = n \log(1+e^{\theta}) < +\infty$ for all $\theta \in \mathbb{R}$, we have $\mathcal{F} = \big\{f \in \mathcal{H}_0 \mid f := \theta \cdot \mathrm{Id}, \theta \in \mathbb{R}\big\}$. ▶

Example 2.2 (Poisson distribution). Consider the family of the pmfs of the Poisson distribution with the mean parameter $\eta > 0$. The general form of the pmf is

$$p_\eta(x) = e^{-\eta}\frac{\eta^x}{x!}, \qquad \text{for all } x \in \mathcal{X} := \{0, 1, 2, \cdots\}.$$

We rewrite p_η in the natural parametrization form as

$$\tilde{q}_\theta(x) = \frac{1}{x!}\exp(x\theta - e^{\theta}), \qquad \text{for all } x \in \mathcal{X},$$

where the natural parameter is $\theta = \log \eta$ and the natural parameter space is $\mathbb{R}$. We recognize

$$\mu(x) = \frac{1}{x!} \quad \text{and} \quad \varphi(x) = x \text{ for all } x \in \mathcal{X}, \qquad \text{and} \qquad B(\theta) = e^{\theta}.$$

Here, similar to the binomial example we have considered earlier, the canonical statistic is $\varphi = \mathrm{Id}$, the identity map. Then, $\mathcal{H}_0$ contains all functions of the form $f(x) = \theta x$ for all $x \in \mathcal{X}$, where $\theta \in \mathbb{R}$. With $\theta_1, \theta_2 \in \mathbb{R}$, the inner product between $f = \theta_1 \cdot \mathrm{Id} \in \mathcal{H}_0$ and $g = \theta_2 \cdot \mathrm{Id} \in \mathcal{H}_0$ is $\langle f, g\rangle_{\mathcal{H}_0} = \theta_1\theta_2$. The reproducing kernel is $k(x, y) = \langle x \cdot \mathrm{Id}, y \cdot \mathrm{Id}\rangle_{\mathcal{H}_0} = xy$ for all $x, y \in \mathcal{X}$.

Furthermore, with $f = \theta \cdot \mathrm{Id}$ for some $\theta \in \mathbb{R}$, we have

$$A(f) = \log\left(\sum_{x=0}^{\infty} \frac{(e^{\theta})^x}{x!}\right) = \log(\exp(e^{\theta})) = e^{\theta} = B(\theta),$$

which is the desired result. Since $A(f) = e^{\theta} < +\infty$ if and only if $\theta \in \mathbb{R}$, the natural parameter space is $\mathcal{F} = \big\{f \in \mathcal{H}_0 \mid f = \theta \cdot \mathrm{Id}, \theta \in \mathbb{R}\big\}$. ▶

Example 2.3 (Exponential distribution). Consider the family of pdfs of the exponential distribution with the scale parameter $\eta > 0$. The general form of the pdf is

$$p_\eta(x) = \frac{1}{\eta}\exp\left(-\frac{x}{\eta}\right), \qquad \text{for all } x \in \mathcal{X} := \big\{x \mid x > 0\big\}.$$

We can rewrite it in the natural parametrization form as

$$\tilde{q}_\theta(x) = \exp\Big(x\theta + \log(-\theta)\Big), \qquad \text{for all } x \in \mathcal{X},$$

where the natural parameter is $\theta := -\frac{1}{\eta}$ and the natural parameter space is $\Theta := (-\infty, 0)$. We recognize that

$$\mu(x) = 1 \quad \text{and} \quad \varphi(x) = x \text{ for all } x \in \mathcal{X}, \qquad \text{and} \qquad B(\theta) = -\log(-\theta).$$

Here, the canonical statistic is $\varphi = \mathrm{Id}$. Then, $\mathcal{H}_0$ contains all functions of the form $f(x) = \theta x$ for all $x \in \mathcal{X}$, for some $\theta \in \Theta$. With $\theta_1, \theta_2 \in \Theta$, the inner product between $f = \theta_1 \cdot \mathrm{Id} \in \mathcal{H}$ and $g = \theta_2 \cdot \mathrm{Id} \in \mathcal{H}$ is $\langle f, g\rangle_{\mathcal{H}_0} = \theta_1\theta_2$. It follows that the reproducing kernel is $k(x, y) = \langle x \cdot \mathrm{Id}, y \cdot \mathrm{Id}\rangle_{\mathcal{H}_0} = xy$, for all $x, y \in \mathcal{X}$.

Finally, with $f = \theta \cdot \mathrm{Id}$ for some $\theta \in \Theta$, we have

$$A(f) = \log\left(\int_{\mathcal{X}} \mu(x)\exp(f(x))\mathrm{d}x\right) = \log\left(\int_0^{+\infty} \exp(\theta x)\mathrm{d}x\right) = \log\left(-\frac{1}{\theta}\right) = B(\theta),$$

which is the desired result. Since $A(f) = A(\theta \cdot \mathrm{Id}) = \log(-\frac{1}{\theta}) < +\infty$ if and only if $\theta < 0$, the natural parameter space is $\mathcal{F} = \big\{f \in \mathcal{H}_0 \mid f = \theta \cdot \mathrm{Id}, \theta \in (-\infty, 0)\big\}$. ▶

Example 2.4 (Univariate normal distribution). Consider the univariate normal distribution with the unknown mean $\eta \in \mathbb{R}$ and the unknown variance $\sigma^2 > 0$. The general form of the pdf is

$$p_{\eta,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\eta)^2\right), \qquad \text{for all } x \in \mathcal{X} := \mathbb{R}.$$

We can rewrite p_{η,σ^2} in the natural parametrization form as

$$\tilde{q}_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\theta_1 x^2 + \theta_2 x - \left(-\frac{\theta_2^2}{4\theta_1} - \frac{1}{2}\log(-2\theta_1)\right)\right),$$

where

$$\theta := (\theta_1, \theta_2)^\top = \left(-\frac{1}{2\sigma^2}, \frac{\eta}{\sigma^2}\right)^\top, \qquad \Theta := \left\{\theta \;\middle|\; \theta := (\theta_1, \theta_2)^\top, \theta_1 < 0, \theta_2 \in \mathbb{R}\right\}.$$

$$\varphi(x) := (\varphi_1(x), \varphi_2(x))^\top = (x^2, x)^\top \in \mathbb{R}^2, \quad \text{and} \quad \mu(x) = \frac{1}{\sqrt{2\pi}} \text{ for all } x \in \mathcal{X},$$

$$B(\theta) = -\frac{\theta_2^2}{4\theta_1} - \frac{1}{2}\log(-2\theta_1).$$

This is an example of a 2-dimensional exponential family.

In this case, $\mathcal{H}_0$ contains all functions f of the form

$$f(x) = \theta_1 x^2 + \theta_2 x, \qquad \text{for all } x \in \mathbb{R},$$

for some $(\theta_1, \theta_2)^\top \in \Theta$. For any $f, g \in \mathcal{H}_0$ with $f(x) = \theta_{1,1}x^2 + \theta_{1,2}x$ and $g(x) = \theta_{2,1}x^2 + \theta_{2,2}x$ for all $x \in \mathcal{X}$, where $(\theta_{1,1}, \theta_{1,2})^\top \in \Theta$ and $(\theta_{2,1}, \theta_{2,2})^\top \in \Theta$, define the inner product between f and g to be

$$\langle f, g \rangle_{\mathcal{H}_0} = \theta_{1,1}\theta_{2,1} + \theta_{1,2}\theta_{2,2}.$$

Then, $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ forms a RKHS with the reproducing kernel

$$k(x, y) = x^2 y^2 + xy, \qquad \text{for all } x, y \in \mathbb{R}.$$

We finally verify $A(f) = B(\theta)$ with $f(x) = \theta_1 x^2 + \theta_2 x$ for all $x \in \mathcal{X}$, where $\theta := (\theta_1, \theta_2)^\top \in \Theta$. Note the following

$$\begin{aligned} A(f) = \log\left(\int_{\mathcal{X}} \mu(x) \exp(f(x)) \mathrm{d}x\right) &= \log\left(\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\bigl(\theta_1 x^2 + \theta_2 x\bigr) \mathrm{d}x\right) \\ &= \log\left(\frac{1}{\sqrt{-2\theta_1}} \exp\left(-\frac{\theta_2^2}{4\theta_1}\right)\right) \\ &= -\frac{\theta_2^2}{4\theta_1} - \frac{1}{2}\log(-2\theta_1) \end{aligned}$$

$$= B(\theta).$$

Since $A(f) < +\infty$ if and only if $\theta_1 < 0$ and $\theta_2 \in \mathbb{R}$, we have $\mathcal{F} = \left\{ f \in \mathcal{H}_0 \mid f(x) = \theta_1 x^2 + \theta_2 x \text{ for all } x \in \mathcal{X}, \theta_1 < 0, \theta_2 \in \mathbb{R} \right\}$. ▶

2.1.5 Assumptions on $\mathcal{H}$ and k and Their Implications

In the rest of this dissertation, we are going to make the following assumptions on $\mathcal{H}$, k and μ, unless explicitly stated otherwise:

(A1) The RKHS $\mathcal{H}$ is infinite-dimensional.

(A2) The kernel function k is continuous and bounded, i.e., $\kappa_1 := \sqrt{k(x,x)} < \infty$.

(A3) The RKHS $\mathcal{H}$ does *not* constant functions.

(A4) The base density μ is a continuously differentiable pdf over $\mathcal{X}$ with the support being $\mathcal{X}$.

If we assume $\mathcal{H}$ to be finite-dimensional, rather than infinite-dimensional as we have done in **(A1)**, we are returning to the case of finite-dimensional exponential families discussed in Section 2.1.1, as we have seen in Section 2.1.4. Classic theories on the estimation in $\mathcal{Q}_{\text{fin}}$ are directly applicable and are not very interesting. Thus, we rule out this case.

The main motivation of **(A2)** is to ensure $\mathcal{F} = \mathcal{H}$, as we have shown in Proposition 2.1. This is going to make all the optimization problems considered in Section 2.2 unconstrained. In addition, since $\mathcal{X} \subseteq \mathbb{R}^d$ and k is continuous, Lemma 4.33 in Steinwart and Christmann (2008) implies that $\mathcal{H}$ is separable and has an orthonormal basis (Theorem 11 in Royden and Fitzpatrick, 2018).

Assumptions **(A3)** and **(A4)** together ensure the identifiability of $\mathcal{Q}_{\text{ker}}$, i.e., $q_{f_1} = q_{f_2}$ if and only if $f_1 = f_2$. One sufficient condition that guarantees $\mathcal{H}$ does not contain

constant functions is that the kernel function k is continuous on $\mathcal{X} \times \mathcal{X}$ and vanishes at infinity (see Remark 3(iii) in Sriperumbudur et al., 2017).

2.2 Nonparametric Density Estimation in $\mathcal{Q}_{\text{ker}}$

We now turn to the nonparametric density estimation problem in $\mathcal{Q}_{\text{ker}}$. Borrowing the framework in Chapter 1, we consider the following minimization problem

$$\underset{q_f \in \mathcal{Q}_{\text{ker}}}{\text{minimize}} \left\{ \widehat{L}(q_f) + \frac{\lambda}{2} \widetilde{P}(f) \right\}. \tag{2.8}$$

The very first question we consider is how rich $\mathcal{Q}_{\text{ker}}$ is as a class of pdfs to estimate p_0. More specifically, we want to understand what class of density functions over $\mathcal{X}$ can be approximated arbitrarily well by those in $\mathcal{Q}_{\text{ker}}$. Proposition 1, Corollary 2 and Proposition 13 in Sriperumbudur et al. (2017) answered this question and state that, under certain regularity conditions, $\mathcal{Q}_{\text{ker}}$ can approximate arbitrarily well any continuous p_0 that vanishes at infinity under the KL-divergence, L^r norm with $r \in [1, \infty]$, the Hellinger distance, and the H-divergence. Hence, $\mathcal{Q}_{\text{ker}}$ is a rather rich class of density functions to estimate p_0.

In the rest of this section, we again consider the two loss functionals, $\widehat{L}_{\text{NLL}}$ and $\widehat{L}_{\text{SM}}$, and give a review of density estimation problem in $\mathcal{Q}_{\text{ker}}$ using them.

2.2.1 Density Estimation in $\mathcal{Q}_{\text{ker}}$ using $\widehat{L}_{\text{NLL}}$

We let $\widehat{L}$ in (2.8) to be the NLL loss functional $\widehat{L}_{\text{NLL}}$. Then, using (2.3), $\widehat{L}_{\text{NLL}}(q_f)$ becomes

$$\widehat{J}_{\text{NLL}}(f) := A(f) - \frac{1}{n} \sum_{i=1}^{n} f(X_i), \qquad \text{for all } f \in \mathcal{F},$$

up to an additive constant.

Since A is convex by Proposition 2.2, and $-\frac{1}{n}\sum_{i=1}^{n} f(X_i) = -\frac{1}{n}\sum_{i=1}^{n} \langle f, k(X_i, \cdot)\rangle_{\mathcal{H}}$ is linear and, hence, convex in f, their sum, $\widehat{J}_{\text{NLL}}$, is convex as well.

We first have a bad news stated in the following proposition.

Proposition 2.6 (Fukumizu (2005)). *Under* ***(A1)*** *and* ***(A2)****, minimizing* $\widehat{J}_{\mathrm{NLL}}$ *over* $\mathcal{H}$ *does* not *have a solution.*

The direct consequence of Proposition 2.6 is that the ML density estimator in $\mathcal{Q}_{\mathrm{ker}}$ does *not* exist.

Remark 2.2. Proposition 2.6 exemplifies a distinct difference between $\mathcal{Q}_{\mathrm{ker}}$ and $\mathcal{Q}_{\mathrm{fin}}$.

Suppose we estimate p_0 using elements in $\mathcal{Q}_{\mathrm{fin}}$ via maximizing the log-likelihood function

$$\left\langle \theta, \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)\right\rangle - B(\theta), \qquad \text{subject to } \theta \in \Theta. \tag{2.9}$$

Under certain regularity conditions, the maximizer of (2.9), denoted by $\hat{\theta}_{\mathrm{ML}}$, exists and is unique, and must satisfy the equation

$$\nabla B(\hat{\theta}_{\mathrm{ML}}) = \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i),$$

which has the moment matching interpretation that the sample mean of the canonical statistic must match the population mean at the MLE, since $\nabla B(\hat{\theta}_{\mathrm{ML}}) = \int_{\mathcal{X}} \tilde{q}_{\hat{\theta}}(x)\varphi(x)\mathrm{d}x$. In particular, the inverse map of ∇B exists and

$$\hat{\theta}_{\mathrm{ML}} = (\nabla B)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)\right).$$

▶

With the result in Proposition 2.6, the density estimation problem in $\mathcal{Q}_{\mathrm{ker}}$ via minimizing $\widehat{J}_{\mathrm{NLL}}$ is ill-posed. In order to make the problem well-posed and obtain a solution, one has to impose certain kind of regularization. Serval ideas have been proposed in the literature to tackle this issue.

The first idea is to regularize the function space over which we minimize $\widehat{J}_{\mathrm{NLL}}$. Rather than minimize it over the entire $\mathcal{H}$, Fukumizu (2005) proposed to construct a sequence of nested finite-dimensional subspaces of $\mathcal{H}$ that enlarges with the sample size n, $\{\mathcal{H}^{(m_n)}\}_{m_n \in \mathbb{N}}$ satisfying $\mathcal{H}^{(m_n)} \subset \mathcal{H}^{(m_{n+1})}$ for all n, and minimize $\widehat{J}_{\mathrm{NLL}}$ over $\mathcal{H}^{(m_n)}$. Supposing $\hat{f}_{\mathrm{ML}}^{(m_n)} := \arg\min_{f \in \mathcal{H}^{(m_n)}} \widehat{J}_{\mathrm{NLL}}(f)$ exists for all n, Fukumizu (2005) showed that, if $p_0 \in \mathcal{Q}_{\mathrm{ker}}$, together with some additional assumptions, $q_{\hat{f}_{\mathrm{ML}}^{(m_n)}}$ is consistent for p_0 under the KL-divergence.

Even though this approach is theoretically interesting, it suffers from several theoretical and practical drawbacks. On the theoretical side, the consistency of $q_{\hat{f}_{\mathrm{ML}}^{(m_n)}}$ relies on the decay rate of the smallest eigenvalue of certain covariance operator, which is hard or even impossible to check in practice. On the practical side, Fukumizu (2005) did not elucidate guidelines on which class of RKHS should be used or how to choose the sequence of nested finite-dimensional subspaces. Moreover, even if such guidelines were provided, the minimization problem is nonlinear by its nature and one has to rely on an iterative optimization algorithm to compute $\hat{f}_{\mathrm{ML}}^{(m_n)}$. Then, it is inevitable to deal with A and its derivative, both of which involve integration over a possibly high-dimensional space and are hard to handle in practice. Thus, the density estimator constructed using this approach is not attractive.

A different approach proposed in the literature is to add a nonzero penalty functional $\widetilde{P}$ to $\widehat{J}_{\mathrm{NLL}}$ and minimize the penalized NLL loss functional. One such work was carried out by Gu and Qiu (1993) who chose $\widetilde{P}$ to be a square seminorm of $\mathcal{H}$. They showed the minimizer of $\widehat{J}_{\mathrm{NLL}}(f) + \lambda \widetilde{P}(f)$ with $f \in \mathcal{H}$ exists and is unique under very mild conditions. However, since $\mathcal{H}$ is infinite-dimensional, this minimizer is *not* computable. They proposed to minimize the penalized NLL loss functional over $\mathcal{H}_0 \oplus \widetilde{\mathcal{H}}_n$, where $\mathcal{H}_0 := \left\{ f \in \mathcal{H} \mid \widetilde{P}(f) = 0 \right\}$ is the null space of $\widetilde{P}$ and $\widetilde{\mathcal{H}}_n := \{ f \mid f :=$

$\sum_{i=1}^{n} \alpha_i k(X_i, \cdot), \alpha_1, \cdots, \alpha_n \in \mathbb{R}\}$ is an n-dimensional subspace of $\mathcal{H}$, and established asymptotic properties of $q_{\tilde{f}_{\mathrm{ML}}^{(\lambda)}}$, where $\tilde{f}_{\mathrm{ML}}^{(\lambda)} := \arg\min_{f \in \mathcal{H}_0 \oplus \widetilde{\mathcal{H}}_n} \{\widehat{J}_{\mathrm{NLL}}(f) + \lambda \widetilde{P}(f)\}$. Gu (1993) proposed an iterative algorithm to compute $\tilde{f}_{\mathrm{ML}}^{(\lambda)}$ and used the quadrature rule over a dense mesh to approximate A and its derivatives at each iteration. All numerical examples therein were limited to cases $d \leq 2$. If d is large, the computation would become prohibitively expensive and the approximations via this approach could be very poor.

In order to avoid working with A directly, Dai et al. (2018) proposed a doubly dual embedding approach. Suppose k satisfies $\int_{\mathcal{X}} k(x, x)\mathrm{d}x < \infty$ so that any $f \in \mathcal{H}$ is square-integrable. We then have

$$A(f) = \sup_{p \in \mathcal{P} \cap L^2(\mathcal{X})} \Big\{ \langle p, f \rangle_{L^2(\mathcal{X})} - \mathrm{KL}(p \| \mu) \Big\}, \qquad \text{for all } f \in \mathcal{H}, \tag{2.10}$$

$$\mathrm{KL}(p \| \mu) = \sup_{g \in \mathcal{H}} \Big\{ \langle p, g \rangle_{L^2(\mathcal{X})} - \int_{\mathcal{X}} e^{g(x)} \mu(x)\mathrm{d}x + 1 \Big\}, \quad \text{for all } p \in \mathcal{P} \cap L^2(\mathcal{X}), \tag{2.11}$$

where $\mathcal{P}$ denotes the set of all pdfs over $\mathcal{X}$, $L^2(\mathcal{X}) := \big\{f : \mathcal{X} \to \mathbb{R} \mid \int_{\mathcal{X}} (f(x))^2 \mathrm{d}x < \infty\big\}$, and $\langle f_1, f_2 \rangle_{L^2(\mathcal{X})} := \int_{\mathcal{X}} f_1(x) f_2(x) \mathrm{d}x$ for all $f_1, f_2 \in L^2(\mathcal{X})$. Plugging (2.10) and (2.11) successively into $\widehat{J}_{\mathrm{NLL}}(f) + \lambda \widetilde{P}(f)$, Dai et al. (2018) proposed to solve the following max-min problem

$$\underset{p \in \mathcal{P} \cap L^2(\mathcal{X})}{\text{maximize}} \ \underset{f, g \in \mathcal{H}}{\text{minimize}} \Big\{ -\frac{1}{n} \sum_{i=1}^{n} f(X_i) + \int_{\mathcal{X}} (f(x) - g(x)) p(x) \mathrm{d}x + \int_{\mathcal{X}} e^{g(x)} \mu(x) \mathrm{d}x + \lambda \widetilde{P}(f) \Big\},$$

and proposed a stochastic gradient ascent-descent algorithm to iterate between the inner and outer optimization problems until convergence. While this approach avoids directly working with A, it incurs additional computational burdens as, in order to compute the minimizer with respect to f, one has to compute the optimal solutions with respect to p and g at the same time.

2.2.2 Density estimation in $\mathcal{Q}_{\text{ker}}$ using $\widehat{L}_{\text{SM}}$

With the discussions in the preceding subsection, we see that the main difficulty in the approach via minimizing $\widehat{J}_{\text{NLL}}$ is that one has to deal with A and its derivative, which is computationally intractable in practice. The SM loss functional, as we have discussed in Chapter 1, can help us avoid this difficulty.

Recall that, under certain conditions, the SM loss functional, in its original form, is

$$\widehat{L}_{\text{SM}}(q) := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\big(\partial_u \log q(X_i)\big)^2 + \partial_u^2 \log q(X_i)\right). \tag{2.12}$$

If we let $q = q_f \in \mathcal{Q}_{\text{ker}}$ in (2.12), under certain additional regularity conditions that will be made explicit in Chapter 3, we can rewrite $\widehat{L}_{\text{SM}}$, up to additive constants, as

$$\begin{aligned}\widehat{J}_{\text{SM}}(f) &:= \frac{1}{2}\sum_{i=1}^{n}\sum_{u=1}^{d}(\partial_u f(X_i))^2 + \sum_{i=1}^{n}\sum_{u=1}^{d}\big(\partial_u \log \mu(X_i)\partial_u f(X_i) + \partial_u^2 f(X_i)\big)\\ &\overset{(\star)}{=} \frac{1}{2}\langle f, \widehat{C} f\rangle_{\mathcal{H}} - \langle f, \hat{z}\rangle_{\mathcal{H}},\end{aligned} \tag{2.13}$$

where $\widehat{C} : \mathcal{H} \to \mathcal{H}$ is given by

$$\widehat{C} := \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\partial_u k(X_i, \cdot) \otimes \partial_u k(X_i, \cdot), \tag{2.14}$$

with $\widehat{C} f = \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\partial_u f(X_i)\partial_u k(X_i, \cdot)$ for all $f \in \mathcal{H}$, and

$$\hat{z} := -\frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\big(\partial_u \log \mu(X_i)\partial_u k(X_i, \cdot) + \partial_u^2 k(X_i, \cdot)\big) \in \mathcal{H}. \tag{2.15}$$

In (2.14) and (2.15), with a fixed $x \in \mathcal{X}$, $\partial_u^s k(x, y) := \frac{\partial^s}{\partial w_u^s} k(x, y)\big|_{w=x}$, for all $y \in \mathcal{X}$, $s = 1, 2$, and $u = 1, \cdots, d$. In $(\star)$, we use $\partial_u^s k(x, \cdot) \in \mathcal{H}$ and the reproducing property of partial derivatives of k, $\partial_u^s f(x) = \langle f, \partial_u^s k(x, \cdot)\rangle_{\mathcal{H}}$, for $s = 1, 2$ and $u = 1, \cdots, d$ (Zhou, 2008). Details about the partial derivatives of k and their reproducing properties can be found in Section A.3 in Appendix A. In particular, notice that $\widehat{J}_{\text{SM}}$ does *not* involve A.

Now, $\widehat{J}_{\mathrm{SM}}$ is a quadratic functional in f. It is not hard to show that the operator $\widehat{C}$ is linear, self-adjoint, and positive semidefinite. Hence, minimizing $\widehat{J}_{\mathrm{SM}}$ over $\mathcal{H}$ is a convex optimization problem. Suppose that $\hat{f}_{\mathrm{SM}} := \arg\min_{f\in\mathcal{H}} \widehat{J}_{\mathrm{SM}}(f)$ exists. Then, $\hat{f}_{\mathrm{SM}}$ must satisfy the first-order optimality condition $\widehat{C}\hat{f}_{\mathrm{SM}} = \hat{z}$; in other words, minimizing $\widehat{J}_{\mathrm{SM}}$ amounts to solving an infinite-dimensional linear system, and $\hat{f}_{\mathrm{SM}} = \widehat{C}^{-1}\hat{z}$. However, since $\widehat{C}$ has finite rank and must be a compact operator in an infinite-dimensional RKHS $\mathcal{H}$, it is *not* invertible (Section 16.5 in Royden and Fitzpatrick, 2018) and, hence, $\hat{f}_{\mathrm{SM}}$ does *not* exist.

Thus, minimizing $\widehat{J}_{\mathrm{SM}}$ is an ill-posed problem. In order to remedy this, certain kind of regularization has to be imposed. Sriperumbudur et al. (2017) proposed to add a penalty term and minimize

$$\widehat{J}_{\mathrm{SM}}(f) + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2, \qquad \text{subject to } f \in \mathcal{H},$$

where we use $\rho > 0$ to denote the penalty parameter associated with the SM loss functional. This is exactly the Tikhonov regularization. Sriperumbudur et al. (2017) showed this penalized SM loss functional has a unique minimizer given by

$$\hat{f}_{\mathrm{SM}}^{(\rho)} := \arg\min_{f\in\mathcal{H}} \left\{ \widehat{J}_{\mathrm{SM}}(f) + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2 \right\} = (\widehat{C} + \lambda I)^{-1}\hat{z}, \tag{2.16}$$

where $I : \mathcal{H} \to \mathcal{H}$ denotes the identity operator in $\mathcal{H}$. In practice, however, it may not be easy to compute the minimizer in the form of (2.16) as it involves solving an infinite-dimensional linear system. With the help of a general representer theorem (Theorem A.2 in Sriperumbudur et al., 2017), it can be shown that

$$\hat{f}_{\mathrm{SM}}^{(\rho)} = \sum_{i=1}^{n}\sum_{u=1}^{d} \alpha_{(i-1)d+u}^{(\rho)} \partial_u k(X_i, \cdot) + \frac{1}{\rho}\hat{z},$$

where $\boldsymbol{\alpha}^{(\rho)} := (\alpha_1^{(\rho)}, \cdots, \alpha_{nd}^{(\rho)})^\top \in \mathbb{R}^{nd}$ can be obtained by solving the linear system

$$(\mathbf{G} + n\rho\mathbf{I}_{nd})\boldsymbol{\alpha}^{(\rho)} = \frac{1}{\rho}\mathbf{h},$$

the $((i-1)d+u,(j-1)+v)$-entry of $\mathbf{G} \in \mathbb{R}^{nd\times nd}$ is $\langle \partial_u k(X_i,\cdot), \partial_v k(X_j,\cdot)\rangle_{\mathcal{H}} = \partial_i\partial_{i+d}k(X_i,X_j)$, and $((i-1)d+u)$-entry of $\mathbf{h} \in \mathbb{R}^{nd}$ is

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

and $\mathbf{I}_{nd}$ denotes the $nd\times nd$ identity matrix. Note that the matrix $\mathbf{G}$ is positive semidefinite and $n\rho\mathbf{I}_{nd}$ is positive definite, so their sum must be positive definite and must be invertible, and it follows that $\boldsymbol{\alpha}^{(\rho)} = \frac{1}{\rho}(\mathbf{G}+n\rho\mathbf{I}_{nd})^{-1}\mathbf{h}$.

Theoretical properties of the penalized SM density estimator $q_{\hat{f}_{\mathrm{SM}}^{(\rho)}}$ was studied by Sriperumbudur et al. (2017). If $p_0 \in \mathcal{Q}_{\mathrm{ker}}$, they established the consistency and convergence rate of $q_{\hat{f}_{\mathrm{SM}}^{(\rho)}}$ under the KL-divergence, the H-divergence, the Hellinger distance, and the total variation distance. If $p_0 \notin \mathcal{Q}_{\mathrm{ker}}$, they showed $q_{\hat{f}_{\mathrm{SM}}^{(\rho)}}$ converges to the element in $\mathcal{Q}_{\mathrm{ker}}$ that has the smallest H-divergence to p_0 and established the corresponding convergence rate under the H-divergence.

In their simulation studies, Sriperumbudur et al. (2017) showed their penalized SM density estimator outperforms the kernel density estimator (Section 6.3 in Wasserman, 2006), especially when d is large. However, no comparison with the (penalized) ML density estimator was conducted.

Note that computing $\hat{f}_{\mathrm{SM}}^{(\rho)}$ requires to solve a linear system of nd equations in nd variables, which requires elementary operations of order $\mathcal{O}(n^3d^3)$. This can be computationally expensive when n or d is large. To alleviate the computational cost, Sutherland et al. (2017) adopted the idea of the Nyström approximation (Williams and Seeger, 2001) and proposed to minimize $\widehat{J}_{\mathrm{SM}}(f) + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2$ subject to

$$f \in \mathrm{Span}\Big\{\partial_u k(Y_j,\cdot) \text{ for all } j=1,\cdots,m \text{ and } u=1,\cdots,d\Big\},$$

where $Y_1,\cdots,Y_m$ are m randomly selected observations from $\{X_1,\cdots,X_n\}$ with $m \ll n$. Then, one only needs to work with a linear system of md equations in md variables,

which requires elementary operations of order $\mathcal{O}(m^3 d^3)$. This penalized SM density estimator is more efficient to compuate and empirically performs very close to the one proposed by Sriperumbudur et al. (2017).

2.3 Proofs

2.3.1 Proof of Proposition 2.1

Proof. Let $f \in \mathcal{H}$ be arbitrary. Due to the Cauchy-Schwartz inequality, we have $\left|\langle f, k(x, \cdot)\rangle_{\mathcal{H}}\right| \leq \|f\|_{\mathcal{H}}\|k(x, \cdot)\|_{\mathcal{H}} \leq \kappa_1\|f\|_{\mathcal{H}}$. Then, we can bound $A(f)$ from above by

$$A(f) \leq \log\left(\exp(\kappa_1\|f\|_{\mathcal{H}})\int_{\mathcal{X}}\mu(x)\mathrm{d}x\right) \overset{(\star)}{=} \kappa_1\|f\|_{\mathcal{H}} < \infty,$$

where we use the assumption that μ is a pdf over $\mathcal{X}$ in deriving $(\star)$. Since the choice of $f \in \mathcal{H}$ is arbitrary, we conclude $\mathcal{H} \subseteq \mathcal{F}$. On the other hand, it is obvious that $\mathcal{F} \subseteq \mathcal{H}$. We conclude that $\mathcal{F} = \mathcal{H}$. ∎

2.3.2 Proof of Proposition 2.2

Proof of Proposition 2.2. We first show the convexity of A. That is, we need to show, for any distinct $f, g \in \mathcal{F}$ and $\alpha \in [0, 1]$, the following inequality holds

$$A(\alpha f + (1-\alpha)g) \leq \alpha A(f) + (1-\alpha)A(g).$$

Notice that if $\alpha = 1$ or $\alpha = 0$, the inequality above becomes an equality and the result holds trivially.

Hence, we assume $\alpha \in (0, 1)$. Then,

$$\begin{aligned}
A(\alpha f + (1-\alpha)g) &= \log\left[\int_{\mathcal{X}}\big(\mu(x)\exp(f(x))\big)^{\alpha}\big(\mu(x)\exp(g(x))\big)^{1-\alpha}\mathrm{d}x\right] \\
&\overset{(\star)}{\leq} \log\left[\left(\int_{\mathcal{X}}\big(\mu(x)\exp(f(x))\big)^{\alpha\cdot\frac{1}{\alpha}}dx\right)^{\alpha}\cdot\left(\int_{\mathcal{X}}\big(\mu(x)\exp(g(x))\big)^{(1-\alpha)\cdot\frac{1}{1-\alpha}}\mathrm{d}x\right)^{1-\alpha}\right] \\
&= \alpha\log\left[\int_{\mathcal{X}}\mu(x)\exp(f(x))\mathrm{d}x\right] + (1-\alpha)\log\left[\int_{\mathcal{X}}\mu(x)\exp(g(x))\mathrm{d}x\right] \\
&= \alpha A(f) + (1-\alpha)A(g),
\end{aligned}$$

where $(\star)$ is due to the Hölder's inequality. This established the convexity of A.

The inequality ($\star$) becomes an equality if and only if there exist $\beta_1 > 0$ and $\beta_2 > 0$ such that $\beta_1 \exp(f(x)) = \beta_2 \exp(g(x))$ for almost all $x \in \mathcal{X}$, or, equivalently,

$$f(x) - g(x) = \langle f - g, k(x, \cdot)\rangle_{\mathcal{H}} = \log \beta_2 - \log \beta_1, \qquad \text{for almost all } x \in \mathcal{X}.$$

Now, if $\mathcal{H}$ does *not* contain constant functions, $f - g$ cannot be a constant function, meaning that the equality cannot hold. Thus, A is strictly convex.

Finally, we show the convexity of $\mathcal{F}$. Let $f, g \in \mathcal{F}$ so that $A(f) < \infty$ and $A(g) < \infty$. By the proof above, we have $A(\alpha f + (1-\alpha)g) < \infty$, i.e, $\alpha f + (1-\alpha)g \in \mathcal{F}$. In other words, $\mathcal{F}$ is convex. ∎

2.3.3 Proof of Lemma 2.1

Proof of Lemma 2.1. Let $f \in \mathcal{H}$ be arbitrary and $g \in \mathcal{H}$ with $g \neq 0$. Note that

$$J_x(f + g) - J_x(f) = \langle f + g, k(x, \cdot)\rangle_{\mathcal{H}} - \langle f, k(x, \cdot)\rangle_{\mathcal{H}} = \langle g, k(x, \cdot)\rangle_{\mathcal{H}}.$$

Then,

$$\frac{\left|J_x(f + g) - J_x(f) - \langle g, k(x, \cdot)\rangle_{\mathcal{H}}\right|}{\|g\|_{\mathcal{H}}} = 0,$$

and, by Definition A.1 in Appendix A, we conjecture $\mathrm{D}J_x(f)(g) = \langle g, k(x, \cdot)\rangle_{\mathcal{H}}$, for all $g \in \mathcal{H}$.

We next verify that the map $\mathrm{D}J_x(f)$ is linear and bounded. The linearity part follows from the linearity of inner product and is omitted. To establish the boundedness, we have, for all $g \in \mathcal{H}$,

$$|\mathrm{D}J_x(f)(g)| = |\langle g, k(x, \cdot)\rangle_{\mathcal{H}}| \leq \|g\|_{\mathcal{H}} \|k(x, \cdot)\|_{\mathcal{H}} \leq \kappa_1 \|g\|_{\mathcal{H}},$$

where $\kappa_1 := \sup_{x \in \mathcal{X}} \sqrt{k(x, x)} < \infty$, by the assumption. Hence, $\mathrm{D}J_x(f)$ is a bounded linear operator. We conclude that J_x is Fréchet differentiable at $f \in \mathcal{H}$ with the

Fréchet derivative at $f \in \mathcal{H}$ being $\mathrm{D}J_x(f)(g) = \langle g, k(x, \cdot)\rangle_{\mathcal{H}}$ for all $g \in \mathcal{H}$. Since the choice of $f \in \mathcal{H}$ is arbitrary, we conclude J_x is Fréchet differentiable over $\mathcal{H}$.

Using the definition of Fréchet gradient, it is easy to see $\nabla J_x(f) = k(x, \cdot)$ for all $f \in \mathcal{H}$. ∎

2.3.4 Proof of Proposition 2.3

Proof of Proposition 2.3. Observe that $A(f) = (J_2 \circ J_1)(f)$ for all $f \in \mathcal{H}$, where

$$\begin{aligned} J_2(x) &:= \log x && \text{for all } x > 0, \\ J_1(f) &:= \int_{\mathcal{X}} \mu(x) \exp(f(x)) \mathrm{d}x && \text{for all } f \in \mathcal{H}. \end{aligned}$$

Since $J_1(f) > 0$ for all $f \in \mathcal{H}$, this composition is well-defined.

Let $f \in \mathcal{H}$ be arbitrary. We first show J_1 is Fréchet differentiable at $f \in \mathcal{H}$ with the Fréchet derivative

$$\mathrm{D}J_1(f)(g) = \left\langle g, \int_{\mathcal{X}} \mu(x) \exp(f(x)) k(x, \cdot) \mathrm{d}x \right\rangle_{\mathcal{H}}, \qquad \text{for all } g \in \mathcal{H}.$$

Note that the integrand of $\int_{\mathcal{X}} \mu(x) \exp(f(x)) k(x, \cdot) \mathrm{d}x$ is an element in $\mathcal{H}$, and

$$\begin{aligned} \left\| \int_{\mathcal{X}} \mu(x) \exp(f(x)) k(x, \cdot) \mathrm{d}x \right\|_{\mathcal{H}} &\leq \int_{\mathcal{X}} \mu(x) \exp(f(x)) \|k(x, \cdot)\|_{\mathcal{H}} \mathrm{d}x \\ &\leq \kappa_1 \exp(A(f)) < \infty, \end{aligned}$$

where $\kappa_1 := \sup_{x \in \mathcal{X}} \sqrt{k(x, x)} < \infty$ by the assumption. Hence, $\int_{\mathcal{X}} \mu(x) \exp(f(x)) k(x, \cdot) \mathrm{d}x$ is Bochner integrable with respect to the Lebesgue measure (by Proposition A.3 in Appendix A), and we can interchange the inner product and the integral (by Proposition A.4(c) in Appendix A) and have

$$\mathrm{D}J_1(f)(g) = \int_{\mathcal{X}} \mu(x) \exp(f(x)) g(x) \mathrm{d}x, \qquad \text{for all } g \in \mathcal{F},$$

Now, let $0 \neq g \in \mathcal{H}$. We have

$$J_1(f + g) - J_1(f) - \mathrm{D}J_1(f)(g)$$

$$
\begin{aligned}
&= \int_{\mathcal{X}} \mu(x) \exp\big(f(x)\big) \Big[\exp(g(x)) - 1 - g(x)\Big] \mathrm{d}x \\
&= \int_{\mathcal{X}} \mu(x) \exp(f(x)) \left[\sum_{j=0}^{\infty} \frac{(g(x))^m}{m!} - 1 - g(x)\right] \mathrm{d}x \\
&= \int_{\mathcal{X}} \mu(x) \exp(f(x)) \left[\sum_{m=2}^{\infty} \frac{(g(x))^m}{m!}\right] \mathrm{d}x \\
&\overset{\text{(i)}}{\leq} \int_{\mathcal{X}} \mu(x) \exp\big(\|f\|_{\mathcal{H}} \sqrt{k(x,\,\cdot\,)}\big) \left[\sum_{m=2}^{\infty} \frac{\|g\|_{\mathcal{H}}^m k(x,x)^{m/2}}{m!}\right] \mathrm{d}x \\
&\overset{\text{(ii)}}{\leq} \int_{\mathcal{X}} \mu(x) \exp(\kappa_1 \|f\|_{\mathcal{H}}) \left[\sum_{m=2}^{\infty} \frac{\|g\|_{\mathcal{H}}^m \kappa_1^{m/2}}{m!}\right] \mathrm{d}x \\
&\overset{\text{(iii)}}{=} \exp(\kappa_1 \|f\|_{\mathcal{H}}) \left[\sum_{m=2}^{\infty} \frac{\|g\|_{\mathcal{H}}^m \kappa_1^{m/2}}{m!}\right],
\end{aligned}
$$

where (i) follows from the Cauchy-Schwartz inequality, (ii) is due to $\sqrt{k(x,x)} \leq \kappa_1$, and (iii) is because μ is a density function over $\mathcal{X}$. To proceed, we have

$$
\frac{\big|J_1(f+g) - J_1(f) - \mathrm{D}J_1(f)(g)\big|}{\|g\|_{\mathcal{H}}} \leq \exp(\kappa_1 \|f\|_{\mathcal{H}}) \left[\sum_{m=2}^{\infty} \frac{\|g\|_{\mathcal{H}}^{m-1} \kappa_1^{m/2}}{m!}\right] \to 0,
$$

as $\|g\|_{\mathcal{H}} \to 0$.

Furthermore, we need to show $\mathrm{D}J_1(f)$ is linear and bounded. The linearity follows from that of the inner product and is omitted. To show the boundedness, we let $g \in \mathcal{H}$ be arbitrary and notice

$$
\begin{aligned}
\big|\mathrm{D}J_1(f)(g)\big|_{\mathcal{H}} &\leq \|g\|_{\mathcal{H}} \int_{\mathcal{X}} \mu(x) \exp\big(\|f\|_{\mathcal{H}} \sqrt{k(x,x)}\big) \sqrt{k(x,x)} \mathrm{d}x \\
&\leq \Big[\kappa_1 \exp(\kappa_1 \|f\|_{\mathcal{H}})\Big] \|g\|_{\mathcal{H}},
\end{aligned}
$$

from which we conclude that $\mathrm{D}J_1(f)$ is a bounded operator. Thus, J_1 is Fréchet differentiable at $f \in \mathcal{H}$ with the desired Fréchet derivative.

Since $A = J_2 \circ J_1$, applying Proposition A.2(b) in Appendix A, we obtain, for any $f \in \mathcal{H}$,

$$
\mathrm{D}A(f)(g) = \frac{1}{J_1(f)} \mathrm{D}J_1(f)(g)
$$

$$= \frac{1}{\exp(A(f))} \left\langle g, \int_{\mathcal{X}} \mu(x) \exp(\langle f, k(x, \cdot)\rangle_{\mathcal{H}}) k(x, \cdot) \mathrm{d}x \right\rangle_{\mathcal{H}}$$
$$= \left\langle g, \int_{\mathcal{X}} q_f(x) k(x, \cdot) \mathrm{d}x \right\rangle_{\mathcal{H}}.$$

Since our choice of $f \in \mathcal{H}$ is arbitrary, we conclude A is Fréchet differentiable over $\mathcal{H}$.

Finally, by the definition of Fréchet gradient operator, we conclude that the Fréchet gradient operator is $\int_{\mathcal{X}} q_f(x) k(x, \cdot) \mathrm{d}x$ as claimed. This completes the proof.

■

2.3.5 Proof of Proposition 2.5

Proof of Proposition 2.5. We will show the desired result by the three steps:

(a) show $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ is an inner product space,

(b) show $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ is a Hilbert space, and

(c) show $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ is a RKHS.

(a) To show $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ is an inner product space, we verify using the definition of an inner product space. Let $f = \sum_{j=1}^m \alpha_j \varphi_j \in \mathcal{H}_0$, $g = \sum_{j=1}^m \beta_j \varphi_j \in \mathcal{H}_0$, $h = \sum_{j=1}^m \gamma_j \varphi_j \in \mathcal{H}_0$, $\alpha_j, \beta_j, \gamma_j \in \mathbb{R}$ for all $j = 1, \cdots, m$, and $a, b \in \mathbb{R}$ be arbitrary. Then, the operation $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is

(i) *symmetric*, since $\langle f, g \rangle_{\mathcal{H}_0} = \sum_{j=1}^m \alpha_j \beta_j = \sum_{j=1}^m \beta_j \alpha_j = \langle g, f \rangle_{\mathcal{H}_0}$;

(ii) *linear*, since

$$\begin{aligned}
\langle af + bg, h \rangle_{\mathcal{H}_0} &= \left\langle a \sum_{j=1}^m \alpha_j \varphi_j + b \sum_{j=1}^m \beta_j \varphi_j, \sum_{j=1}^m \gamma_j \varphi_j \right\rangle_{\mathcal{H}_0} \\
&= \left\langle \sum_{j=1}^m (a\alpha_j + b\beta_j) \varphi_j, \sum_{j=1}^m \gamma_j \varphi_j \right\rangle_{\mathcal{H}_0} \\
&= \sum_{j=1}^m (a\alpha_j + b\beta_j) \gamma_j
\end{aligned}$$

$$= a\sum_{j=1}^{m}\alpha_j\gamma_j + b\sum_{j=1}^{m}\beta_j\gamma_j$$
$$= a\langle f, h\rangle_{\mathcal{H}_0} + b\langle g, h\rangle_{\mathcal{H}_0};$$

(iii) *positive definite*, since $\langle f, f\rangle_{\mathcal{H}_0} = \sum_{j=1}^{m}\alpha_j^2 \geq 0$ for all $f \in \mathcal{H}_0$, and $\langle f, f\rangle_{\mathcal{H}_0} = 0$ if and only if $\alpha_j = 0$ for all $j = 1, \cdots, m$, implying $f = 0$.

(b) To show $(\mathcal{H}_0, \langle\,\cdot\,,\,\cdot\,\rangle_{\mathcal{H}_0})$ is a Hilbert space, first note that the inner product space $(\mathcal{H}_0, \langle\,\cdot\,,\,\cdot\,\rangle_{\mathcal{H}_0})$ is m-dimensional. The desired result follows directly from the fact that any finite-dimensional inner product space over $\mathbb{R}$ is complete and the definition that a Hilbert space is a complete inner product space.

(c) Finally, to show $(\mathcal{H}_0, \langle\,\cdot\,,\,\cdot\,\rangle_{\mathcal{H}_0})$ is a RKHS with the reproducing kernel (2.7), we let $x \in \mathcal{X}$ be fixed and $E_x : \mathcal{H}_0 \to \mathbb{R}$ be the evaluation functional, i.e., $E_x(f) = f(x)$ for all $f \in \mathcal{H}_0$, and need to show E_x is a bounded operator.

To this end, let $f, g \in \mathcal{H}_0$ be arbitrary and $a, b \in \mathbb{R}$, and note the following

$$|E_x(f)| = |f(x)| = \left|\sum_{j=1}^{m}\alpha_j\varphi_j(x)\right| \leq \sqrt{\sum_{j=1}^{m}\alpha_j^2}\sqrt{\sum_{j=1}^{m}(\varphi_j(x))^2}$$
$$= \|f\|_{\mathcal{H}_0}\sqrt{\sum_{j=1}^{m}(\varphi_j(x))^2} = M_x\|f\|_{\mathcal{H}_0},$$

where $M_x := \sqrt{\sum_{j=1}^{m}(\varphi_j(x))^2} < \infty$. Hence, the evaluation functional E_x is a bounded operator, from which we conclude $\mathcal{H}_0$ is a RKHS.

Finally, we identify the reproducing kernel associated with $\mathcal{H}_0$. Let $k(x, \cdot\,) = \sum_{j=1}^{m}\varphi_j(x)\varphi_j \in \mathcal{H}_0$, so that $k(x, y)$ is given by (2.7). Letting $f = \sum_{j=1}^{m}\alpha_j\varphi_j \in \mathcal{H}_0$, we have

$$\langle f, k(x, \cdot\,)\rangle_{\mathcal{H}_0} = \left\langle \sum_{j=1}^{m}\alpha_j\varphi_j, \sum_{j=1}^{m}\varphi_j(x)\varphi_j \right\rangle_{\mathcal{H}_0} = \sum_{j=1}^{m}\alpha_j\varphi_j(x) = f(x).$$

Thus, k defined in (2.7) satisfies the reproducing property and is the reproducing kernel of $\mathcal{H}_0$. ∎

Chapter 3: Early Stopping Score Matching Density Estimator

In this chapter, we introduce the early stopping SM density estimator and discuss its statistical 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 adopted by Sriperumbudur et al. (2017) was to add a penalty functional and minimize the penalized SM loss functional, which 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 with the penalized SM density estimator 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.

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 resulting 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 small variance, but stopping it too late leads to a model with a small 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, the 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 $\{\tau_t\}_{t\in\mathbb{N}}$, 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 sequence $\{g(x^{(t)})\}_{t\in\mathbb{N}}$ is guaranteed to converge to $g(x^*)$ as $t \to \infty$ (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.

(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, 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, we can derive the H-divergence between p_0 and $q_f \in \mathcal{Q}_{\text{ker}}$ and the corresponding SM matching loss that we have seen in Chapter 2.

Theorem 3.1 (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*

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

where $\widehat{C} : \mathcal{H} \to \mathcal{H}$ and $\hat{z} \in \mathcal{H}$ are given by (2.14) and (2.15) 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 2.2.2 in Chapter 2, minimizing $\widehat{J}_{\mathrm{SM}}$ 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 to minimizing $\widehat{J}_{\mathrm{SM}}$ and terminate it early to regularize. Since $\widehat{J}_{\mathrm{SM}}$ 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 A.2 in Appendix A) and derive the Fréchet gradient of $\widehat{J}_{\mathrm{SM}}$ in the following theorem.

Theorem 3.2 (Fréchet gradient of $\widehat{J}_{\mathrm{SM}}$)**.** *Under the same assumptions in Theorem 3.1, the Fréchet gradient of $\widehat{J}_{\mathrm{SM}}$, denoted by $\nabla \widehat{J}_{\mathrm{SM}}$, is a map from $\mathcal{H}$ to $\mathcal{H}$ given by*

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

The proof of Theorem 3.2 can be found in Section 3.5.1.

With the result of Theorem 3.2, starting from some $\hat{f}_{\mathrm{SM}}^{(0)} \in \mathcal{H}$, the gradient descent iterates are

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

where $\tau_t \in (0, 1/(d\kappa_2^2))$ is the step size at the t-th iterate. We omit the subscript "SM" associated with $\hat{f}_{\mathrm{SM}}^{(t)}$ in the rest of this chapter for notational simplicity.

Using the definition of the second-order Fréchet derivative and gradient (Definition A.3 in Appendix A), we know $\widehat{J}_{\mathrm{SM}}$ 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 3.3. *Starting the gradient descent algorithm from $\hat{f}^{(0)} \in \mathcal{H}$, the $(t+1)$-st gradient descent iterate is*

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

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

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

The proof of Theorem 3.3 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 $\hat{f}^{(0)} = 0$, we can ignore the term $\prod_{i=0}^{t}(I - \tau_i \widehat{C})\hat{f}^{(0)}$, and (3.4) becomes

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

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 $\rho \to \infty$, $q_{\hat{f}_{\mathrm{SM}}^{(\rho)}} \to \mu$.

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 early stopping SM density estimates correspond to the penalized ones with smaller ρ values.

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

Even though Theorem 3.3 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 $\hat{f}^{(t)}$, 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 derive 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 $(t+1)$-st 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 $\hat{f}^{(t+1)}$ that helps compute it.

Theorem 3.4. *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*

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

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

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

In addition, let $\mathbf{G} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ *be 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* $\widetilde{\mathbf{\Lambda}}^{(t)} \in \mathbb{R}^{nd\times nd}$ *be a diagonal matrix with diagonal elements* $\tilde{\lambda}_1^{(t)}, \cdots, \tilde{\lambda}_{nd}^{(t)}$ *being*

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

for all $w = 1, \cdots, nd$. *Then, for all* $t \in \mathbb{N}_0$, *we have*

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

where $\alpha_{(i-1)d+u}^{(t)} = -\frac{\tau^2}{n}\left[\mathbf{Q}\widetilde{\mathbf{\Lambda}}^{(t)}\mathbf{Q}^\top\mathbf{h}\right]_{(i-1)d+u}$, *for all* $i = 1, \cdots, n$ *and* $u = 1, \cdots, d$.

The proof of Theorem 3.4 is provided in Section 3.5.3. Thus, from Theorem 3.4, $\hat{f}^{(t+1)}$ can be obtained by addition and multiplication of certain matrices.

Remark 3.1. We discuss an alternative implementation of the gradient descent algorithm.

Note that, with $\hat{f}^{(0)} = 0$, gradient descent iterates

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

belong 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, $\hat{f}^{(t+1)}$ for all $t \in \mathbb{N}_0$, must lie in the linear subspace

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

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 $\widehat{J}_{\mathrm{SM}}$ yields

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

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

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

Since the gradient vector of $\widetilde{J}_{\mathrm{SM}}$ at $\boldsymbol{\beta}$ is $\nabla\widetilde{J}_{\mathrm{SM}}(\boldsymbol{\beta}) = \frac{1}{n}\widetilde{\mathbf{G}}^\top\widetilde{\mathbf{G}}\boldsymbol{\beta} - \widetilde{\mathbf{h}}$, 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 $\tilde{\tau} \in \left(0, n/\lambda_{\max}(\widetilde{\mathbf{G}}^\top\widetilde{\mathbf{G}})\right)$ 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

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

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 Estimators

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

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 36 and 2, respectively. Plots of μ

and its logarithm are shown in Figure 3.1. We choose the kernel function to be the Gaussian kernel function

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

with the bandwidth parameter $\sigma = 5$. 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 $\tau = 20 \in (0, 1/(d\kappa_2^2)) = (0, 25)$.

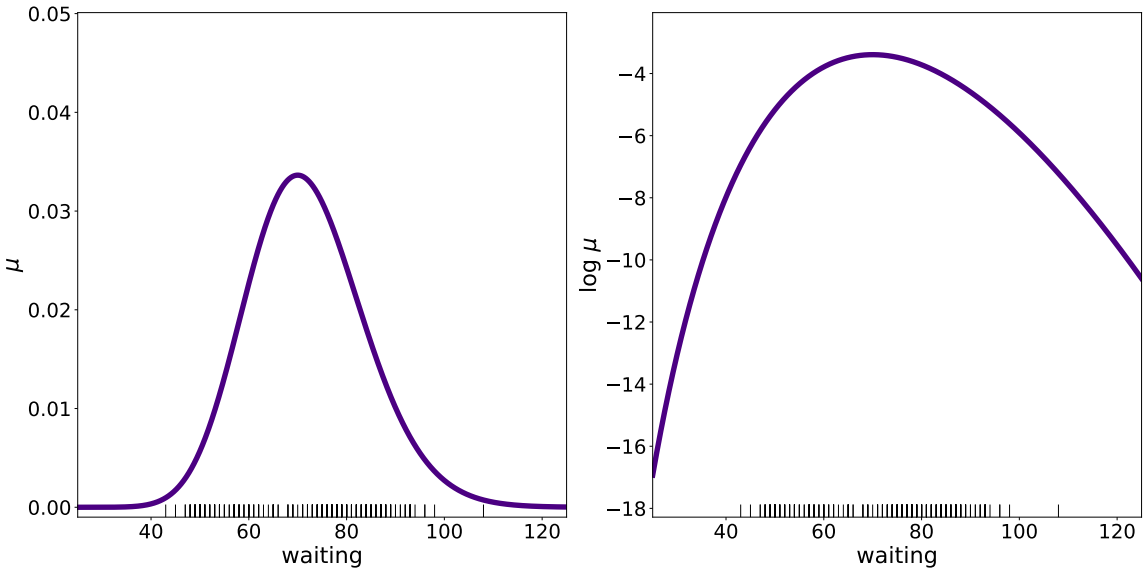

Figure 3.1: Left panel shows μ and right panel shows $\log \mu$. The rug plot indicates the location of data.

The resulting early stopping SM density estimates are shown in Figure 3.2. 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 isolated observation 108. We will

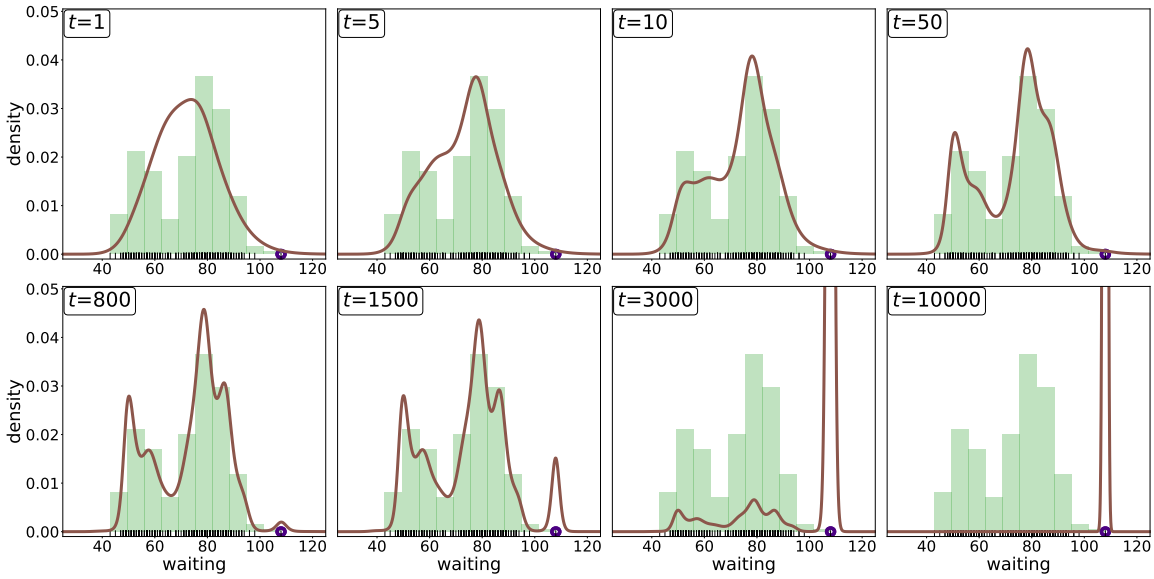

Figure 3.2: Early stopping SM density estimates for different values of number of iterations labeled at the upper left corner. Histogram of data with the bin width chosen by the Freedman-Diaconis rule is shown in green. The rug plot indicates the location of data and the purple circle indicates the location of the observation 108.

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 the other

part, denoted by $\mathbf{S}_{\text{test}}$, is used to evaluate the performance of the corresponding gradient descent iterate. Let $\widehat{J}_{\text{SM,train}}$ and $\widehat{J}_{\text{SM,test}}$ be the SM loss functionals constructed using data in $\mathbf{S}_{\text{train}}$ and data in $\mathbf{S}_{\text{test}}$, 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 3.1.

Algorithm 3.1 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.

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

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

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 data in the remaining fold at this gradient descent iterate. Note that, for this particular t_0, 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 t_0. Repeat this process for each number of iterations candidate, and obtain an estimated risk from 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 3.2.

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 iterations 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.

Algorithm 3.2 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 of (roughly) equal size, denoted by $\mathbf{S}_1, \cdots, \mathbf{S}_K$;
Set `metric` to be an empty list to record the estimated risks for different number of iterations candidates;
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 data in $\mathbf{S}_{\text{train}}$, 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 data in $\mathbf{S}_{\text{test}}$;
    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 $\hat{f}^{(t_{m^*})}$.

3.3.1 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 3.5. *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 3.2. 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 all 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 3.5 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 $\text{range}(\widehat{C})$ and the other resides in $\text{range}(\widehat{C})^\perp$, the orthogonal complement of $\text{range}(\widehat{C})$.

As we have discussed in Remark 3.1, $\text{range}(\widehat{C})$ 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 $\Pi_S(\hat{z})$ denotes the orthogonal projection of $\hat{z}$ 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.

Proposition 3.1. *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 3.2. *Let* $\mathbf{G} \in \mathbb{R}^{nd\times nd}$ *and* $\mathbf{h} \in \mathbb{R}^{nd}$ *be the quantities defined in Theorem 3.4 and assume* $\mathbf{G}$ *is positive definite. 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 $\mathcal{X}$ for all $t \in \mathbb{N}_0$.

Proposition 3.3. *Under the same assumptions in Theorem 3.5, there exists* $M > 0$ *such that, for all* $x \in \mathcal{X}$ *and all* $t \in \mathbb{N}_0$, $|\langle \hat{f}_1^{(t+1)}, k(x, \cdot)\rangle_{\mathcal{H}}| \le M$.

The proofs of Propositions 3.1 - 3.3 can be found in Section 3.5.4.

Even though $\hat{f}_1^{(t+1)}$ is bounded over $\mathcal{X}$ for all $t \in \mathbb{N}_0$, $\hat{f}_2^{(t+1)}$ is *not*. In particular, since $\hat{f}_2^{(t+1)} = (t+1)\tau\hat{z}_2$, we have

$$|\langle \hat{f}_2^{(t+1)}, k(x, \cdot)\rangle_{\mathcal{H}}| = (t+1)\tau|\langle \hat{z}_2, k(x, \cdot)\rangle_{\mathcal{H}}| \to \infty, \qquad \text{as } t \to \infty,$$

unless $\langle \hat{z}_2, k(x, \cdot)\rangle_{\mathcal{H}} = 0$.

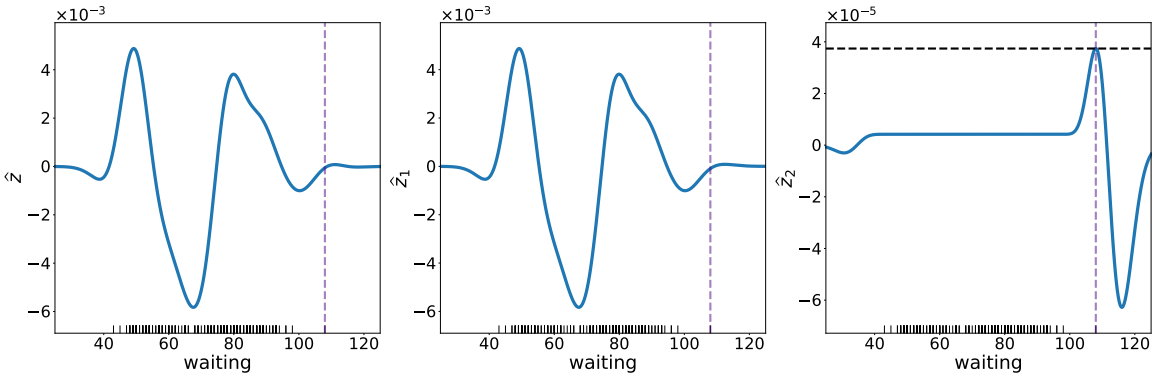

Figure 3.3: Plots of $\hat{z}$ (left panel), $\hat{z}_1$ (middle panel) and $\hat{z}_2$ (right panel). The rug plot indicates the location of data.

3.3.1.2 Numerical Illustration of Theorem 3.5

We use the `waiting` variable in the Old Faithful Geyser dataset to illustrate Theorem 3.5.

With the help of (2.15) in Chapter 2 and Proposition 3.2, we plot $\hat{z}$, $\hat{z}_1$ and $\hat{z}_2$ in Figure 3.3. In particular, note that, from the right panel of Figure 3.3, $\hat{z}_2$ achieves the maximum at 108. With Theorem 3.5, we expect see that, as the number of iterations t keeps increasing, the density value at 108 also keeps increasing. Figure 3.4, the plot of the density value at 108 against the number of iterations, coincides with our expectation.

3.3.2 Rate of Convergence

We study the rate of convergence of the early stopping SM density estimator in this section. More specifically, we attempt to answer the following questions:

1. Suppose $p_0 = q_{f_0} \in \mathcal{Q}_{\text{ker}}$ 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 $t \in \mathbb{N}$?

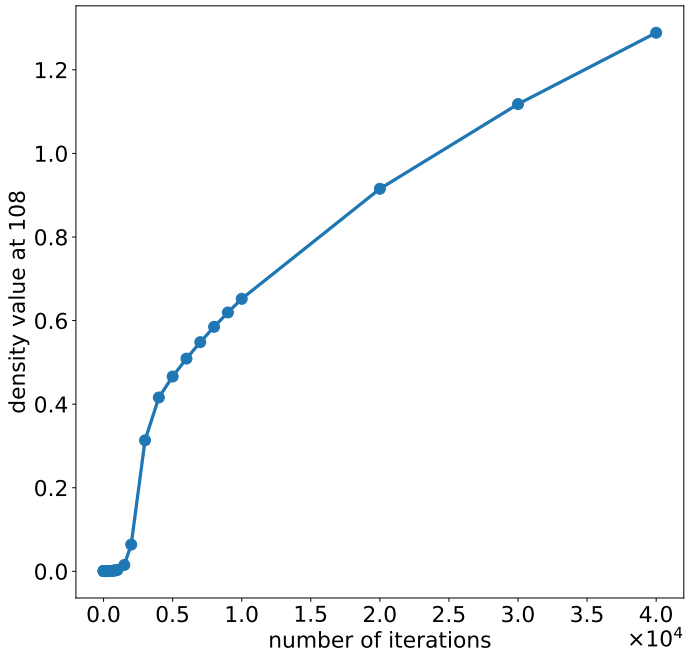

Figure 3.4: Density value at 108 against the number of iterations.

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 3.6. *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 $C_1 := 2d\tau(\kappa_3 + \kappa_4)(\tau d\kappa_2^2 + 1)\sqrt{2\log(2/\delta)}$ *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 iterates derived from the SM loss functional $\widehat{J}_{\text{SM}}$ (which depends on finitely many samples) and the other one is the gradient descent iterates derived from the H-divergence $J_{\text{SM}}(f)$ which can be viewed as the SM loss functional with infinitely many samples, i.e., the population-version SM loss functional. 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 intermediary.

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 an early 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 3.6.

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 derive an upper bound for the approximation error $\|f^{(t)} - f_0\|_{\mathcal{H}}$.

Proposition 3.4 (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 3.4 is almost identical to that of Theorem 3.2 and is omitted here.

With Proposition 3.4, 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 3.3, 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 3.7 (An upper bound on the approximation error)**.** *Choose $f^{(0)} = 0$ and the constant step size $\tau \in (0, 1/(d\kappa_2^2))$. Under the same assumptions in Theorem 3.6, 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 3.7 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 $f^{(t)} \to f_0$.

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 3.8 (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 3.6, 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 3.8 will be provided in Section 3.5.7. Note, in particular, that the gap between $\hat{f}^{(t)}$ and $f^{(t)}$ expands with t.

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

Theorem 3.6 establishes the stopping rule t^* and the convergence rate of $\hat{f}^{(t^*(n))}$. We can carry it over to bounding various distances between p_0 and $q_{\hat{f}^{(t^*(n))}}$. The main result is the following corollary.

Corollary 3.1 (Various distances between p_0 and $q_{\hat{f}^{(t^*(n))}}$)**.** *Under the same assumptions as in Theorem 3.6, 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,

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

(b) in terms of the KL-divergence,

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

(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* $p, q : \mathcal{X} \to [0, \infty)$*,*

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

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

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

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 3.1 will be provided in Section 3.5.8.

3.3.2.4 Discussion on (B7)

We end this section with 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, note 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 (Theorem A.1 in Appendix A), there exists an orthonormal basis for $\overline{\text{range}(C)}$, $\{\psi_\nu\}_{\nu=1}^{\infty}$, 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)** imply $\overline{\text{range}(C)} = \mathcal{H}$. Thus, $\{\psi_\nu\}_{\nu=1}^{\infty}$ 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 has been reviewed in Chapter 2. We omit the subscript "SM" in $\hat{f}_{\mathrm{SM}}^{(\rho)}$ in this section for notational simplicity.

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

From Theorem 3.3, 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 estimator 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 3.5, 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 3.9. *Let $\hat{z}_2 \in \mathcal{H}$ be the orthogonal projection of $\hat{z}$ onto* range$(\widehat{C})^{\perp}$. *Under* ***(A1)*** - ***(A4)*** *in Chapter 2,* ***(B1)*** - ***(B5)***, *and* ***(C1)*** *in Theorem 3.5, we have* $\lim_{\rho \to 0^+} q_{\hat{f}^{(\rho)}}(x^*) = \infty$.

The proof of Theorem 3.9 (given in Section 3.5.9) depends on the decomposition of $\hat{f}^{(\rho)}$ into two components, which we now state.

Proposition 3.5. *For each $\rho > 0$, 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 $\hat{f}_1^{(\rho)}$ is bounded over $\mathcal{X}$ for all $\rho > 0$ as below.

Proposition 3.6. *Under the same assumptions as in Theorem 3.9, there exists $M > 0$ such that for all $x \in \mathcal{X}$ and all $\rho > 0$, $\left|\langle \hat{f}_1^{(\rho)}, k(x, \cdot)\rangle_{\mathcal{H}}\right| \leq M$.*

However, $\hat{f}_2^{(\rho)}$ is *not* bounded over $\mathcal{X}$ for all $\rho > 0$ by noting that

$$|\langle \hat{f}_2^{(\rho)}, k(x, \cdot)\rangle_{\mathcal{H}}| = \rho^{-1} |\langle \hat{z}_2, k(x, \cdot)\rangle_{\mathcal{H}}| \to \infty, \qquad \text{as } \rho \to 0^+,$$

unless $\langle \hat{z}_2, k(x, \cdot)\rangle_{\mathcal{H}} = 0$.

The proofs of Propositions 3.5 and 3.6 will be given in Section 3.5.9 as well.

We now use the `waiting` variable to illustrate Theorem 3.9. As we have seen from Figure 3.3, $\hat{z}_2$ achieves the maximum at 108. With the result of Theorem 3.9, we expect to see that, as the value of the penalty parameter ρ keeps decreasing, the density value at 108 keeps increasing. Figure 3.5, the plot of the density value at 108 against $\log \rho$, coincides with our expectation.

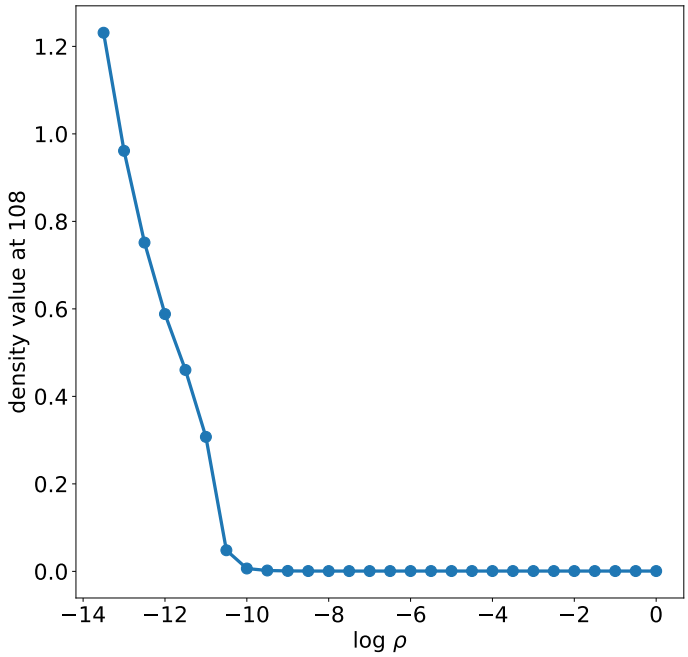

Figure 3.5: Density value at 108 against $\log \rho$.

3.4.3 Comparison through Eigen-decomposition

We show the similarities of the penalized and early stopping SM density estimators through their 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 $R < \infty$ denotes the rank of $\widehat{C}$.

By (2.16) in Chapter 2 and Proposition 3.5, we have

$$\begin{aligned} \hat{f}^{(\rho)} = (\widehat{C} + \rho I)^{-1}\hat{z} &= (\widehat{C} + \rho I)^{-1}\widehat{C}\hat{g}_1 + \frac{1}{\rho}\hat{z}_2 \\ &= \sum_{\nu=1}^{R} \frac{\hat{\xi}_\nu}{\hat{\xi}_\nu + \rho} \langle \hat{g}_1, \widehat{\psi}_\nu \rangle_{\mathcal{H}} \widehat{\psi}_\nu + \frac{1}{\rho}\hat{z}_2, \end{aligned} \tag{3.13}$$

where $\hat{g}_1 \in \mathcal{H}$ satisfying $\hat{z}_1 = \widehat{C}\hat{g}_1$. In addition, by Proposition 3.1, we have

$$\hat{f}^{(t)} = \tau \sum_{j=0}^{t-1} (I - \tau\widehat{C})^j \hat{z} = (I - (I - \tau\widehat{C})^t)\hat{g}_1 + t\tau\hat{z}_2$$
$$= \sum_{\nu=1}^{R} (1 - (1 - \tau\hat{\xi}_\nu)^t)\langle \hat{g}_1, \widehat{\psi}_\nu \rangle_{\mathcal{H}} \widehat{\psi}_\nu + t\tau\hat{z}_2. \tag{3.14}$$

Comparing (3.13) and (3.14), 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 density estimators. We stress a key difference of them by examining their convergence rates.

It has been shown in Sriperumbudur et al. (2017) that, under the same assumptions as in Theorem 3.6, 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 the value of γ is becoming large, 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 a larger value of γ implies that f_0 is smoother. 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 3.6 implies that with the stopping rule t^*, we have $\|\hat{f}^{(t^*(n))} - f_0\|_{\mathcal{H}} = \mathcal{O}_{p_0}(n^{-\frac{\gamma}{2(\gamma+2)}})$. 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` variable in the Old Faithful Geyser dataset.

The resulting penalized and early stopping SM density estimates with different choices of ρ and t are shown in Figure 3.6. 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.

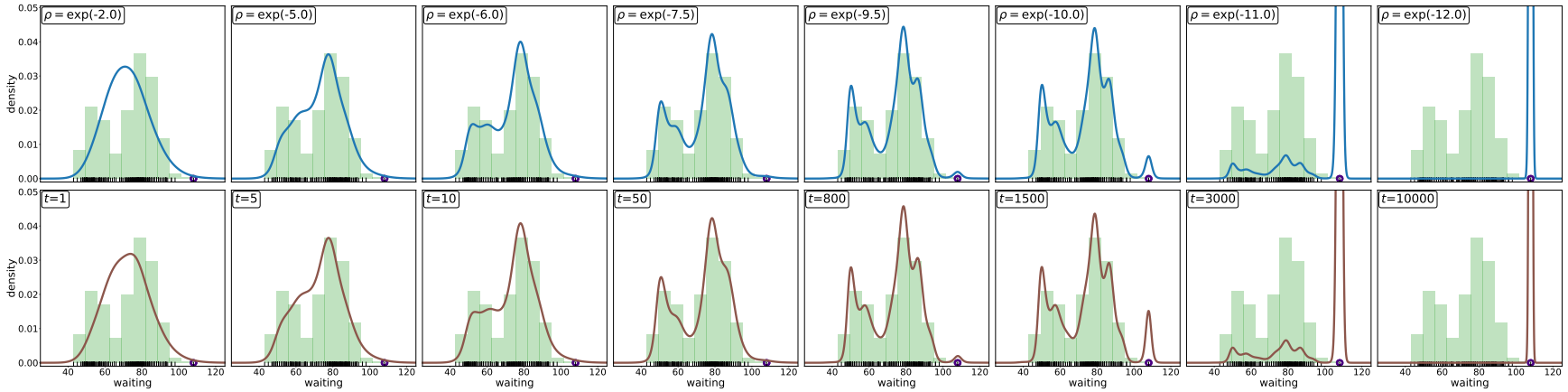

Figure 3.6: The penalized (first row) and early stopping (second row) SM density estimates with various choices of ρ and t, respectively, shown at the upper left corner. Histogram of data with the bin width chosen by the Freedman-Diaconis rule is shown in green. The rug plot indicates the location of data and the purple circle indicates the location of the observation 108.

3.5 Auxiliary Results and Proofs

3.5.1 Proof of Theorem 3.2

Proof of Theorem 3.2. 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, by the definition of the Fréchet gradient, we know $\nabla \widehat{J}_{\mathrm{SM}}(f) = \widehat{C}f - \hat{z}$ for all $f \in \mathcal{H}$. ∎

3.5.2 Proof of Theorem 3.3

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

$$\hat{f}^{(1)} = \hat{f}^{(0)} - \tau_0(\widehat{C}\hat{f}^{(0)} - \hat{z}) = (I - \tau_0\widehat{C})\hat{f}_0 + \tau_0\hat{z}. \tag{3.15}$$

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 (3.15).

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

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

which is the desired result.

The case of the constant step size is straightforward. ∎

3.5.3 Proof of Theorem 3.4

In order to prove Theorem 3.4, we need the following lemma, which is essentially a generalized version of the binomial theorem with elements in a Hilbert space.

Lemma 3.1. *For all $j \in \mathbb{N}_0$, we have*

$$(I - \tau\widehat{C})^j = \sum_{\ell=0}^{j} \binom{j}{\ell} (-\tau)^\ell \widehat{C}^\ell. \tag{3.16}$$

Proof of Lemma 3.1. We prove by induction. When $j = 0$, the left-hand side of (3.16) 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, (3.16) holds for $j = 0$.

Now, we assume (3.16) holds for $j = s$, and we show it also holds for $j = s + 1$. With $j = s + 1$, the left-hand side of (3.16) 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 3.2. *For all $\ell \in \mathbb{N}$, we have*

$$\widehat{C}^{\ell}\hat{z} = \frac{1}{n^{\ell}} \sum_{i=1}^{n} \sum_{u=1}^{d} \big[\mathbf{G}^{\ell-1}\mathbf{h}\big]_{(i-1)d+u} \partial_u k(X_i, \cdot), \tag{3.17}$$

where $\mathbf{G} \in \mathbb{R}^{nd\times nd}$ and $\mathbf{h} \in \mathbb{R}^{nd}$ are the same as those defined in Theorem 3.4.

Proof of Lemma 3.2. 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 (3.17) by induction. Note when $\ell = 1$, the left-hand side of (3.17) 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, (3.17) holds for $\ell = 1$.

Now, suppose that (3.17) 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 3.4.

Proof of Theorem 3.4. 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 3.1, 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 3.2, 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 $\{\partial_u k(X_i, \cdot) \text{ for } i = 1, \cdots, n \text{ and } u = 1, \cdots, d\}$, 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,

$$\begin{aligned}
\sum_{j=1}^{t} \sum_{\ell=1}^{j} \binom{j}{\ell} \left(\frac{(-\tau)^\ell}{n^\ell} \mathbf{G}^{\ell-1} \mathbf{h} \right) &= -\frac{\tau}{n} \sum_{j=1}^{t} \sum_{\ell=1}^{j} \binom{j}{\ell} \left(\left(-\frac{\tau}{n} \right)^{\ell-1} (\mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top)^{\ell-1} \mathbf{h} \right) \\
&= -\frac{\tau}{n} \sum_{j=1}^{t} \sum_{\ell=1}^{j} \binom{j}{\ell} \left(\left(-\frac{\tau}{n} \right)^{\ell-1} \mathbf{Q} \mathbf{\Lambda}^{\ell-1} \mathbf{Q}^\top \mathbf{h} \right) \\
&= -\frac{\tau}{n} \mathbf{Q} \left[\sum_{j=1}^{t} \sum_{\ell=1}^{j} \binom{j}{\ell} \left(-\frac{\tau}{n} \mathbf{\Lambda} \right)^{\ell-1} \right] \mathbf{Q}^\top \mathbf{h}. \qquad (3.18)
\end{aligned}$$

Now, since $\mathbf{\Lambda}$ is a diagonal matrix, the sum and the power in (3.18) 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 3.1. Since we can write $\hat{z} = \hat{z}_1 + \hat{z}_2$ with $\hat{z}_1 = \widehat{C}\hat{g}_1 \in \text{range}(\widehat{C})$ and $\hat{z}_2 \in \text{range}(\widehat{C})^\perp$, we can write the $(t+1)$-st 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

$$\hat{f}^{(t+1)} - \hat{g}_1 = (I - \tau\widehat{C})(\hat{f}^{(t)} - \hat{g}_1) + \tau\hat{z}_2. \tag{3.19}$$

Projecting both sides of (3.19) onto $\text{range}(\widehat{C})$, we obtain

$$\begin{aligned}
\Pi_{\text{range}(\widehat{C})}(\hat{f}^{(t+1)} - \hat{g}_1) &= \Pi_{\text{range}(\widehat{C})}\big((I - \tau\widehat{C})(\hat{f}^{(t)} - \hat{g}_1) + \tau\hat{z}_2\big) \\
&= (I - \tau\widehat{C})\Pi_{\text{range}(\widehat{C})}(\hat{f}^{(t)} - \hat{g}_1) \\
&= (I - \tau\widehat{C})^{t+1}\Pi_{\text{range}(\widehat{C})}(\hat{f}^{(0)} - \hat{g}_1) \\
&= (I - \tau\widehat{C})^{t+1}\Pi_{\text{range}(\widehat{C})}(-\hat{g}_1) \\
&= -(I - \tau\widehat{C})^{t+1}\hat{g}_1,
\end{aligned}$$

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

Now, project both sides of (3.19) onto $\text{range}(\widehat{C})^\perp$, we obtain

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

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 3.2. (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 definite by assumption, 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 exist and 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

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

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 $\frac{1}{n}\mathbf{G}\boldsymbol{\beta}^* = \boldsymbol{\alpha}^*$. ∎

Proof of Proposition 3.3. Recall from Proposition 3.1 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, (iii) follows from **(A2)**. ∎

We are now ready to prove Theorem 3.5.

Proof of Theorem 3.5. Let $\hat{z}_2(x) := \langle \hat{z}_2, k(x, \cdot)\rangle_{\mathcal{H}}$ for all $x \in \mathcal{X}$. Notice that

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

Using Proposition 3.3, we can bound $q_{\hat{f}^{(t)}}(x^*)$ from below by

$$q_{\hat{f}^{(t)}}(x^*) \geq \frac{\mu(x^*) \exp\big(-M + t\tau\hat{z}_2(x^*)\big)}{\int_{\mathcal{X}} \mu(x) \exp\big(M + t\tau\hat{z}_2(x)\big) \mathrm{d}x} = \frac{\exp(-2M)\mu(x^*)}{\int_{\mathcal{X}} \mu(x) \frac{\exp(t\tau\hat{z}_2(x))}{\exp(t\tau\hat{z}_2(x^*))} \mathrm{d}x}, \tag{3.20}$$

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 $x = x^*$, by **(C1)**. 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

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

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

Taking the limit of $t \to \infty$ of both sides of (3.20), 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 3.6

Proof of Theorem 3.6. Based on the inequality (3.9) and Theorems 3.7 and 3.8, we know that with probability at least $1 - \delta$ for $\delta \in (0, 1)$, the following inequality holds

$$\|\hat{f}^{(t)} - f_0\|_{\mathcal{H}} \leq C_1 \frac{t^2}{\sqrt{n}} + C_2 t^{-\gamma}, \tag{3.21}$$

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 (3.21), 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 3.7

In order to prove Theorem 3.7, 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 3.3 (Sriperumbudur et al., 2017, Theorem 4(ii))**.** *Under* ***(A1)*** *-* ***(A4)*** *in Chapter 2 and* ***(B1)*** *-* ***(B7)****, we have* $Cf_0 = z$.

Lemma 3.4. *Assume* $f^{(0)} = 0$. *For all* $t \in \mathbb{N}_0$, $f^{(t)} - f_0 = -(I - \tau C)^t f_0$.

Proof of Lemma 3.4. 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 3.3, 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

$$f^{(t)} - f_0 = f^{(t-1)} - f_0 - \tau C(f^{(t-1)} - f_0) = (I - \tau C)(f^{(t-1)} - f_0)$$

$$= (I - \tau C)^t (f^{(0)} - f_0).$$

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

We now prove Theorem 3.7.

Proof of Theorem 3.7. By Lemma 3.4, 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 $\{\psi_\nu\}_{\nu=1}^\infty$ are the eigenvectors of C that form an orthonormal basis of $\overline{\text{range}(C)}$ and $\{\xi_\nu\}_{\nu=1}^\infty$ 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

$$\begin{aligned} \|(I - \tau C)^t C^\gamma\| &\leq \sup_\nu \left\{ (1 - \tau\xi_\nu)^t \xi_\nu^\gamma \right\} = \sup_\nu \exp\left(t \log(1 - \tau\xi_\nu) + \gamma \log \xi_\nu \right) \\ &\leq \sup_\nu \exp\left(-t\tau\xi_\nu + \gamma \log \xi_\nu \right), \end{aligned} \tag{3.22}$$

where the last inequality follows from the basic inequality $\log(1 + x) \leq x$ for all $x > -1$. Also, note that $\log(1 - \tau\xi_\nu)$ is well-defined since $\tau < 1/(d\kappa_2^2) \leq 1/\|C\|$ 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 (3.22), we have

$$\|(I - \tau C)^t C^\gamma\| \leq \left(\frac{\gamma}{\tau e t} \right)^\gamma, \tag{3.23}$$

and the desired result follows. ∎

3.5.7 Proof of Theorem 3.8

In order to prove Theorem 3.8, we first need the following proposition that gives an equivalent expression of $\hat{f}^{(t)}$ that links operators $\widehat{C}$ and C.

Proposition 3.7. *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 3.7. 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 3.3, 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 probabilistic upper bounds for $\|\widehat{C} - C\|$ and $\|\hat{z} - z\|_{\mathcal{H}}$.

Lemma 3.5 (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 $i = 1, \cdots, n$, $\|W_i\|_{\mathcal{H}} \leq c_i < \infty$ almost surely. Then, for any $\varepsilon > 0$,*

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

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 3.5 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 3.6. *Under* ***(B1)*** *-* ***(B5)****, the operator $C : \mathcal{H} \to \mathcal{H}$ defined in (3.1) is a Hilbert-Schmidt operator with $\|C\|_{\mathrm{HS}} \leq d\kappa_2^2$, where $\|\cdot\|_{\mathrm{HS}}$ 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 A.4 in Appendix A.

Proof of Lemma 3.6. 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

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

where (i) follows from **(B5)**. By the relationship between the Hilbert-Schmidt norm and the trace (see Proposition A.10(c) in Appendix A), we conclude that $\|C\|_{\mathrm{HS}} \leq \mathrm{trace}(C) \leq d\kappa_2^2$.

As for the result on z, using Proposition A.4(b) in Appendix A, 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 3.7. *Under* ***(B1)*** *-* ***(B5)****, for all $i = 1, \cdots, n$, the operator $\widehat{C}_i : \mathcal{H} \to \mathcal{H}$ defined by*

$$\widehat{C}_i := \sum_{u=1}^{d} \partial_u k(X_i,\cdot) \otimes \partial_u k(X_i,\cdot) \tag{3.24}$$

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

$$\hat{z}_i := -\sum_{u=1}^{d} \big(\partial_u \log \mu(X_i)\partial_u k(X_i, \cdot) + \partial_u^2 k(X_i, \cdot)\big) \in \mathcal{H} \tag{3.25}$$

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

Proof of Lemma 3.7. 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

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

and hence, $\|\widehat{C}_i\|_{\mathrm{HS}} \le d\kappa_2^2$. 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 of the partial derivatives of k in (iv). We use **(B5)** in (v).

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

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

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

We now use Lemmas 3.5 - 3.7 to derive upper bounds of $\|\widehat{C} - C\|$ in Proposition 3.8 and $\|\hat{z} - z\|_{\mathcal{H}}$ in Proposition 3.9.

Proposition 3.8. *Under the same assumptions in Theorem 3.8, 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 (3.24). 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 A.10(a) in Appendix A,

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

Then, the desired result follows from Lemma A.10(b) in Appendix A and Lemma 3.5. ∎

Proposition 3.9. *Under the same assumptions in Theorem 3.8, 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 (3.25), we have $\hat{z} = \frac{1}{n}\sum_{i=1}^n z_i$ and $\mathbb{E}[z_i] = \mathbb{E}[\hat{z}] = z$. Define $W_i := \hat{z}_i - z$ for all $i = 1, \cdots, n$. It follows that $\{W_i\}_{i=1}^n$ is a 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 3.5. ∎

In order to prove Theorem 3.8, 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 3.10. *Let $\hat{f}^{(0)} = 0$. Under the same assumptions in Theorem 3.8, 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 3.3 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 3.8, and Lemma 3.7, 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 3.8.

Proof of Theorem 3.8. 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

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

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

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

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

Using Propositions 3.8 and 3.9, we have, with probability at least $1 - \delta$ for $\delta \in (0, 1)$,

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

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

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

■

3.5.8 Proof of Corollary 3.1

In order to prove Corollary 3.1, we need the following lemma (Lemma A.1 in Sriperumbudur et al., 2017).

Lemma 3.8. *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

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

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

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

(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 3.1. (a) By Theorem 4(i) in Sriperumbudur et al. (2017), we know that $\mathrm{H}(p_0\|q_f) = \frac{1}{2}\langle f - f_0, C(f-f_0)\rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$. In addition, by Lemma 3.6 and Lemma A.10 in Appendix A, we know $\|C\| \le \|C\|_{\mathrm{HS}} \le d\kappa_2^2$. Thus, we have

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

(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 3.8(a), there exists a universal constant $c > 0$ such that

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

Since

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

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

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

(c) By Lemma 3.8(b), we have

$$\begin{aligned}\mathrm{He}(p_0\|q_{\hat{f}^{(t^*(n))}}) &\le \exp\left(\frac{1}{2}\kappa_1(4C_1+C_2)n^{-\frac{\gamma}{2(\gamma+2)}}\right)\kappa_1(4C_1+C_2)n^{-\frac{\gamma}{2(\gamma+2)}} \\ &\le \underbrace{\exp\left(\frac{1}{2}\kappa_1(4C_1+C_2)\right)\kappa_1(4C_1+C_2)}_{=:C_5} n^{-\frac{\gamma}{2(\gamma+2)}}.\end{aligned}$$

(d) By Lemma 3.8(c), we have

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

■

3.5.9 Proof of Results in Section 3.4

We first prove the following lemma that will help the proof of Proposition 3.5.

Lemma 3.9. *The following identities hold*

$$\begin{aligned}\Pi_{\text{range}(\widehat{C})}(\widehat{C} + \rho I)^{-1} &= (\widehat{C} + \rho I)^{-1}\Pi_{\text{range}(\widehat{C})}, \\ \Pi_{\text{range}(\widehat{C})^\perp}(\widehat{C} + \rho I)^{-1} &= \rho^{-1}\Pi_{\text{range}(\widehat{C})^\perp}.\end{aligned}$$

Proof of Lemma 3.9. In order to show the desired results, 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,

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

■

Proof of Proposition 3.5. 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 Lemma 3.9 in (i) and the definitions of $\hat{z}_1$ and $\hat{z}_2$ in (ii). ∎

Proof of Proposition 3.6. First, note by Proposition 3.5 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 $\text{range}(\widehat{C})$, and $R < \infty$ 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,

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

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 3.9. Let $\hat{z}_2(x) := \langle \hat{z}_2, k(x, \cdot)\rangle_{\mathcal{H}}$ for all $x \in \mathcal{X}$. By Proposition 3.5, we know, for any $x \in \mathcal{X}$,

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

and, hence,

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

By Proposition 3.6, we can bound $q_{\hat{f}^{(\rho)}}(x^*)$ from below by

$$\begin{aligned} q_{\hat{f}^{(\rho)}}(x^*) &\geq \frac{\mu(x^*)\exp(-M + \rho^{-1}\hat{z}_2(x^*))}{\int_{\mathcal{X}} \mu(x)\exp(M + \rho^{-1}\hat{z}_2(x))\,\mathrm{d}x} \\ &= \frac{\mu(x^*)\exp(-2M)}{\int_{\mathcal{X}} \mu(x)\frac{\exp(\rho^{-1}\hat{z}_2(x))}{\exp(\rho^{-1}\hat{z}_2(x^*))}\,\mathrm{d}x} \end{aligned} \tag{3.26}$$

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

$$R_\rho(x) := \frac{\exp(\rho^{-1}\hat{z}_2(x))}{\exp(\rho^{-1}\hat{z}_2(x^*))} = \left(\frac{\exp(\hat{z}_2(x))}{\exp(\hat{z}_2(x^*))}\right)^{\frac{1}{\rho}},$$

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 (3.26), 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. ■

Chapter 4: Comparison of Regularized ML and SM Density Estimators in $\mathcal{Q}_{\text{ker}}$

After discussing the early stopping SM density estimator and comparing it with the penalized SM density estimator in the preceding chapter, we now turn to comparing these two kinds of regularized SM density estimators with the penalized ML density estimator. In particular, we would like to understand what differences the regularized SM density estimators and the penalized ML density estimator have and why they have such differences.

In Section 4.1, we focus on the penalized ML density estimator. We will establish the existence and the uniqueness of the minimizer of the penalized NLL loss functional. In spite of its existence, similar to the discussion in Gu and Qiu (1993), such a minimizer in an infinite-dimensional RKHS is *not* computable. Instead, we seek a finite-dimensional subspace of $\mathcal{H}$ to approximate this minimizer. In order to ensure the comparability of SM and ML density estimators, we minimize the (penalized) SM loss functional over this finite-dimensional subspace as well, and discuss how to compute the penalized and early stopping SM density estimators in it in Section 4.2. We then discuss the similarities and differences between regularized SM density estimators and penalized ML density estimator via numerical examples in Section 4.3. Finally, we explain why they have the observed differences in Section 4.4.

4.1 Penalized ML Density Estimator

In this section, we look at the penalized ML density estimator, which is obtained by minimizing

$$\widehat{J}_{\mathrm{NLL}}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2, \qquad \text{subject to } f \in \mathcal{F}, \tag{4.1}$$

where $\widehat{J}_{\mathrm{NLL}}(f) = A(f) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)$ is the NLL loss functional we have seen in Chapter 2, and $\lambda \geq 0$ is the penalty parameter. Note that in (4.1), we choose the penalty functional to be $\widetilde{P}(f) := \frac{1}{2}\|f\|_{\mathcal{H}}^2$, for all $f \in \mathcal{H}$, which is to ensure the comparability with the regularized SM density estimators.

The following proposition established the existence and the uniqueness of the minimizer of (4.1).

Proposition 4.1. *Under Assumptions **(A1)** - **(A4)** in Chapter 2, for every $\lambda > 0$, (4.1) has a unique minimizer in $\mathcal{F}$.*

The proof of Proposition 4.1 is given in Section 4.5.1.

4.1.1 Failure of the Representer Theorem

Note that Proposition 4.1 only shows (4.1) has a unique minimizer, but gives no indication what such a minimizer is like.

Since minimizing (4.1) is a convex minimization problem over an infinite-dimensional RKHS, a classic tool of characterizing its minimizer is the representer theorem (Kimeldorf and Wahba, 1971; Schölkopf, Herbrich, and Smola, 2001), which states that, if $\widetilde{J} : \mathcal{H} \to \mathbb{R}$ is a convex loss functional that depends *only* on the evaluation of $f \in \mathcal{H}$ at data points $X_1, \cdots, X_n$, then

$$\min_{f\in\mathcal{H}}\left\{\widetilde{J}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2\right\} = \min_{f\in\mathrm{Span}\{k(X_1,\cdot),\cdots,k(X_n,\cdot)\}}\left\{\widetilde{J}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2\right\};$$

in other words, the minimizer of $\widetilde{J}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$ over $\mathcal{H}$ can be represented as a linear combination of $k(X_1, \cdot), \cdots, k(X_n, \cdot)$. Here, we only consider a specific form of the penalty functional $\frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$. The representer theorem covers the case where a general penalty functional is used; see Schölkopf, Herbrich, and Smola (2001) for a discussion.

The representer theorem is a powerful tool and reduces a convex minimization problem in an infinite-dimensional RKHS to the one in a finite-dimensional subspace, which effectively reduces the computational burden and improves the efficiency. However, note that, due the presence of A in (4.1), $\widehat{J}_{\mathrm{NLL}}$ does not only depend on the evaluation of f at data points $X_1, \cdots, X_n$, but also on that of f at *all* points in $\mathcal{X}$. This suggests the representer theorem may fail in characterizing the minimizer of (4.1). In fact, as the following example demonstrates, the representer theorem does fail.

Example 4.1. Let $\mathcal{X} = \mathbb{R}^2$ and the kernel function be $k(x, y) = (x^\top y)^2$ for all $x, y \in \mathbb{R}^2$, i.e., the homogeneous polynomial kernel of degree 2. We choose the base density to be $\mu(x) = \frac{1}{2\pi}\exp(-\frac{\|x\|_2^2}{2})$ for all $x \in \mathbb{R}^2$, i.e., the pdf of 2-dimensional Gaussian distribution with the zero mean and the identity covariance matrix. We fix $n = 1$ and let the single data point be $X_1 = (a, 0)^\top \in \mathbb{R}^2$. Also, let $\lambda > 0$.

Then, it can be shown that the minimizer of the corresponding penalized NLL loss functional

$$A(f) - f(X_1) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2, \qquad \text{subject to } f \in \mathcal{F}, \tag{4.2}$$

takes on the form of

$$f^*(x) = b^* x_1^2 + c^* x_2^2, \qquad \text{for all } x = (x_1, x_2)^\top \in \mathbb{R}^2,$$

for some $b^* \neq 0$ and $c^* \neq 0$, but $\mathrm{Span}\{k(X_1, \cdot)\} \cap \mathcal{F}$ contains all functions of the form

$$f(x) = \gamma a^2 x_1^2, \quad \text{for all } x := (x_1, x_2)^\top \in \mathbb{R}^2 \text{ and } \gamma \in \mathbb{R} \text{ satisfying } \gamma a^2 < \frac{1}{2}.$$

Thus, $f^* \notin \text{Span}\{k(X_1, \cdot)\} \cap \mathcal{F}$ and we conclude the representer theorem fails.

Details about this example can be found in Section 4.5.2. ▶

4.1.2 Construction of a Finite-dimensional Approximating Space

Now, we have seen the representer theorem fails in characterizing the minimizer of (4.1). In addition, as is discussed by Gu and Qiu (1993), despite its existence and the uniqueness, such a minimizer in an infinite-dimensional $\mathcal{H}$ is *not* computable. In order to compute the penalized ML density estimator, we propose to seek a finite-dimensional subspace in $\mathcal{H}$ to approximate the minimizer of (4.1), which is the focus of the current section.

One idea is to choose the following n-dimensional subspace spanned by the kernel functions centered at $X_1, \cdots, X_n$,

$$\left\{ f \;\middle|\; f := \sum_{i=1}^{n} \beta_i k(X_i, \cdot), \beta_1, \cdots, \beta_n \in \mathbb{R} \right\}, \tag{4.3}$$

as the approximating space. This is the approach taken by Gu and Qiu (1993) and Gu (1993).

We propose a different approach here. Suppose $\mathcal{X} \subseteq \mathbb{R}$ for the moment and start with a set of grid points over $\mathcal{X}$, denoted by $\mathbf{X}_1 := \{w_1, \cdots, w_m\}$, that covers the range of data. We minimize the objective functional in (4.1) over

$$\widetilde{\mathcal{H}}_1 := \left\{ f \;\middle|\; f := \sum_{j=1}^{m} \beta_j k(w_j, \cdot), \beta_j \in \mathbb{R}, w_j \in \mathbf{X}_1, \text{ for all } j = 1, \cdots, m \right\}.$$

Then, we insert an additional grid point at the midpoint of two adjacent points in $\mathbf{X}_1$, forming

$$\mathbf{X}_2 := \left\{ w_1, w_{1.5}, w_2, \cdots, w_{m-1}, w_{\frac{2m-1}{2}}, w_m \right\},$$

where $w_{\frac{2j-1}{2}} = \frac{1}{2}(w_j + w_{j+1})$ for all $j = 1, \cdots, m-1$, and minimize (4.1) over

$$\widetilde{\mathcal{H}}_2 := \left\{ f \;\middle|\; f := \sum_{j \in \{1, 1.5, 2, \cdots, m\}} \beta_j k(w_j, \cdot), \beta_j \in \mathbb{R}, w_j \in \mathbf{X}_2, \text{ for all } j = 1, 1.5, 2, \cdots, m \right\}.$$

Let $\tilde{f}_{\mathrm{ML},\ell} := \arg\min_{f \in \widetilde{\mathcal{H}}_\ell} \left\{ \widehat{J}_{\mathrm{NLL}}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \right\}$ for $\ell = 1, 2$. Then, we compute the following relative difference

$$\left| \frac{\left[\widehat{J}_{\mathrm{NLL}}(\tilde{f}_{\mathrm{ML},2}) + \frac{\lambda}{2}\|\tilde{f}_{\mathrm{ML},2}\|_{\mathcal{H}}^2 \right] - \left[\widehat{J}_{\mathrm{NLL}}(\tilde{f}_{\mathrm{ML},1}) + \frac{\lambda}{2}\|\tilde{f}_{\mathrm{ML},1}\|_{\mathcal{H}}^2 \right]}{\widehat{J}_{\mathrm{NLL}}(\tilde{f}_{\mathrm{ML},1}) + \frac{\lambda}{2}\|\tilde{f}_{\mathrm{ML},1}\|_{\mathcal{H}}^2} \right|. \tag{4.4}$$

Since $\mathbf{X}_1 \subset \mathbf{X}_2$ and $\widetilde{\mathcal{H}}_1 \subset \widetilde{\mathcal{H}}_2$, we must have $\widehat{J}_{\mathrm{NLL}}(\tilde{f}_{\mathrm{ML},1}) + \frac{\lambda}{2}\|\tilde{f}_{\mathrm{ML},1}\|_{\mathcal{H}}^2 \geq \widehat{J}_{\mathrm{NLL}}(\tilde{f}_{\mathrm{ML},2}) + \frac{\lambda}{2}\|\tilde{f}_{\mathrm{ML},2}\|_{\mathcal{H}}^2$.

We repeat the process described above: keep inserting additional grid points at the midpoint of two adjacent points, minimize the penalized NLL loss functional over the new finite-dimensional subspace with the updated grid points, and stop until the relative difference of the values of the penalized NLL loss functional between two consecutive sets of grid points does not exceed a pre-specified tolerance parameter. The algorithm is provided in Algorithm 4.1.

If $\mathcal{X} \subseteq \mathbb{R}^d$ for $d > 1$, we can generalize the procedure described above accordingly. However, one has to be careful when adding additional points, since, with $d > 1$, each grid point has up to 2^d adjacent points, and the number of grid points being added each time can grow exponentially. The computational burden will also increase accordingly as d increases. This is one example of the infamous curse of dimensionality and is an unsatisfactory feature of our approach.

One advantage of our approach, comparing to the approach by Gu and Qiu (1993) and Gu (1993), is that our approach can yield an even smaller minimum of the penalized NLL loss functional than theirs, as we will see via numerical examples in Section 4.1.4.

Algorithm 4.1 Determining the finite-dimensional subspace of approximating the minimizer of $\widehat{J}_{\mathrm{NLL}}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$ over $f \in \mathcal{H}$

Require:

- $X_1, \cdots, X_n$, the data from p_0;
- $\mathbf{X}_1 := \{w_1, \cdots, w_m\}$, the first set of grid points;
- $\epsilon > 0$, the tolerance parameter.

Let $\widetilde{\mathcal{H}}_1 = \left\{f \mid f := \sum_j \beta_j k(w_j, \cdot), \beta_j \in \mathbb{R}, w_j \in \mathbf{X}_1, \text{ for all } j = 1, 2, \cdots, m\right\}$ and compute $\tilde{f}_{\mathrm{ML},1} := \arg\min_{f \in \widetilde{\mathcal{H}}_1} \left\{\widehat{J}_{\mathrm{NLL}}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2\right\}$;
Form $\mathbf{X}_2$, which is obtained by adding a grid point at the midpoint of two adjacent points in $\mathbf{X}_1$;
Compute $\tilde{f}_{\mathrm{ML},2} := \arg\min_{f \in \widetilde{\mathcal{H}}_2} \left\{\widehat{J}_{\mathrm{NLL}}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2\right\}$, where $\widetilde{\mathcal{H}}_2 := \{f \mid f := \sum_j \beta_j k(w_j, \cdot), \beta_j \in \mathbb{R}, w_j \in \mathbf{X}_2\}$;
Compute `error` using (4.4);
while `error` $> \epsilon$ **do**
    Let $\mathbf{X}_1 = \mathbf{X}_2$, and form a new $\mathbf{X}_2$ by adding a grid point at the midpoint of two adjacent points in $\mathbf{X}_1$;
    Let $\tilde{f}_{\mathrm{ML},1} = \tilde{f}_{\mathrm{ML},2}$, and compute $\tilde{f}_{\mathrm{ML},2} := \arg\min_{f \in \widetilde{\mathcal{H}}_2} \left\{\widehat{J}_{\mathrm{NLL}}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2\right\}$, where $\widetilde{\mathcal{H}}_2 := \{f \mid f := \sum_j \beta_j k(w_j, \cdot), \beta_j \in \mathbb{R}, w_j \in \mathbf{X}_2\}$;
    Update `error` using (4.4).
return $\mathbf{X}_1$.

In our description of the process above, one key component we have not discussed is how to compute the minimizer of the penalized NLL loss functional for a given finite-dimensional approximating subspace. As we have discussed in Chapter 2, the main difficulty is to handle A and its derivatives efficiently. In the next section, we propose an algorithm to approximate the values of A and its derivatives and compute the minimizer of the penalized NLL loss functional.

4.1.3 Computation of the Minimizer of the Penalized NLL Loss Functional

Our goal of this section is to provide an adaptive algorithm to compute the minimizer of $\widehat{J}_{\mathrm{NLL}}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$ over a given finite-dimensional approximating subspace

$$\widetilde{\mathcal{H}} := \left\{ f \;\middle|\; f := \sum_{j=1}^{m} \beta_j k(w_j, \cdot), \beta_j \in \mathbb{R} \text{ for all } j = 1, 2, \cdots, m \right\},$$

where $\{w_1, \cdots, w_m\} \subset \mathcal{X}$ is a set of specified grid points.

Since any $f \in \widetilde{\mathcal{H}}$ can be written as $f = \sum_{j=1}^{m} \beta_j k(w_j, \cdot)$, we can rewrite the penalized NLL loss functional as

$$\widetilde{J}_{\mathrm{NLL},\lambda}(\boldsymbol{\beta}) := \widehat{J}_{\mathrm{NLL}}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 = \widetilde{A}(\boldsymbol{\beta}) - \boldsymbol{\beta}^\top \left(\frac{1}{n} \mathbf{K}_1 \mathbf{1}_n \right) + \frac{\lambda}{2} \boldsymbol{\beta}^\top \mathbf{K}_2 \boldsymbol{\beta}, \tag{4.5}$$

for all $\boldsymbol{\beta} := (\beta_1, \cdots, \beta_m)^\top \in \mathbb{R}^m$, where

$$\widetilde{A}(\boldsymbol{\beta}) := A(f) = \log \left(\int_{\mathcal{X}} \mu(x) \exp \left(\sum_{j=1}^{m} \beta_j k(w_j, x) \right) \mathrm{d}x \right),$$

the (j, i)-entry of $\mathbf{K}_1 \in \mathbb{R}^{m \times n}$ is $k(w_j, X_i)$, the (j, j')-entry of $\mathbf{K}_2 \in \mathbb{R}^{m \times m}$ is $k(w_j, w_{j'})$, for all $j, j' = 1, \cdots, m$, $i = 1, \cdots, n$, and $\mathbf{1}_n := (1, \cdots, 1)^\top \in \mathbb{R}^n$.

By a similar argument to Proposition 4.1, we can show $\widetilde{J}_{\mathrm{NLL},\lambda}$ has a unique minimizer in $\mathbb{R}^m$. To compute this minimizer, we use the gradient descent algorithm with the constant step size. Starting from $\boldsymbol{\beta}_{\mathrm{NLL}}^{(0)} \in \mathbb{R}^m$, the gradient descent iterates are

$$\boldsymbol{\beta}_{\mathrm{NLL}}^{(t+1)} = \boldsymbol{\beta}_{\mathrm{NLL}}^{(t)} - \tau \nabla \widetilde{J}_{\mathrm{NLL},\lambda}(\boldsymbol{\beta}_{\mathrm{NLL}}^{(t)}), \qquad \text{for all } t \in \mathbb{N}_0,$$

where $\tau > 0$ is an appropriately chosen constant step size, and

$$\nabla \widetilde{J}_{\mathrm{NLL},\lambda}(\boldsymbol{\beta}) = \nabla \widetilde{A}(\boldsymbol{\beta}) - \frac{1}{n}\mathbf{K}_1\mathbf{1}_n + \lambda \mathbf{K}_2\boldsymbol{\beta}. \tag{4.6}$$

We terminate the algorithm when $\|\nabla \widetilde{J}_{\mathrm{NLL},\lambda}(\boldsymbol{\beta}^{(t)}_{\mathrm{NLL}})\|_2/\sqrt{m}$ is less than a pre-specified tolerance parameter.

The j-th component of $\nabla \widetilde{A}(\boldsymbol{\beta})$ in (4.6) is

$$\int_{\mathcal{X}} k(w_j, x)\mu(x) \exp\left(\sum_{\ell=1}^{m} \beta_\ell k(w_\ell, x) - \widetilde{A}(\boldsymbol{\beta})\right) \mathrm{d}x, \tag{4.7}$$

which is $\mathbb{E}_{q_f}[k(w_j, X)]$ with $f \in \widetilde{\mathcal{H}}$, for all $j = 1, \cdots, m$. Exact computation of (4.7) can be difficult. We propose the batch Monte Carlo method in the next section to approximate them.

4.1.3.1 Batch Monte Carlo Approximation of $\nabla \widetilde{A}(\boldsymbol{\beta})$

Since μ is a density function, let $Y_1, \cdots, Y_B$ be i.i.d samples from μ. We can approximate $\exp(\widetilde{A}(\boldsymbol{\beta}))$ and (4.7) using the Monte Carlo method by

$$\widehat{\exp(\widetilde{A}(\boldsymbol{\beta}))} := \frac{1}{B}\sum_{b=1}^{B} \exp\left(\sum_{\ell=1}^{m} \beta_\ell k(w_\ell, Y_b)\right) \tag{4.8}$$

and

$$\frac{1}{B}\sum_{b=1}^{B} k(w_j, Y_b) \exp\left(\sum_{\ell=1}^{m} \beta_\ell k(w_\ell, Y_b) - \log\left(\widehat{\exp(\widetilde{A}(\boldsymbol{\beta}))}\right)\right), \tag{4.9}$$

respectively.

What we have so far is just the *vanilla* Monte Carlo method. It requires us to choose an appropriate value of B so that we can obtain satisfactory approximations of the desired quantities. This may not be a trivial task in practice.

In order to better choose the value of B under different scenarios and control the quality of the approximations, we propose the batch Monte Carlo method. The

main idea is to keep drawing a batch of random samples from μ and updating the approximations until the standard deviation of approximations is less than a pre-specified tolerance parameter (see Steps 4, 10, 15 and 21 in Algorithm 4.2 for the calculation of the stopping criterion). The details are given in Algorithm 4.2.

Algorithm 4.2 Batch Monte Carlo method

Require:

- $w_1, \cdots, w_m$, grid points at which kernel functions are centered;
- a sampler to draw i.i.d samples from μ;
- $\boldsymbol{\beta}$, the coefficient vector of kernel functions;
- B, batch size;
- $\epsilon > 0$, tolerance parameter to determine when to stop sampling.

`/* Approximation of` $\exp(\widetilde{A}(\boldsymbol{\beta}))$ `*/`

1: Draw $Y_1, \cdots, Y_B$ from μ and set `batch_cnt` $= 1$; ▷ *Draw the first batch*

2: Approximate $\exp(\widetilde{A}(\boldsymbol{\beta}))$ using $Y_1, \cdots, Y_B$ according to (4.8); let the resulting approximation be `output1`;

3: Set `mean_sq` $= \frac{1}{B}\sum_{b=1}^{B} \exp\big(2\sum_{\ell=1}^{m} \beta_\ell k(w_\ell, Y_b)\big)$;

4: Set `se1` $= \sqrt{\texttt{mean\_sq} - \texttt{output1}^2}$;

5: **while** `se1` $> \epsilon$ **do**

6: Draw $Y_1, \cdots, Y_B$ from μ; ▷ *Draw more batches*

7: Approximate $\exp(\widetilde{A}(\boldsymbol{\beta}))$ using $Y_1, \cdots, Y_B$ according to (4.8); let the resulting approximation be `output1_inter`;

8: Update

$$\texttt{output1} = (\texttt{output1\_inter} + \texttt{output1} \times \texttt{batch\_cnt})/(\texttt{batch\_cnt} + 1);$$

9: Update

$$\texttt{mean\_sq} = \left(\frac{1}{B}\sum_{b=1}^{B} \exp\left(2\sum_{\ell=1}^{m} \beta_\ell k(w_\ell, Y_b)\right) + \texttt{mean\_sq} \times \texttt{batch\_cnt}\right)\Big/(\texttt{batch\_cnt} + 1);$$

10: Update `se1` $= \sqrt{\texttt{mean\_sq} - \texttt{output1}^2}$;

11: Update `batch_cnt` $=$ `batch_cnt` $+ 1$;

`/* Approximation of` $\exp(\widetilde{A}(\boldsymbol{\beta}))$ `is output1 */`

Algorithm 2 Batch Monte Carlo method (continued)

`/* Approximation of` $\nabla\widetilde{A}(\boldsymbol{\beta})$ `*/`

Draw $Y_1, \cdots, Y_B$ from μ and set $\texttt{batch\_cnt} = 1$; ▷ *Draw the first batch*

Approximate $[\nabla\widetilde{A}(\boldsymbol{\beta})]_j$ using $Y_1, \cdots, Y_B$ according to (4.9), for all $j = 1, \cdots, m$; let the resulting approximation of $\nabla\widetilde{A}(\boldsymbol{\beta})$ be `output2`;

Let $\texttt{mean\_sq}_j = \frac{1}{B}\sum_{b=1}^{B}\big(k(w_j, Y_b)\exp\big(\sum_{\ell=1}^{m}\beta_\ell k(w_\ell, Y_b) - \log(\texttt{output1})\big)\big)^2$ for all $j = 1, \cdots, m$;

Let $\texttt{se\_grad}_j = \sqrt{\texttt{mean\_sq}_j - \texttt{output2}_j^2}$ and $\texttt{se\_grad} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\texttt{se\_grad}_j^2}$;

while $\texttt{se\_grad} > \epsilon$ **do**

  Draw $Y_1, \cdots, Y_B$ from μ; ▷ *Draw more batches*

  Approximate $[\nabla\widetilde{A}(\boldsymbol{\beta})]_j$ using $Y_1, \cdots, Y_B$ according to (4.9), for all $j = 1, \cdots, m$; let the resulting approximation of $\nabla\widetilde{A}(\boldsymbol{\beta})$ be `output2_inter`;

  Update

$$\texttt{output2} = \big(\texttt{output2\_inter} + \texttt{output2} \times \texttt{batch\_cnt}\big)\big/(\texttt{batch\_out} + 1);$$

  Update

$$\texttt{mean\_sq}_j = \left(\frac{1}{B}\sum_{b=1}^{B}\left(k(w_j, Y_b)\exp\left(\sum_{\ell=1}^{m}\beta_\ell k(w_\ell, Y_b) - \log(\texttt{output1})\right)\right)^2 + \texttt{mean\_sq}_j \times \texttt{batch\_out}\right)\Big/(\texttt{batch\_out} + 1),$$

  for all $j = 1, \cdots, m$;

  Update

$$\texttt{se\_grad}_j = \sqrt{\texttt{mean\_sq}_j - \texttt{output2}_j^2}, \quad \text{and} \quad \texttt{se\_grad} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\texttt{se\_grad}_j^2};$$

  Update $\texttt{batch\_cnt} = \texttt{batch\_cnt} + 1$;

return $\texttt{output1} \in \mathbb{R}$ and $\texttt{output2} \in \mathbb{R}^m$.

4.1.3.2 Gradient Descent Algorithm to Minimize $\widetilde{J}_{\mathrm{NLL},\lambda}$

With the batch Monte Carlo method described in the preceding section, we provide the complete gradient descent algorithm to minimize $\widetilde{J}_{\mathrm{NLL},\lambda}$ over $\mathbb{R}^m$ in Algorithm 4.3.

Algorithm 4.3 Gradient descent algorithm to minimize $\widetilde{J}_{\mathrm{NLL},\lambda}$ over $\mathbb{R}^m$

Require:

- $X_1, \cdots, X_n$, data;
- $w_1, \cdots, w_m$, points at which kernel functions are centered;
- $\lambda \geq 0$, penalty parameter;
- $\boldsymbol{\beta}_{\mathrm{NLL}}^{(0)} \in \mathbb{R}^m$, starting point;
- $\tau > 0$, step size;
- $\epsilon > 0$, the tolerance parameter to determine when to terminate the gradient descent algorithm.

Compute $\mathbf{K}_1 \in \mathbb{R}^{m \times n}$ and $\mathbf{K}_2 \in \mathbb{R}^{m \times m}$ that appear in (4.5);
Set $\texttt{coef} = \boldsymbol{\beta}_{\mathrm{NLL}}^{(0)}$;
Compute $\nabla \widetilde{A}(\texttt{coef})$ using Algorithm 4.2;
Compute $\nabla \widetilde{J}_{\mathrm{NLL},\lambda}(\texttt{coef})$ by (4.6);
Compute $\texttt{error} = \|\nabla \widetilde{J}_{\mathrm{NLL},\lambda}(\texttt{coef})\|_2 / \sqrt{m}$;
while $\texttt{error} > \varepsilon$ **do**
    Update $\texttt{coef} = \texttt{coef} - \tau \nabla \widetilde{J}_{\mathrm{NLL},\lambda}(\texttt{coef})$;
    Compute $\nabla \widetilde{A}(\texttt{coef})$ using Algorithm 4.2;
    Compute $\nabla \widetilde{J}_{\mathrm{NLL},\lambda}(\texttt{coef})$ by (4.6);
    Update $\texttt{error} = \|\nabla \widetilde{J}_{\mathrm{NLL},\lambda}(\texttt{coef})\|_2 / \sqrt{m}$.
return coef.

4.1.4 Numerical Illustration

We use the `waiting` variable in the Old Faithful Geyser dataset to illustrate the algorithms introduced in the preceding two sections — one on seeking an appropriate finite-dimensional approximating space, and the other on computing the minimizer of the penalized NLL loss functional.

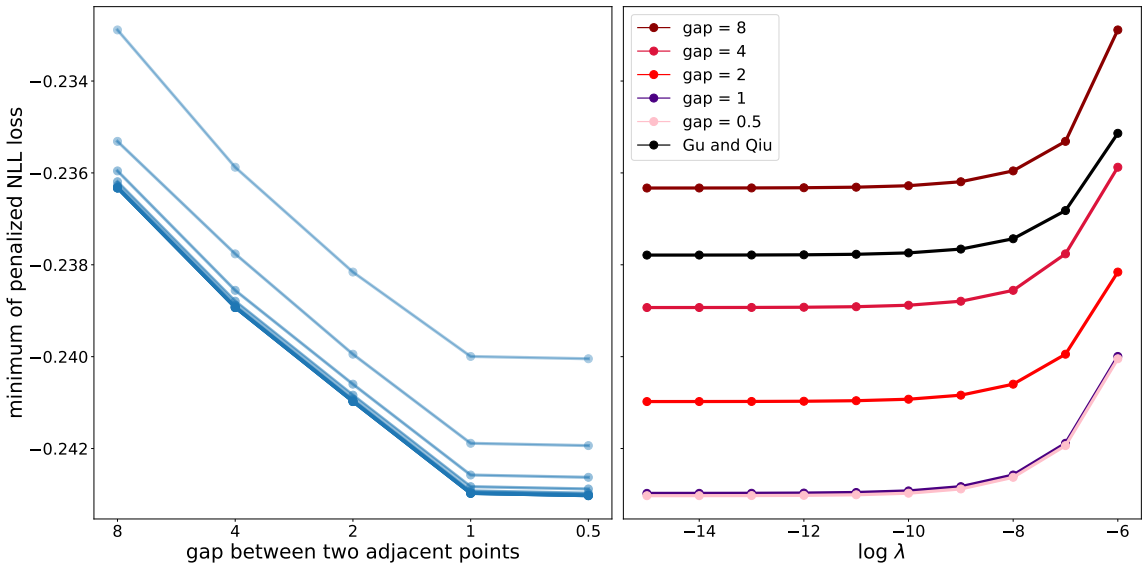

Figure 4.1: Left panel: the minimum of $\widetilde{J}_{\mathrm{NLL},\lambda}$ against the gap between two adjacent points at which kernel functions are centered in different choices of finite-dimensional approximating subspace. Different opacity indicates different values of λ, and the more opaque line indicates the smaller λ value. Right panel: the minimum of $\widetilde{J}_{\mathrm{NLL},\lambda}$ against different values of $\log \lambda$.

We start from the set of grid points $\{1, 9, 17, \cdots, 201\}$, where the difference between two adjacent grid points is 8. Note that this starting set of grid points contains 26 points and covers the range of raw data. In approximating $\nabla \widetilde{A}$ using Algorithm 4.2, we choose the batch size to be $B = 5000$ and the tolerance parameter to be 10^{-3}. In applying Algorithm 4.3, we choose the tolerance parameter to be 10^{-4}. The gradient descent algorithm always starts from the zero vector. The values of λ we choose are $e^{-15}, e^{-14}, \cdots, e^{-6}$.

The left panel of Figure 4.1 shows the minimum of $\widetilde{J}_{\mathrm{NLL},\lambda}$ against the gap between two adjacent grid points at which kernel functions are centered in different choices of finite-dimensional approximating subspace. For each value of λ, we see that, as the gap between two adjacent grid points becomes smaller and smaller, the minimum of

$\widetilde{J}_{\text{NLL},\lambda}$ keeps decreasing. When the difference between two adjacent grid points is 1, if we further insert additional grid points, the further reduction on the minimum of $\widetilde{J}_{\text{NLL},\lambda}$ becomes negligible. Thus, for this particular `waiting` data, we use the following subspace as our final choice of the finite-dimensional approximating subspace

$$\widetilde{\mathcal{H}} := \left\{ f \;\middle|\; f := \sum_{j=1}^{201} \beta_j k(w_j, \cdot), \beta_j \in \mathbb{R}, w_j = j, \text{ for all } j = 1, \cdots, 201 \right\}. \tag{4.10}$$

The right panel of Figure 4.1 shows the minimum of $\widetilde{J}_{\text{NLL},\lambda}$ against different values of $\log \lambda$. Similar to our observations from the left panel, as the gap between two adjacent grid points becomes smaller and smaller, the minimum of $\widetilde{J}_{\text{NLL},\lambda}$ keeps decreasing for all choices of λ. Comparing the minimums of $\widetilde{J}_{\text{NLL},\lambda}$ over the finite-dimensional subspaces with the gap between two adjacent grid points being 1 and 0.5, the difference is very tiny. In addition, the black curve in this panel shows the minimums of $\widetilde{J}_{\text{NLL},\lambda}$ using the finite-dimensional approximating subspace (4.3) that Gu and Qiu (1993) and Gu (1993) proposed. It is obvious that our approach yields a smaller minimum of $\widetilde{J}_{\text{NLL},\lambda}$ than their approach does, implying the superiority of our approach.

4.2 Regularized SM Density Estimators with $f \in \widetilde{\mathcal{H}}$

We minimize the penalized NLL loss functional over a finite-dimensional approximating space. Since our ultimate goal of this chapter is to compare the regularized SM density estimators and the penalized ML density estimator and understand their similarities and differences, in order to ensure the comparability, we also minimize the (penalized) SM loss functional over the same finite-dimensional subspace of $\mathcal{H}$. We describe how to achieve this in the current section.

Throughout this section, we again let

$$\widetilde{\mathcal{H}} := \left\{ f \;\middle|\; f := \sum_{j=1}^{m} \beta_j k(w_j, \cdot), \beta_j \in \mathbb{R} \text{ for all } j = 1, 2, \cdots, m \right\}$$

be a finite-dimensional approximating subspace, where $\{w_1, \cdots, w_m\} \subset \mathcal{X}$ is a set of specified grid points. We first derive the expression of $\widehat{J}_{\mathrm{SM}}(f)$ with $f \in \widetilde{\mathcal{H}}$, where $\widehat{J}_{\mathrm{SM}}$ is the SM loss functional given by (2.13) in Chapter 2.

Proposition 4.2. *Let $f = \sum_{j=1}^{m} w_j k(w_j, \cdot) \in \widetilde{\mathcal{H}}$. Then, the SM loss functional can be written as*

$$\widetilde{J}_{\mathrm{SM}}(\boldsymbol{\beta}) := \frac{1}{2}\boldsymbol{\beta}^\top \left(\frac{1}{n} \mathbf{S}\mathbf{S}^\top \right) \boldsymbol{\beta} - \boldsymbol{\beta}^\top \mathbf{t},$$

where $\boldsymbol{\beta} := (\beta_1, \cdots, \beta_m)^\top \in \mathbb{R}^m$, the $(j, (i-1)d + u)$-entry of $\mathbf{S} \in \mathbb{R}^{m \times (nd)}$ is $\langle k(w_j, \cdot), \partial_u k(X_i, \cdot) \rangle_{\mathcal{H}} = \partial_u k(X_i, w_j)$, and the j-th entry of $\mathbf{t} \in \mathbb{R}^m$ is

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

for all $j = 1, \cdots, m$, $i = 1, \cdots, n$ and $u = 1, \cdots, d$.

Proof of Proposition 4.2 can be found in Section 4.5.3.

4.2.1 Penalized SM Density Estimator with $f \in \widetilde{\mathcal{H}}$

In order to compute the penalized SM density estimator under the constraint $f \in \widetilde{\mathcal{H}}$, we minimize the penalized SM loss functional $\widehat{J}_{\mathrm{SM}}(f) + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2$ subject to $f \in \widetilde{\mathcal{H}}$. The following proposition establishes the existence and the uniqueness of this minimizer, and discusses how to compute it.

Proposition 4.3. *The minimizer of $\widehat{J}_{\mathrm{SM}}(f) + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2$ subject to $f \in \widetilde{\mathcal{H}}$, denoted by $\tilde{f}_{\mathrm{SM}}^{(\rho)}$, exists and is unique, and is given by*

$$\tilde{f}_{\mathrm{SM}}^{(\rho)} = \sum_{j=1}^{m} \beta_j^{(\rho)} k(w_j, \cdot),$$

where $\boldsymbol{\beta}_{\mathrm{SM}}^{(\rho)} := (\beta_1^{(\rho)}, \beta_2^{(\rho)}, \cdots, \beta_m^{(\rho)})^\top \in \mathbb{R}^m$ *satisfies*

$$\left(\frac{1}{n}\mathbf{S}\mathbf{S}^\top + \lambda \mathbf{K}_2\right)\boldsymbol{\beta}_{\mathrm{SM}}^{(\rho)} = \mathbf{t},$$

where $\mathbf{S} \in \mathbb{R}^{m\times(nd)}$ *and* $\mathbf{t} \in \mathbb{R}^m$ *are the same as those defined in Proposition 4.2, and* $\mathbf{K}_2 \in \mathbb{R}^{m\times m}$ *is the same as the one used in* (4.5).

Proof of Proposition 4.3 can be found in Section 4.5.4.

4.2.2 Early Stopping SM Density Estimator with $f \in \widetilde{\mathcal{H}}$

We now discuss how to compute the early stopping SM density estimator with $f \in \widetilde{\mathcal{H}}$. We apply the gradient descent algorithm to minimizing $\widetilde{J}_{\mathrm{SM}}$ over $\mathbb{R}^m$. The corresponding gradient descent iterates are

$$\boldsymbol{\beta}_{\mathrm{SM}}^{(t+1)} = \boldsymbol{\beta}_{\mathrm{SM}}^{(t)} - \tau[\mathbf{M}\boldsymbol{\beta}_{\mathrm{SM}}^{(t)} - \mathbf{t}] = (\mathbf{I} - \tau\mathbf{M})\boldsymbol{\beta}_{\mathrm{SM}}^{(t)} + \tau\mathbf{t}, \tag{4.11}$$

where $\mathbf{M} := \frac{1}{n}\mathbf{S}\mathbf{S}^\top$ for notational simplicity, $\boldsymbol{\beta}_{\mathrm{SM}}^{(0)} \in \mathbb{R}^m$ is the starting point of the algorithm, and $\tau > 0$ is the appropriately chosen constant step size.

The following proposition links $\boldsymbol{\beta}_{\mathrm{SM}}^{(t)}$ to $\boldsymbol{\beta}_{\mathrm{SM}}^{(0)}$.

Proposition 4.4. *For all* $t \in \mathbb{N}_0$, *we have*

$$\boldsymbol{\beta}_{\mathrm{SM}}^{(t+1)} = (\mathbf{I} - \tau\mathbf{M})^{t+1}\boldsymbol{\beta}_{\mathrm{SM}}^{(0)} + \tau\sum_{\ell=0}^{t}(\mathbf{I} - \tau\mathbf{M})^{\ell}\mathbf{t}. \tag{4.12}$$

If, in particular, we choose $\boldsymbol{\beta}_{\mathrm{SM}}^{(0)} = \mathbf{0}_m$, *the* m*-dimensional zero vector, we have*

$$\boldsymbol{\beta}_{\mathrm{SM}}^{(t+1)} = \tau\mathbf{Q}\widetilde{\boldsymbol{\Lambda}}^{(t+1)}\mathbf{Q}^\top\mathbf{t}, \qquad \text{for all } t \in \mathbb{N}_0, \tag{4.13}$$

where $\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top = \mathbf{M}$ *be the eigen-decomposition of* $\mathbf{M}$, $\mathbf{Q}$ *is an orthogonal matrix,* $\boldsymbol{\Lambda}$ *is a diagonal matrix with all eigenvalues of* $\mathbf{M}$, $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_m$, *on the diagonal, and* $\widetilde{\boldsymbol{\Lambda}}^{(t+1)}$ *is the diagonal matrix with elements on the diagonal being*

$$\tilde{\lambda}_j^{(t+1)} = \begin{cases} \dfrac{1-(1-\tau\lambda_j)^{t+1}}{\tau\lambda_j}, & \text{if } \lambda_j \neq 0, \\ t+1, & \text{if } \lambda_j = 0, \end{cases}$$

for all $j = 1, \cdots, m$ *and* $t \in \mathbb{N}_0$.

Proof of Proposition 4.4 can be found in Section 4.5.5.

Using Proposition 4.4, the $(t+1)$-st gradient descent iterate of $\widehat{J}_{\mathrm{SM}}$ in $\widetilde{\mathcal{H}}$ is

$$\sum_{j=1}^{m} \beta_j^{(t+1)} k(w_j, \cdot),$$

where $\beta_j^{(t+1)}$ is the j-th component of $\boldsymbol{\beta}_{\mathrm{SM}}^{(t+1)}$.

4.3 Comparison of Regularized ML and SM Density Estimators

With the algorithms of minimizing the penalized NLL loss functional and the (penalized) SM loss functional over a finite-dimensional approximating subspace of $\mathcal{H}$ shown in the preceding sections, we now use numerical examples to compare the corresponding density estimators and understand their similarities and differences.

We again use the `waiting` variable in the Old Faithful Geyser dataset. Figure 4.2 shows the resulting density estimates with different choices of penalty parameters or numbers of iterations. All these density estimates are computed over the finite-dimensional subspace (4.10). We use the same $\mathcal{X}$, μ, and k as in Chapter 3.

From the second and third rows of Figure 4.2, we see that, similar to our findings in Chapter 3, two kinds of regularized SM density estimates are still qualitatively very similar.

The penalized ML and regularized SM density estimates are qualitatively very similar when there is a large or intermediate amount of regularization. However, they are very different when there is very small regularization (small penalty parameter values or a large number of iterations). Specifically, when a very small regularization is imposed, regularized SM density estimates contains a bump or becomes a spike at the isolated observation, but penalized ML density estimates do not, even when there is no penalty (the case when $\lambda = 0$).

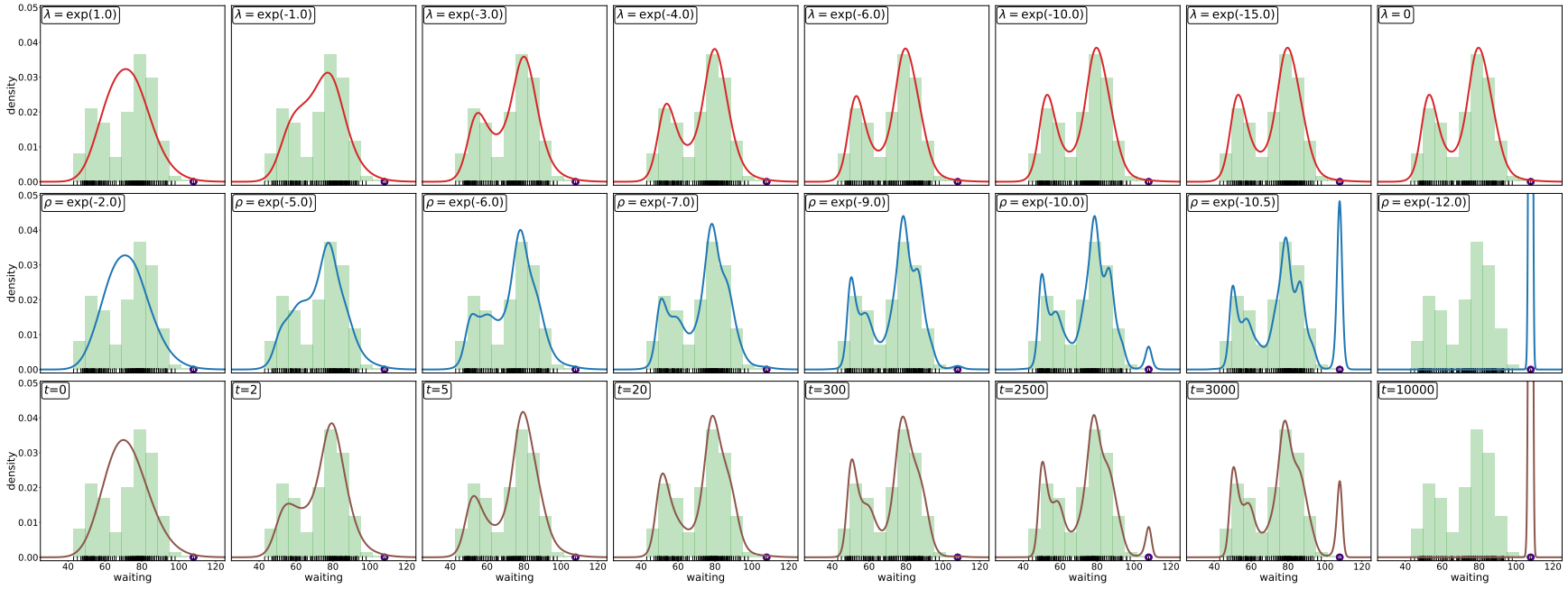

Figure 4.2: Penalized ML (first row), penalized SM (second row), and early stopping SM (third row) density estimates of the `waiting` variable. Histogram of data with the bin width chosen by the Freedman-Diaconis rule is shown in green. The rug plot indicates the location of data and the purple circle indicates the location of the isolated observation 108.

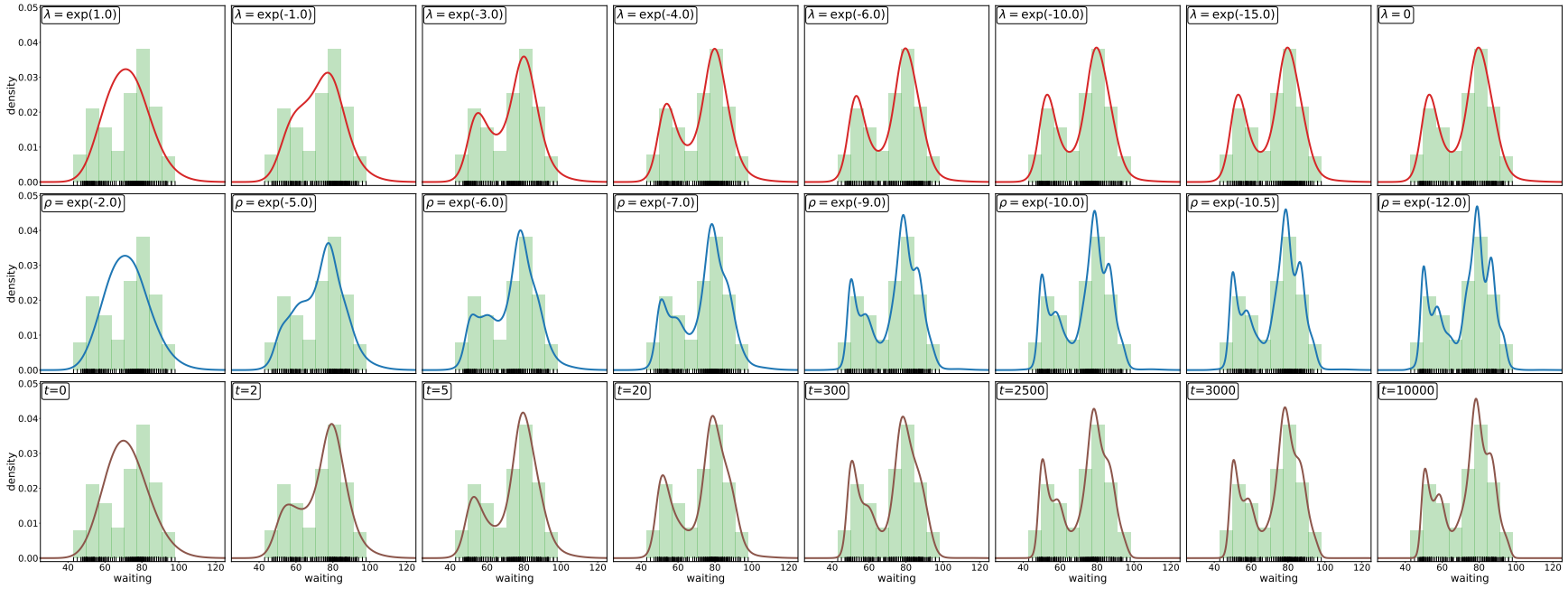

Figure 4.3: Penalized ML (first row), penalized SM (second row), and early stopping SM (third row) density estimates of the `waiting` variable with the isolated observation 108 removed. Histogram of data with the bin width chosen by the Freedman-Diaconis rule is shown in green. The rug plot indicates the location of data.

If we remove this isolated observation 108, as is shown in Figure 4.3, regularized SM density estimates do not contain a spike when the regularization is small.

Based on these numerical examples, we conclude that the regularized SM density estimators are very sensitive to the presence of an isolated observation, especially when there is a small amount of regularization. Additional numerical examples confirm this.

4.4 Discussion on the Presence of a Spike in SM Density Estimates

We now attempt to explain why the regularized SM density estimates put a spike at the isolated observation when regularization is very small. With $d = 1$, we can write the SM loss functional as

$$\widehat{L}_{\mathrm{SM}}(q_f) = \underbrace{\frac{1}{n}\sum_{i\neq i^*}\left(\frac{1}{2}\big((\log q_f)'(X_i)\big)^2 + (\log q_f)''(X_i)\right)}_{=:\mathrm{(I)}} + \underbrace{\frac{1}{n}\left(\frac{1}{2}\big((\log q_f)'(X_{i^*})\big)^2 + (\log q_f)''(X_{i^*})\right)}_{=:\mathrm{(II)}},$$

where X_{i^*} denotes the isolated observation.

Recall that the penalized SM density estimator is obtained by minimizing $\widehat{L}_{\mathrm{SM}}(q_f) + \frac{\rho}{2}\|f\|^2_{\mathcal{H}}$. When ρ is tiny, we are effectively minimizing the $\widehat{L}_{\mathrm{SM}}$ part. Notice $\log q_f$ is a linear combination of $\log \mu$ and Gaussian kernel functions centered at a set of grid points or the first two derivatives of Gaussian kernels centered at data points (depending on which basis functions we use), and these Gaussian kernel functions and derivatives of Gaussian kernel functions are local basis functions. Also, note that a spike is essentially a local maximum. Consequently, putting a spike in $\log q_f$ at X_{i^*} has the effects of forcing $((\log q_f)'(X_{i^*}))^2 \approx 0$ and reducing the value of $(\log q_f)''(X_{i^*})$

and, hence, that of (II) a lot, without affecting (I) much. Thus, putting a spike at X_{i^*} can reduce the value of $\widehat{L}_{\mathrm{SM}}$, coinciding with the goal of minimizing $\widehat{L}_{\mathrm{SM}}$.

A similar explanation holds for the early stopping SM density estimator. As we keep running the gradient descent algorithm, we are searching for a $q_f \in \mathcal{Q}_{\mathrm{ker}}$ that can reduce the value of $\widehat{L}_{\mathrm{SM}}$ as much as possible. With an isolated observation X_{i^*} being present and local basis functions being used, putting a spike at X_{i^*} can achieve this goal.

4.5 Proofs

4.5.1 Proof of Proposition 4.1

Proof of Proposition 4.1. First, by **(A2)** in Chapter 2, we know $\mathcal{F} = \mathcal{H}$ so that minimizing (4.1) over $\mathcal{H}$ is an unconstrained minimization problem.

By **(A3)** and Theorem 2.2 in Chapter 2, we know A is strictly convex. In addition, since $\frac{1}{n}\sum_{i=1}^n f(X_i) = \frac{1}{n}\sum_{i=1}^n \langle f, k(X_i, \cdot)\rangle$ is convex in f, we conclude $\widehat{J}_{\text{NLL}}$ is convex. In addition, note that the functional $f \mapsto \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$ is strongly convex with constant $\lambda > 0$. It follows that the objective functional in (4.1) is strongly convex with constant $\lambda > 0$ (Proposition 10.8 in Bauschke and Combettes, 2011). Finally, Corollary 11.17 in Bauschke and Combettes (2011) implies (4.1) has exactly one minimizer in $\mathcal{H}$. ■

4.5.2 Details about Example 4.1

We provide details about Example 4.1 here, and will follow three steps:

Step 1. Identify the RKHS $\mathcal{H}$ associated with the choice of k and the corresponding natural parameter space $\mathcal{F}$;

Step 2. Identify $\text{Span}\{k(X_1, \cdot)\}$ and $\text{Span}\{k(X_1, \cdot)\} \cap \mathcal{F}$;

Step 3. Explcitly compute the minimizer of (4.2) over $\mathcal{F}$ and show this minimizer does not reside in $\text{Span}\{k(X_1, \cdot)\} \cap \mathcal{F}$.

Step 1. Letting $x = (x_1, x_2)^\top \in \mathbb{R}^2$ and $y = (y_1, y_2)^\top \in \mathbb{R}^2$ be arbitrary, we have

$$k(x, y) = (x^\top y)^2 = x_1^2 y_1^2 + 2x_1 x_2 y_1 y_2 + x_2^2 y_2^2 = \langle \Phi(x), \Phi(y)\rangle, \qquad \text{for all } x, y \in \mathbb{R},$$

where $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$ denotes the feature map of k and is given by

$$\Phi(x) = \left(x_1^2, \sqrt{2}x_1 x_2, x_2^2\right)^\top,$$

and the inner product $\langle \cdot, \cdot \rangle$ is the dot product in $\mathbb{R}^3$. Then, the resulting $\mathcal{H}$ contains all functions mapping from $\mathbb{R}^2$ to $\mathbb{R}$ of the form

$$f(x) = w_1 x_1^2 + \sqrt{2} w_2 x_1 x_2 + w_3 x_2^2, \qquad \text{for all } x = (x_1, x_2)^\top \in \mathbb{R}^2,$$

for some $w_1, w_2, w_3 \in \mathbb{R}$ with the RKHS norm of f being $\|f\|_{\mathcal{H}} = \sqrt{w_1^2 + w_2^2 + w_3^2}$ (Theorem 4.21 in Steinwart and Christmann, 2008).

We now derive the natural parameter space $\mathcal{F}$. Let $f(x) = w_1 x_1^2 + \sqrt{2} w_2 x_1 x_2 + w_3 x_2^2$ for some $w_1, w_2, w_3 \in \mathbb{R}$ and all $x = (x_1, x_2)^\top \in \mathbb{R}^2$. By simple algebra, we have

$$e^{A(f)} = \int \int_{\mathbb{R}^2} \frac{1}{2\pi} \exp\left(-\frac{\|x\|_2^2}{2}\right) \exp\big(f(x)\big) \mathrm{d}x_1 \mathrm{d}x_2 = |W|^{1/2},$$

where $|W|$ denotes the determinant of the matrix W,

$$W^{-1} = \begin{pmatrix} 1 - 2w_1 & -\sqrt{2} w_2 \\ -\sqrt{2} w_2 & 1 - 2w_3 \end{pmatrix},$$

and we assume $(1 - 2w_1)(1 - 2w_3) - 2w_2^2 > 0$. Then,

$$A(f) = \log\big(|W|^{1/2}\big) = -\frac{1}{2} \log\big((1 - 2w_1)(1 - 2w_3) - 2w_2^2\big).$$

Note that $A(f) < \infty$ if and only if W is positive definite if and only if $1 - 2w_1 > 0$ and $1 - 2w_3 > 0$ and $(1 - 2w_1)(1 - 2w_3) - 2w_2^2 > 0$ if and only if $w_1 < \frac{1}{2}$ and $w_3 < \frac{1}{2}$ and $(1 - 2w_1)(1 - 2w_3) - 2w_2^2 > 0$.

As the conclusion of Step 1, we have

$$\mathcal{H} = \Big\{ f : \mathbb{R}^2 \to \mathbb{R} \;\Big|\; f(x) := w_1 x_1^2 + \sqrt{2} w_2 x_1 x_2 + w_3 x_2^2 \text{ for all } x = (x_1, x_2)^\top \in \mathbb{R}^2, \\ w_1, w_2, w_3 \in \mathbb{R} \Big\}, \tag{4.14}$$

$$\mathcal{F} = \Big\{ f : \mathbb{R}^2 \to \mathbb{R} \;\Big|\; f(x) = w_1 x_1^2 + \sqrt{2} w_2 x_1 x_2 + w_3 x_2^2 \text{ for all } x = (x_1, x_2)^\top \in \mathbb{R}^2, \\ w_1 < \frac{1}{2}, w_3 < \frac{1}{2}, (1 - 2w_1)(1 - 2w_3) - 2w_2^2 > 0, w_1, w_2, w_3 \in \mathbb{R} \Big\}. \tag{4.15}$$

Step 2. With $X_1 = (a, 0)^\top$, it is easy to see $\mathrm{Span}\{k(X_1, \cdot)\}$ contains all functions of the form

$$f(x) = \gamma k(X_1, x) = \gamma (X_1^\top x)^2 = \gamma a^2 x_1^2, \qquad \text{for all } x := (x_1, x_2)^\top \in \mathbb{R}^2 \text{ and } \gamma \in \mathbb{R}.$$

Then, we can characterize $\mathrm{Span}\{k(X_1, \cdot)\} \cap \mathcal{F}$ as

$$\mathrm{Span}\{k(X_1, \cdot)\} \cap \mathcal{F} = \left\{ f : \mathbb{R}^2 \to \mathbb{R} \;\middle|\; f(x) = \gamma a^2 x_1^2 \text{ for all } x = (x_1, x_2)^\top \in \mathbb{R}^2, \gamma a^2 < \frac{1}{2} \right\}.$$

Step 3. Let $f \in \mathcal{F}$ be

$$f(x) = w_1 x_1^2 + \sqrt{2} w_2 x_1 x_2 + w_3 x_2^2, \qquad \text{for all } x = (x_1, x_2)^\top \in \mathbb{R}^2,$$

where $w_1, w_2, w_3 \in \mathbb{R}$ satisfy the constraints in (4.15). Then, with $\lambda > 0$, $\widehat{J}_{\mathrm{NLL}}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$ becomes

$$\widetilde{J}_\lambda(w_1, w_2, w_3) := -\frac{1}{2}\log\big((1-2w_1)(1-2w_3) - 2w_2^2\big) - w_1 a^2 + \frac{\lambda}{2}(w_1^2 + w_2^2 + w_3^2).$$

By calculus, we have

$$\nabla \widetilde{J}_\lambda(w_1, w_2, w_3) = \begin{pmatrix} \frac{1-2w_3}{(1-2w_1)(1-2w_3)-2w_2^2} - a^2 + \lambda w_1 \\ \frac{2w_2}{(1-2w_1)(1-2w_3)-2w_2^2} + \lambda w_2 \\ \frac{1-2w_1}{(1-2w_1)(1-2w_3)-2w_2^2} + \lambda w_3 \end{pmatrix}.$$

Any stationary points, denoted by (w_1^*, w_2^*, w_3^*), must satisfy $\nabla \widetilde{J}_\lambda(w_1^*, w_2^*, w_3^*) = 0$, from which we obtain the following system of equations

$$\begin{cases} (1-2w_3^*) + (\lambda w_1^* - a^2)\big((1-2w_1^*)(1-2w_3^*) - 2w_2^{*2}\big) & = 0, \\ 2w_2^* + \lambda w_2^*\big((1-2w_1^*)(1-2w_3^*) - 2w_2^{*2}\big) & = 0, \\ (1-2w_1^*) + \lambda w_3^*\big((1-2w_1^*)(1-2w_3^*) - 2w_2^{*2}\big) & = 0. \end{cases}$$

Solving the preceding system for (w_1^*, w_2^*, w_3^*) yields the following solutions:

$$\left(\frac{1}{2}, 0, \frac{1}{2}\right), \quad \big(b_1^*, 0, c_1^*\big), \quad \big(b_1^*, 0, c_2^*\big), \quad \big(b_2^*, 0, c_1^*\big), \quad \big(b_2^*, 0, c_2^*\big), \tag{4.16}$$

where

$$b_1^* := \frac{(\lambda + 2a^2) + \sqrt{(\lambda - 2a^2)^2 + 8\lambda}}{4\lambda}, \qquad b_2^* := \frac{-(\lambda + 2a^2) + \sqrt{(\lambda - 2a^2)^2 + 8\lambda}}{-4\lambda},$$

$$c_1^* := \frac{\lambda + \sqrt{\lambda^2 + 8\lambda}}{4\lambda}, \qquad\qquad c_2^* := \frac{\lambda - \sqrt{\lambda^2 + 8\lambda}}{4\lambda}.$$

We next check whether the corresponding functions belong to $\mathcal{F}$ or not.

First, notice that the pair of $(w_1^*, w_2^*, w_3^*) = (\frac{1}{2}, 0, \frac{1}{2})$ leads to the function $f(x) = \frac{1}{2}(x_1^2 + x_2^2)$, which does *not* belong to $\mathcal{F}$ since $w_1^* = \frac{1}{2}$ and $w_3^* = \frac{1}{2}$. We discard it.

In addition, since $\lambda > 0$, we have

$$b_1^* = \frac{(\lambda + 2a^2) + \sqrt{(\lambda - 2a^2)^2 + 8\lambda}}{4\lambda} > \frac{(\lambda + 2a^2) + |\lambda - 2a^2|}{4\lambda}.$$

Consider the following two cases:

(i) If $2a^2 - \lambda \geq 0$, that is, $a^2/\lambda \geq \frac{1}{2}$, we have $b_1^* > \frac{\lambda + 2a^2 + 2a^2 - \lambda}{4\lambda} = \frac{a^2}{\lambda} \geq \frac{1}{2}$;

(ii) If $2a^2 - \lambda < 0$, we have $b_1^* > \frac{\lambda + 2a^2 + \lambda - 2a^2}{4\lambda} = \frac{2\lambda}{4\lambda} \geq \frac{1}{2}$.

In both cases, we have $b_1^* > \frac{1}{2}$, and any function f with $w_1^* = b_1^*$ cannot belong to $\mathcal{F}$.

Next, we show $b_2^* < \frac{1}{2}$, which is equivalent to showing $1 - 2b_2^* > 0$. Since

$$\begin{aligned} 1 - 2b_2^* &= 1 - \frac{-(\lambda + 2a^2) + \sqrt{(\lambda - 2a^2)^2 + 8\lambda}}{-2\lambda} \\ &= 1 + \frac{\sqrt{(\lambda - 2a^2)^2 + 8\lambda} - (\lambda + 2a^2)}{2\lambda} \\ &> 1 + \frac{|\lambda - 2a^2| - (\lambda + 2a^2)}{2\lambda}. \end{aligned}$$

Again, we consider the following two cases:

(i) If $2a^2 - \lambda \geq 0$, we have $1 - 2b_2^* > 1 + \frac{2a^2 - \lambda - \lambda - 2a^2}{2\lambda} = 1 + \frac{-2\lambda}{2\lambda} = 0$;

(ii) If $2a^2 - \lambda < 0$, that is, $0 \leq a^2/\lambda < \frac{1}{2}$, we have $1 - 2b_2^* > 1 + \frac{\lambda - 2a^2 - \lambda - 2a^2}{2\lambda} = 1 - \frac{2a^2}{\lambda} > 0$.

As for c_1^* and c_2^*, since $\lambda > 0$, we have

$$c_1^* = \frac{\lambda + \sqrt{\lambda^2 + 8\lambda}}{4\lambda} > \frac{\lambda + \sqrt{\lambda^2}}{4\lambda} = \frac{1}{2},$$

$$c_2^* = \frac{\lambda - \sqrt{\lambda^2 + 8\lambda}}{4\lambda} < \frac{\lambda}{4\lambda} = \frac{1}{4} < \frac{1}{2}.$$

That is, any function f with $w_3^* = c_1^*$ cannot belong to $\mathcal{F}$.

As a conclusion, the only solution in (4.16) such that the resulting f belongs to $\mathcal{F}$ is $(w_1^*, w_2^*, w_3^*) = (b_2^*, 0, c_2^*)$.

In addition, it is easy to see that the Hessian matrix of $\widetilde{J}_\lambda$ at $(w_1^*, w_2^*, w_3^*) = (b_2^*, 0, c_2^*)$ is

$$\begin{pmatrix} \frac{2(1-c_2^*)^2}{(1-2b_2^*)^2(1-2c_2^*)^2} + \lambda & 0 & 0 \\ 0 & \frac{2(1-b_2^*)(1-c_2^*)}{(1-2b_2^*)^2(1-2c_2^*)^2} + \lambda & 0 \\ 0 & 0 & \frac{2(1-b_2^*)^2}{(1-2b_2^*)^2(1-2c_2^*)^2} + \lambda \end{pmatrix},$$

which is positive definite. We conclude that $\widetilde{J}_\lambda$ achieves the minimum at $(w_1^*, w_2^*, w_3^*) = (b_2^*, 0, c_2^*)$, and the corresponding function, $f^* := \arg\min_{f \in \mathcal{F}} \{\widehat{J}_{\mathrm{NLL}}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2\}$, is

$$f^*(x) = b_2^* x_1^2 + c_2^* x_2^2, \qquad \text{for all } x = (x_1, x_2)^\top \in \mathbb{R}^2.$$

Since $c_2^* \neq 0$ for all $\lambda > 0$, f^* does *not* belong to $\mathrm{Span}\{k(X_1, \cdot)\} \cap \mathcal{F}$. ∎

4.5.3 Proof of Proposition 4.2

Proof of Proposition 4.2. Since $f \in \widetilde{\mathcal{H}}$ takes on the form $f = \sum_{j=1}^m \beta_j k(w_j, \cdot)$ for $\beta_1, \cdots, \beta_m \in \mathbb{R}$, we first have

$$\begin{aligned} &\langle f, (\partial_u k(X_i, \cdot) \otimes \partial_u k(X_i, \cdot)) f \rangle_{\mathcal{H}} \\ &= \left\langle \sum_{\ell=1}^m \beta_\ell k(w_\ell, \cdot), (\partial_u k(X_i, \cdot) \otimes \partial_u k(X_i, \cdot)) \left(\sum_{j=1}^m \beta_j k(w_j, \cdot) \right) \right\rangle_{\mathcal{H}} \\ &= \sum_{\ell=1}^m \sum_{j=1}^m \beta_\ell \beta_j \langle \partial_u k(X_i, \cdot), k(w_\ell, \cdot) \rangle_{\mathcal{H}} \langle \partial_u k(X_i, \cdot), k(w_j, \cdot) \rangle_{\mathcal{H}} \\ &= \sum_{\ell=1}^m \sum_{j=1}^m \beta_\ell \beta_j \partial_u k(X_i, w_\ell) \partial_u k(X_i, w_j), \end{aligned}$$

where the last equality is due to the reproducing property of the partial derivative of k (see Proposition A.5 in Appendix A). Then, since $\widehat{C} = \frac{1}{n} \sum_{i=1}^n \sum_{u=1}^d \partial_u k(X_i, \cdot) \otimes$

$\partial_u k(X_i, \cdot)$, by the linearity of the inner product, we have

$$\begin{aligned}\langle f, \widehat{C} f\rangle_{\mathcal{H}} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\big\langle f, (\partial_u k(X_i, \cdot) \otimes \partial_u k(X_i, \cdot)) f\big\rangle_{\mathcal{H}} \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\sum_{\ell=1}^{m}\sum_{j=1}^{m}\beta_\ell \beta_j \partial_u k(X_i, w_\ell)\partial_u k(X_i, w_j) \\ &= \boldsymbol{\beta}^\top\left(\frac{1}{n}\mathbf{S}\mathbf{S}^\top\right)\boldsymbol{\beta}.\end{aligned}$$

As for $\langle f, \hat{z}\rangle_{\mathcal{H}}$, we have the following

$$\begin{aligned}\langle f, \hat{z}\rangle_{\mathcal{H}} &= \left\langle \sum_{j=1}^{m}\beta_j k(w_j, \cdot), -\frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\big(\partial_u^2 k(X_i, \cdot) + (\partial_u \log \mu)(X_i)\partial_u k(X_i, \cdot)\big)\right\rangle_{\mathcal{H}} \\ &= \sum_{j=1}^{m}\beta_j\left(-\frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\big(\partial_u^2 k(X_i, w_j) + (\partial_u \log \mu)(X_i)\partial_u k(X_i, w_j)\big)\right) \\ &= \boldsymbol{\beta}^\top \mathbf{t},\end{aligned}$$

where we use the reproducing property of the partial derivative of k in the second equality. The desired result follows by combining all pieces above together. ■

4.5.4 Proof of Proposition 4.3

Proof of Proposition 4.3. Recall from Chapters 2 and 3 that the operator $\widehat{C} : \mathcal{H} \to \mathcal{H}$ is positive semidefinite, $\widehat{J}_{\mathrm{SM}}$ is convex over $\mathcal{H}$. In addition, note that the functional $f \mapsto \frac{\rho}{2}\|f\|_{\mathcal{H}}^2$ is strongly convex with constant ρ. We conclude that the penalized SM loss functional $f \mapsto \widehat{J}_{\mathrm{SM}}(f) + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2$ is strongly convex with parameter $\lambda > 0$, and that it has exactly one minimizer (Corollary 11.17 in Bauschke and Combettes, 2011).

With $f = \sum_{j=1}^{m}\beta_j k(w_j, \cdot) \in \widetilde{\mathcal{H}}$, we have $\frac{\rho}{2}\|f\|_{\mathcal{H}}^2 = \frac{\rho}{2}\boldsymbol{\beta}^\top \mathbf{K}_2 \boldsymbol{\beta}$. Thus, together with Proposition 4.2, we can write the functional $\widehat{J}_{\mathrm{SM}}(f) + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2$ with $f \in \widetilde{\mathcal{H}}$ as

$$\widetilde{J}_{\mathrm{SM},\rho}(\boldsymbol{\beta}) := \frac{1}{2}\boldsymbol{\beta}^\top\left(\frac{1}{n}\mathbf{S}\mathbf{S}^\top\right)\boldsymbol{\beta} - \boldsymbol{\beta}^\top \mathbf{t} + \frac{\rho}{2}\boldsymbol{\beta}^\top \mathbf{K}_2 \boldsymbol{\beta}.$$

Then, $\boldsymbol{\beta}_{\mathrm{SM}}^{(\rho)} := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^m} \widetilde{J}_{\mathrm{SM},\rho}(\boldsymbol{\beta})$ must satisfy the first-order optimality condition

$$\left(\frac{1}{n}\mathbf{S}\mathbf{S}^\top + \rho\mathbf{K}_2\right)\boldsymbol{\beta}_{\mathrm{SM}}^{(\rho)} = \mathbf{t}.$$

As a consequence, $\arg\min_{f\in\widetilde{\mathcal{H}}}\{\widehat{J}_{\mathrm{SM}}(f)+\frac{\rho}{2}\|f\|_{\mathcal{H}}^2\}$ is $\sum_{j=1}^m \beta_j^{(\rho)}k(w_j,\cdot)$, where $\beta_j^{(\rho)}$ is the j-th element of $\boldsymbol{\beta}_{\mathrm{SM}}^{(\rho)}$. ■

4.5.5 Proof of Proposition 4.4

Proof of Proposition 4.4. We first prove (4.12) by induction. When $t = 0$, from (4.11), we have

$$\boldsymbol{\beta}_{\mathrm{SM}}^{(1)} = (\mathbf{I}-\tau\mathbf{M})\boldsymbol{\beta}_{\mathrm{SM}}^{(0)} + \tau\mathbf{t}. \tag{4.17}$$

From (4.12), we plug in $t = 0$ and obtain

$$\boldsymbol{\beta}_{\mathrm{SM}}^{(1)} = (\mathbf{I}-\tau\mathbf{M})\boldsymbol{\beta}_{\mathrm{SM}}^{(0)} + \tau\sum_{\ell=0}^{0}(\mathbf{I}-\tau\mathbf{M})^i\mathbf{t} = (\mathbf{I}-\tau\mathbf{M})\boldsymbol{\beta}_{\mathrm{SM}}^{(0)} + \tau\mathbf{t},$$

which is identical to (4.17).

Now, supposing (4.12) holds for $t = s$, we are going to show it also holds for $t = s + 1$. Note the following

$$\begin{aligned}
\boldsymbol{\beta}_{\mathrm{SM}}^{(s+1)} &= (\mathbf{I}-\tau\mathbf{M})\boldsymbol{\beta}_{\mathrm{SM}}^{(s)} + \tau\mathbf{t} \\
&= (\mathbf{I}-\tau\mathbf{M})\left[(\mathbf{I}-\tau\mathbf{M})^s\boldsymbol{\beta}_{\mathrm{SM}}^{(0)} + \tau\sum_{\ell=0}^{s-1}(\mathbf{I}-\tau\mathbf{M})^\ell\mathbf{t}\right] + \tau\mathbf{t} \\
&= (\mathbf{I}-\tau\mathbf{M})^{s+1}\boldsymbol{\beta}_{\mathrm{SM}}^{(0)} + \tau\sum_{\ell=0}^{s-1}(\mathbf{I}-\tau\mathbf{M})^{\ell+1}\mathbf{t} + \tau\mathbf{t} \\
&= (\mathbf{I}-\tau\mathbf{M})^{s+1}\boldsymbol{\beta}_{\mathrm{SM}}^{(0)} + \tau\sum_{\ell=1}^{s}(\mathbf{I}-\tau\mathbf{M})^{\ell}\mathbf{t} + \tau\mathbf{t} \\
&= (\mathbf{I}-\tau\mathbf{M})^{s+1}\boldsymbol{\beta}_{\mathrm{SM}}^{(0)} + \tau\sum_{\ell=0}^{s}(\mathbf{I}-\tau\mathbf{M})^{\ell}\mathbf{t},
\end{aligned}$$

which is the desired result.

We now prove (4.13). With $\boldsymbol{\beta}_{\mathrm{SM}}^{(0)} = \mathbf{0}_m$, we can simplify (4.12) to

$$\boldsymbol{\beta}_{\mathrm{SM}}^{(t+1)} = \tau\sum_{\ell=0}^{t}(\mathbf{I}-\tau\mathbf{M})^{\ell}\mathbf{t}, \qquad \text{for all } t\in\mathbb{N}_0.$$

Let $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top = \mathbf{M}$ be the eigen-decomposition of $\mathbf{M}$, where $\mathbf{Q}$ and $\mathbf{\Lambda}$ are defined in the proposition. Then,

$$\mathbf{I} - \tau\mathbf{M} = \mathbf{I} - \tau\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top = \mathbf{Q}\mathbf{Q}^\top - \tau\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top = \mathbf{Q}(\mathbf{I} - \tau\mathbf{\Lambda})\mathbf{Q}^\top,$$

and we have

$$\sum_{\ell=0}^{t}(\mathbf{I} - \tau\mathbf{M})^\ell = \sum_{\ell=0}^{t}\mathbf{Q}(\mathbf{I} - \tau\mathbf{\Lambda})^\ell\mathbf{Q}^\top = \mathbf{Q}\left[\sum_{\ell=0}^{t}(\mathbf{I} - \tau\mathbf{\Lambda})^\ell\right]\mathbf{Q}^\top.$$

Now, if $\lambda_j \neq 0$, we have

$$\sum_{\ell=0}^{t}(1 - \tau\lambda_j)^\ell = \frac{1 - (1 - \tau\lambda_j)^{t+1}}{\tau\lambda_j}; \tag{4.18}$$

and if $\lambda_j = 0$, we have

$$\sum_{\ell=0}^{t}(1 - \tau\lambda_j)^\ell = t + 1. \tag{4.19}$$

Therefore, if we let $\widetilde{\mathbf{\Lambda}}^{(t+1)}$ be the diagonal matrix with elements on the diagonal being (4.18) and (4.19), the desired result follows. ■

Chapter 5: Influence Function of a (Log-)Density Function and Its Properties

From numerical examples and discussions in Chapter 4, we have seen that the regularized SM density estimates with and without the presence of an isolated observation can be very different. This motivates us to study the sensitivity of the density estimators to the input data. The tool we use an extension of the classic influence function.

We will give a review of the influence function and its applications in Section 5.1, and then introduce our extension of this classic notion to the studies of density estimators in Section 5.2. We will focus on the finite-dimensional exponential family and the infinite-dimensional kernel exponential family in Sections 5.3 and 5.4, respectively, and discuss various properties of the influence functions of ML and SM (log-)density projections (to be defined) in them.

5.1 Influence Function and Its Applications in Statistics

Influence function, a classical notion from robust statistics, was first introduced by Hampel (1968) to investigate the infinitesimal behavior of real- and vector-valued statistical functionals and has become one of the most important tools in robust statistics (Hampel et al., 1986).

Let F be a distribution over the sample space $\mathcal{X}$, and T to be a statistical functional, any function of F. In this section, we assume that T is vector-valued so that $T(F) \in \mathbb{R}^m$. The *influence function* of T at F is defined to be

$$\mathrm{IF}(T, F, y) := \lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon}\Big(T\big((1-\varepsilon)F + \varepsilon\delta_y\big) - T(F)\Big) \tag{5.1}$$

$$= \frac{\mathrm{d}}{\mathrm{d}\varepsilon} T\big((1-\varepsilon)F + \varepsilon\delta_y\big)\Big|_{\varepsilon=0}, \tag{5.2}$$

where δ_y denotes the point mass 1 at $y \in \mathcal{X}$. Inspecting (5.1) and (5.2), we see $\mathrm{IF}(T, F, y)$ is nothing but the Gâteaux derivative of the statistical functional T at the distribution F in the direction of the point mass δ_y. Various properties of $\mathrm{IF}(T, F, y)$ have been discussed by Hampel (1974), Hampel et al. (1986), and Wasserman (2006).

The influence function has a nice interpretation. It measures the impact of an infinitesimally small amount of contamination of the original distribution F at y on the quantity of interest $T(F)$.

The *empirical influence function* (Definition 2.18 in Wasserman, 2006) is defined by letting F in $\mathrm{IF}(T, F, y)$ be F_n. In addition, the *sample influence function* (Hampel, 1974; Cook and Weisberg, 1982) is defined to be

$$\mathrm{SIF}_\varepsilon(T, F_n, y) := \frac{1}{\varepsilon}\Big(T\big((1-\varepsilon)F_n + \varepsilon\delta_y\big) - T(F_n)\Big), \tag{5.3}$$

where $\varepsilon > 0$ is tiny. Essentially, $\mathrm{SIF}_\varepsilon(T, F_n, y)$ is a finite-difference approximation of $\mathrm{IF}(T, F_n, y)$, and can be very useful when $T((1-\varepsilon)F_n + \varepsilon\delta_y)$ and $T(F_n)$ are analytically intractable or the limit by letting $\varepsilon \to 0^+$ is hard to handle. If, in particular, we let $\varepsilon = \frac{1}{n+1}$ in (5.3), we obtain Tukey's *sensitivity curve* which is used to assess the sensitivity of an estimator to the position of an additional observation *not* present in the sample (Hampel et al., 1986); and, if we let $\varepsilon = -\frac{1}{n-1}$ in (5.3), the resulting sample influence function has a close connection with the leave-one-out cross validation and the jackknife (see Example 3.18 in Wasserman, 2006, for a discussion).

We provide three examples below to illustrate the concept of the influence function.

Example 5.1 (Mean). Let $\mathcal{X} = \mathbb{R}$ and consider the statistical function defined by $T_1(F) = \int_{\mathbb{R}} x \mathrm{d}F(x) =: \mathbb{E}_F[X]$, where we assume $\mathbb{E}_F[X]$ exists. Since

$$\begin{aligned}\frac{1}{\varepsilon}\Big(T_1\big((1-\varepsilon)F + \varepsilon\delta_y\big) - T_1(F)\Big) &= \frac{1}{\varepsilon}\left(\int_{\mathcal{X}} x \mathrm{d}\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - \int_{\mathcal{X}} x \mathrm{d}F(x)\right) \\ &= \frac{1}{\varepsilon}\int_{\mathcal{X}} x \mathrm{d}(-\varepsilon F + \varepsilon\delta_y)(x) \\ &= \int_{\mathcal{X}} x \mathrm{d}(\delta_y - F)(x) \\ &= y - \mathbb{E}_F[X],\end{aligned}$$

we have

$$\mathrm{IF}(T_1, F, y) = y - \mathbb{E}_F[X]. \tag{5.4}$$

▶

Example 5.2 (Median). Let $\mathcal{X} = \mathbb{R}$ and consider the statistical function defined by $T_2(F) = F^{-1}(\frac{1}{2})$, the median of the distribution F. Here, we assume F has a density function p_0 that is symmetric around 0 and $p_0(0) \neq 0$. Then, we have $F^{-1}(\frac{1}{2}) = 0$. Let $F_{\varepsilon,y} := (1-\varepsilon)F + \varepsilon\delta_y$.

First consider the case $y = 0$. Then, for any $\varepsilon \in (0,1)$, we have $F_{\varepsilon,0}(x) < \frac{1}{2}(1-\varepsilon)$ for all $x < 0$ and $F_{\varepsilon,0}(x) > \frac{1}{2}(1+\varepsilon)$ for all $x > 0$. It follows that $F_{\varepsilon,0}^{-1}(\frac{1}{2}) = 0$ and $\mathrm{IF}(T_2, F, 0) = 0$.

Now, suppose $y \neq 0$. We must have $\frac{1}{2} = F_{\varepsilon,y}\big(F_{\varepsilon,y}^{-1}(\frac{1}{2})\big)$, which is equivalent to

$$\frac{1}{2} = (1-\varepsilon)F\left(F_{\varepsilon,y}^{-1}\left(\frac{1}{2}\right)\right) + \varepsilon\delta_y\left(F_{\varepsilon,y}^{-1}\left(\frac{1}{2}\right)\right).$$

Differentiating both sides of the preceding equation with respect to ε and evaluating at $\varepsilon = 0$, we have

$$0 = -F\left(F^{-1}\left(\frac{1}{2}\right)\right) + p_0\left(F^{-1}\left(\frac{1}{2}\right)\right)\left[\frac{\mathrm{d}F_{\varepsilon,y}^{-1}(\frac{1}{2})}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0}\right] + \delta_y\left(F^{-1}\left(\frac{1}{2}\right)\right)$$

$$= -\frac{1}{2} + p_0(0)\mathrm{IF}(T_2, F, y) + \delta_y(0).$$

Rearranging the preceding equation yields $\mathrm{IF}(T_2, F, y) = \frac{1-2\delta_y(0)}{2p_0(0)}$.

Now, if $y > 0$, $\delta_y(0) = 0$ and $\mathrm{IF}(T_2, F, y) = \frac{1}{2f(0)}$; if $y < 0$, $\delta_y(0) = 1$ and $\mathrm{IF}(T_2, F, y) = \frac{1-2}{2f(0)} = \frac{-1}{2f(0)}$. Summarizing all cases above, we have

$$\mathrm{IF}(T_2, F, y) = \frac{\mathrm{sign}(y)}{2f(0)}. \tag{5.5}$$

▶

Example 5.3 (M-estimator). The M-estimator, first proposed by Huber (1964), is a generalization of the maximum likelihood estimator and aims at designing new robust estimators. Here, we define an *M-estimator*, denoted by $T(F)$, to be the one that satisfies

$$\int_{\mathcal{X}} \psi(x, T(F))\mathrm{d}F(x) = 0,$$

where $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^m$ and $\Theta \subseteq \mathbb{R}^m$ is the parameter space. If, in particular, we let $\psi(x, \theta) = \frac{\partial}{\partial \eta} \log f_\eta(x)|_{\eta=\theta}$ for all $x \in \mathcal{X}$ and all $\theta \in \Theta$, where $\big\{f_\theta : \mathcal{X} \to [0, \infty) \mid \theta \in \Theta\big\}$ is the statistical model we assume, we obtain the maximum likelihood estimator.

We now derive the influence function of the M-estimator. Let $\varepsilon > 0$ be fixed. Under the distribution $(1-\varepsilon)F + \varepsilon\delta_y$, the corresponding M-estimator must satisfy

$$\int_{\mathcal{X}} \psi\big(x, T((1-\varepsilon)F + \varepsilon\delta_y)\big)\mathrm{d}\big((1-\varepsilon)F(x) + \varepsilon\delta_y(x)\big) = 0.$$

Differentiating both sides of the preceding equation with respect to ε and using the chain rule yield

$$\int_{\mathcal{X}} \psi\big(x, T((1-\varepsilon)F + \varepsilon\delta_y)\big)\mathrm{d}(\delta_y(x) - F(x)) + \left[\int_{\mathcal{X}} \frac{\partial}{\partial t}\psi(x, t)\Big|_{t=T((1-\varepsilon)F+\varepsilon\delta_y)} \mathrm{d}\big((1-\varepsilon)F(x) + \varepsilon\delta_y(x)\big)\right] \frac{\mathrm{d}}{\mathrm{d}\varepsilon} T\big((1-\varepsilon)F + \varepsilon\delta_y\big) = 0.$$

Evaluating the preceding equation at $\varepsilon = 0$, we obtain

$$\int_{\mathcal{X}} \psi(x, \theta(F)) \mathrm{d}(\delta_y(x) - F(x)) + \left[\int_{\mathcal{X}} \frac{\partial}{\partial t} \psi(x,t) \Big|_{t=T(F)} \mathrm{d}F(x) \right] \left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon} T\big((1-\varepsilon)F + \varepsilon\delta_y\big) \Big|_{\varepsilon=0} \right] = 0.$$

Since $\int_{\mathcal{X}} \psi(x, T(F)) \mathrm{d}F(x) = 0$ by definition and $\int_{\mathcal{X}} \psi(x, T(F)) \mathrm{d}\delta_y(x) = \psi(y, T(F))$, we can simplify the preceding equation as

$$\psi(y, T(F)) + \left[\int_{\mathcal{X}} \frac{\partial}{\partial t} \psi(y,t) \Big|_{t=T(F)} \mathrm{d}F(x) \right] \mathrm{IF}(T, F, x) = 0.$$

Finally, if we assume the $m \times m$ matrix $-\int_{\mathcal{X}} \frac{\partial}{\partial t} \psi(x,t)\big|_{t=T(F)} \mathrm{d}F(x)$ is invertible, we obtain

$$\mathrm{IF}(T, F, y) = \left(- \int_{\mathcal{X}} \frac{\partial}{\partial t} \psi(x,t) \Big|_{t=T(F)} \mathrm{d}F(x) \right)^{-1} \psi(y, \theta(F)).$$

▶

In the influence function approach to robustness, an estimator is regarded to be robust if its *gross-error sensitivity*, $\sup_{y \in \mathcal{X}} \|\mathrm{IF}(T, F, y)\|_2$, is finite. Here, the gross-error sensitivity measures the worst influence which a small amount of contamination can have on the value of the estimator (Hampel et al., 1986).

Let us go back to Examples 5.1 and 5.2 discussed earlier and look their gross-error sensitivities. From (5.4) and (5.5), we have $\sup_{y \in \mathcal{X}} |\mathrm{IF}(T_1, F, y)| = \infty$ and $\sup_{y \in \mathcal{X}} |\mathrm{IF}(T_2, F, y)| = \frac{1}{2f(0)} < \infty$, respectively, from which we conclude that the median is more robust than the mean.

Since being introduced into the world of statistics in the 1960s, the influence function has found a wide range of applications. They can be used to understand the robustness properties of various estimators, to design new estimators with certain robustness properties (Hampel et al., 1986), to identify influential observations in

model fitting (Cook, 1977; Cook and Weisberg, 1982), to perform model validation (Debruyne, Hubert, and Suykens, 2008; Koh and Liang, 2017), and to design efficient subsampling algorithms to reduce the computational load in training machine learning models (Ting and Brochu, 2018; Raj et al., 2020).

5.2 Extension of the Influence Function in Density Estimation Problem

Even though the influence function has been used in various statistical applications, it has not been used much in the density estimation problem.

The main difficulty is that the influence function was traditionally defined for real- or vector-valued statistical functionals. But, the object of primary interest in the density estimation problem is a function, or more precisely, a pdf. We need to extend the definition of the influence function by allowing the statistical functional therein to be function-valued.

We now present our approach. Let T be a map from the collection of distribution functions over $\mathcal{X}$ to the class of log-density functions over $\mathcal{X}$, and $\widetilde{T}$ be a map from the collection of distribution functions over $\mathcal{X}$ to the class of density functions over $\mathcal{X}$; that is, if F is a distribution function over $\mathcal{X}$, $T(F)$ is a log-density function and $\widetilde{T}(F)$ is a density function over $\mathcal{X}$. Then, we define the *influence functions of* $T(F)$ *and* $\widetilde{T}(F)$ *evaluated at* $x \in \mathcal{X}$ to be

$$\mathrm{IF}_x(T, F, y) := \lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon}\Big(T\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - T(F)(x)\Big), \tag{5.6}$$

$$\mathrm{IF}_x(\widetilde{T}, F, y) := \lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon}\Big(\widetilde{T}\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - \widetilde{T}(F)(x)\Big), \tag{5.7}$$

respectively. Note that both T and $\widetilde{T}$ are function-valued statistical functionals, and assign to a distribution over $\mathcal{X}$ a log-density function and density function over $\mathcal{X}$, respectively. In addition, $\mathrm{IF}_x(T, F, y)$ and $\mathrm{IF}_x(\widetilde{T}, F, y)$ are both real-valued as they

only depend on the evaluations of T and $\widetilde{T}$, respectively, at different distribution functions.

In (5.6) (*resp.* (5.7)), $\mathrm{IF}_x(T, F, y)$ (*resp.* $\mathrm{IF}_x(\widetilde{T}, F, y)$) is the Gâteaux derivative of T (*resp.* $\widetilde{T}$) at F in the direction of δ_y evaluated at x, and describes how the value of $T(F)$ (*resp.* $\widetilde{T}(F)$) at x is affected by y. In particular, $\mathrm{IF}_x(T, F, y) > 0$ means the value of the log-density function at x increases with the presence of y, and $\mathrm{IF}_x(T, F, y) < 0$ means the value of the log-density function at x decreases with the presence of y. A similar interpretation can be extended to $\mathrm{IF}_x(\widetilde{T}, F, y)$, simply by replacing "log-density function" with "density function".

The following proposition establishes the relationship between $\mathrm{IF}_x(\widetilde{T}, F, y)$ and $\mathrm{IF}_x(T, F, y)$.

Proposition 5.1. *Suppose* $\widetilde{T}(F)(x) = \exp(T(F)(x))$ *for all* $x \in \mathcal{X}$. *Then, we have*

$$\mathrm{IF}_x(\widetilde{T}, F, y) = \widetilde{T}(F)(x)\mathrm{IF}_x(\widetilde{T}, F, y), \qquad \text{for all } x \in \mathcal{X}. \tag{5.8}$$

The proof of Proposition 5.1 can be found in Section 5.5.1.

From (5.8), we can view $\mathrm{IF}_x(\widetilde{T}, F, y)$ as a weighted version of $\mathrm{IF}_x(T, F, y)$, where the weight is $\widetilde{T}(x)$, the value of the unperturbed density function at x.

In addition, the following proposition establishes the relationship among the KL-divergence, $\mathrm{IF}_x(T, F, y)$, and $\mathrm{IF}_x(\widetilde{T}, F, y)$.

Proposition 5.2. *Suppose* $\widetilde{T}(F)(x) = \exp(T(F)(x))$ *for all* $x \in \mathcal{X}$, $\int_{\mathcal{X}} \widetilde{T}(F)(x)|T((1-\varepsilon)F + \varepsilon\delta_y)(x)|\mathrm{d}x < \infty$ *for all* $\varepsilon \in [0, 1]$, *and* $\int_{\mathcal{X}} |\mathrm{IF}_x(\widetilde{T}, F, y)|\mathrm{d}x < \infty$. *Then, we have*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{KL}\Big(\widetilde{T}(F)\Big\|\widetilde{T}((1-\varepsilon)F + \varepsilon\delta_y)\Big)\Big|_{\varepsilon=0} &= -\int_{\mathcal{X}} \widetilde{T}(F)(x)\mathrm{IF}_x(T, F, y)\mathrm{d}x \\ &= -\int_{\mathcal{X}} \mathrm{IF}_x(\widetilde{T}, F, y)\mathrm{d}x.\end{aligned}$$

We next use a simple example to illustrate the definitions of the influence functions of $T(F)$ and $\widetilde{T}(F)$ evaluated at $x \in \mathcal{X}$.

Example 5.4 (Normal location model). Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{Q}$ contain all pdfs of the form

$$q_\theta(x) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right), \text{ for all } x \in \mathcal{X}, \qquad \theta \in \mathbb{R}.$$

Let $T(F) = \log q_{\theta(F)}$ and $\widetilde{T}(F) = q_{\theta(F)}$, where

$$\theta(F) := \arg\max_{\theta \in \mathbb{R}} \left\{ \int_{\mathcal{X}} \log q_\theta(x) \mathrm{d}F(x) \right\} = \mathbb{E}_F[X],$$

and we assume $\mathbb{E}_F[X]$ exists. If we suppose F has a pdf, say p_0, then $q_{\theta(F)}$ is the ML density projection of p_0 onto the family $\mathcal{Q}$. To see this is a *projection*, suppose p_0 satisfies $\int_{\mathcal{X}} p_0(x)|\log p_0(x)|\mathrm{d}x < \infty$ and $\int_{\mathcal{X}} p_0(x)|\log q_\theta(x)|\mathrm{d}x < \infty$ for all $\theta \in \mathbb{R}$, and notice that $q_{\theta(F)}$ minimizes $\mathrm{KL}(p_0 \| q_\theta)$ over all $q_\theta \in \mathcal{Q}$; that is, $q_{\theta(F)}$ has the smallest KL-divergence to p_0 among all pdfs in $\mathcal{Q}$.

By simple algebra, we have, for all $x \in \mathcal{X}$,

$$\begin{aligned} \mathrm{IF}_x(T, F, y) &= (y - \mathbb{E}_F[X])(x - \mathbb{E}_F[X]), \\ \mathrm{IF}_x(\widetilde{T}, F, y) &= q_{\theta(F)}(x)(y - \mathbb{E}_F[X])(x - \mathbb{E}_F[X]). \end{aligned}$$

If we let $\mathbb{E}_F[X] = 0$ and $y = 2$, $\mathrm{IF}_x(T, F, y)$ and $\mathrm{IF}_x(\widetilde{T}, F, y)$ evaluated at different values of x are shown in Figure 5.1. ▶

If we fix y and F and view $\mathrm{IF}_x(T, F, y)$ and $\mathrm{IF}_x(\widetilde{T}, F, y)$ as functions of the evaluation point x, both can vary with x, which is obvious from Figure 5.1. In other words, with a fixed F, a fixed y can have different effects on each of $T(F)$ and $\widetilde{T}(F)$ at different evaluation points. In order to have a summarizing quantity to describe the maximal possible effect of y on $T(F)$ and $\widetilde{T}(F)$, we use

$$M(T, F, y) := \sup_{x \in \mathcal{X}} |\mathrm{IF}_x(T, F, y)|, \qquad \text{and} \qquad M(\widetilde{T}, F, y) := \sup_{x \in \mathcal{X}} |\mathrm{IF}_x(\widetilde{T}, F, y)|,$$

and call them the *overall influences* of y on $T(F)$ and $\widetilde{T}(F)$, respectively. They describe the maximal possible effects of y on $T(F)$ and $\widetilde{T}(F)$, respectively.

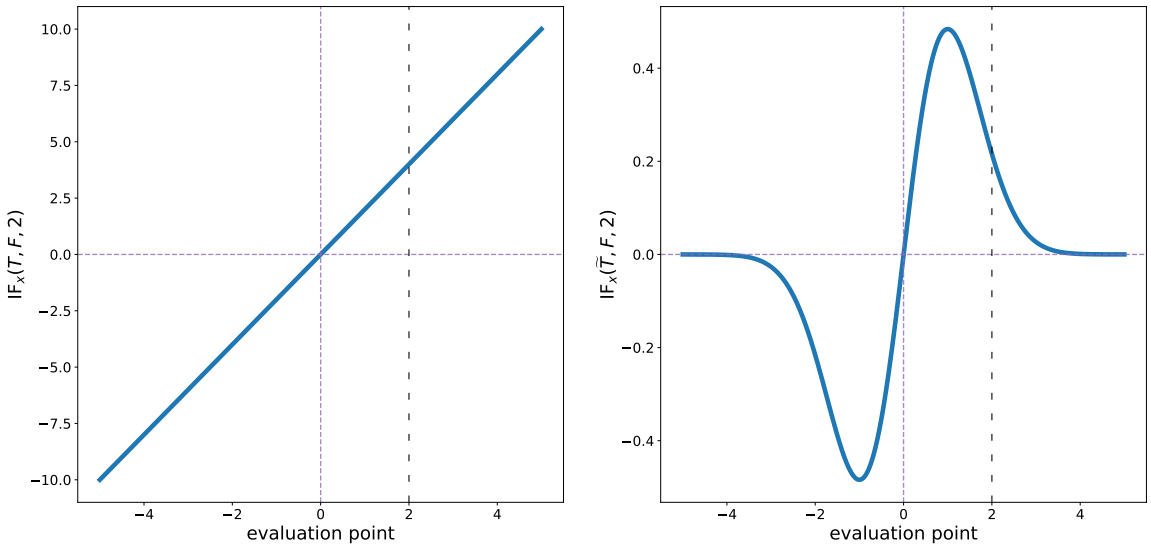

Figure 5.1: $\mathrm{IF}_x(T, F, y)$ (left panel) and $\mathrm{IF}_x(\widetilde{T}, F, y)$ (right panel) evaluated at different $x \in \mathcal{X}$ with $\mathbb{E}_F[X] = 0$ and $y = 2$. The black dashed vertical line indicates the location of the contaminant y.

Example 5.4 (Normal location model, continued). The overall influences of y on $T(F)$ and $\widetilde{T}(F)$ are

$$M(T, F, y) = \begin{cases} 0, & \text{if } y = \mathbb{E}_F[X] \\ \infty, & \text{otherwise} \end{cases}, \quad \text{and} \quad M(\widetilde{T}, F, y) = \frac{1}{\sqrt{2e\pi}}\big|y - \mathbb{E}_F[X]\big|,$$

respectively. ▶

Finally, we define the *sample influence functions of* $T(F)$ *and* $\widetilde{T}(F)$ *evaluated at* $x \in \mathcal{X}$ to be

$$\mathrm{SIF}_{x,\varepsilon}(T, F, y) := \frac{1}{\varepsilon}\Big(T\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - T(F)(x)\Big),$$
$$\mathrm{SIF}_{x,\varepsilon}(\widetilde{T}, F, y) := \frac{1}{\varepsilon}\Big(\widetilde{T}\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - \widetilde{T}(F)(x)\Big),$$

respectively, and the corresponding *sample overall influences* to be

$$\widehat{M}_\varepsilon(T, F, y) := \sup_{x\in\mathcal{X}}|\mathrm{SIF}_{x,\varepsilon}(T, F, y)|, \qquad \text{and} \qquad \widehat{M}_\varepsilon(\widetilde{T}, F, y) := \sup_{x\in\mathcal{X}}|\mathrm{SIF}_{x,\varepsilon}(\widetilde{T}, F, y)|$$

respectively. Similarly to (5.3), $\mathrm{SIF}_{x,\varepsilon}(T,F,y)$ (*resp.* $\mathrm{SIF}_{x,\varepsilon}(\widetilde{T},F,y)$) is the finite-difference approximation of $\mathrm{IF}_x(T,F,y)$ (*resp.* $\mathrm{IF}_{x,\varepsilon}(\widetilde{T},F,y)$).

5.3 Influence Function of (Log-)Density Projection in a Finite-dimensional Exponential Family

With the introduction of $\mathrm{IF}_x(T,F,y)$ and $\mathrm{IF}_x(\widetilde{T},F,y)$ in the preceding section, we focus on the influence functions of the ML and SM (log-)density projections (to be defined later) in an m-dimensional exponential family $\mathcal{Q}_{\mathrm{fin}}$ (introduced in Chapter 2) in this section.

Recall $\mathcal{Q}_{\mathrm{fin}}$ contains all pdfs of the form

$$\tilde{q}_\theta(x) := \mu(x)\exp\big(\langle\theta,\varphi(x)\rangle - B(\theta)\big) \text{ for all } x\in\mathcal{X}, \qquad \theta\in\Theta,$$

where we assume each component of φ, $\varphi_j : \mathcal{X}\to\mathbb{R}$, is twice continuously differentiable. We define the following maps

$$\begin{aligned} T(F) &= \log\tilde{q}_{\theta_{\mathrm{ML},F}}, && \text{and} && S(F) = \log\tilde{q}_{\theta_{\mathrm{SM},F}},\\ \widetilde{T}(F) &= \tilde{q}_{\theta_{\mathrm{ML},F}}, && \text{and} && \widetilde{S}(F) = \tilde{q}_{\theta_{\mathrm{SM},F}}, \end{aligned}$$

where F is a distribution function over $\mathcal{X}$,

$$\begin{aligned} \theta_{\mathrm{ML},F} &:= \operatorname*{arg\,min}_{\theta\in\Theta}\left\{\int_{\mathcal{X}} -\log\tilde{q}_\theta(x)\mathrm{d}F(x)\right\}\\ &= \operatorname*{arg\,min}_{\theta\in\Theta}\left\{B(\theta) - \Big\langle\theta,\mathbb{E}_F[\varphi(X)]\Big\rangle\right\},\\ \theta_{\mathrm{SM},F} &:= \operatorname*{arg\,min}_{\theta\in\Theta}\left\{\int_{\mathcal{X}}\sum_{u=1}^{d}\left(\frac{1}{2}\big(\partial_u\log\tilde{q}_\theta(x)\big)^2 + \partial_u\log\tilde{q}_\theta(x)\right)\mathrm{d}F(x)\right\}\\ &= \operatorname*{arg\,min}_{\theta\in\Theta}\left\{\frac{1}{2}\theta^\top\mathbb{E}_F[D_1(X)D_1(X)^\top]\theta - \theta^\top\mathbb{E}_F[W(X)]\right\}, \end{aligned}$$

D_1 is a map from $\mathcal{X}$ to $\mathbb{R}^{m\times d}$ with $[D_1(x)]_{j,u} = \partial_u\varphi_j(x)$ for all $x\in\mathcal{X}$, D_2 is a map from $\mathcal{X}$ to $\mathbb{R}^{m\times d}$ with $[D_2(x)]_{j,u} = \partial_u^2\varphi_j(x)$ for all $x\in\mathcal{X}$, W is a map from $\mathcal{X}$ to $\mathbb{R}^m$ with

$W(x) = -(D_1(x)\nabla \log \mu(x) + D_2(x)\mathbf{1}_d)$ for all $x \in \mathcal{X}$, and $\mathbf{1}_d := (1, \cdots, 1)^\top \in \mathbb{R}^d$. Supposing that the pdf of F, denoted by p_0, exists, we observe that $\widetilde{T}(F)$ and $\widetilde{S}(F)$ are the ML and SM density projections of p_0 onto the family $\mathcal{Q}_{\text{fin}}$, as they have the smallest KL- and H-divergences to p_0 among all $\tilde{q}_\theta \in \mathcal{Q}_{\text{fin}}$, respectively.

Then, the following theorem provides explicit expressions for $\text{IF}_x(T, F, y)$ and $\text{IF}_x(S, F, y)$. The expressions for $\text{IF}_x(\widetilde{T}, F, y)$ and $\text{IF}_x(\widetilde{S}, F, y)$ are easy to obtain by multiplying $\text{IF}_x(T, F, y)$ and $\text{IF}_x(S, F, y)$ with $\widetilde{T}(F)(x)$ and $\widetilde{S}(F)(x)$, respectively, using Proposition 5.1.

Theorem 5.1 (Influence functions of ML and SM log-density projections in $\mathcal{Q}_{\text{fin}}$ evaluated at $x \in \mathcal{X}$)**.**

(a) Assume $\mathbb{E}_F[\varphi(X)]$ *exists and belongs to* $\text{int}(\Theta)$, $\int_{\mathcal{X}} \tilde{q}_{\theta_{\text{ML},F}}(x)\|\varphi(x)\|_2^2 \text{d}x < \infty$, *and* $\nabla^2 B(\theta_{\text{ML},F})$ *is invertible. Then, for all* $x \in \mathcal{X}$,

$$\text{IF}_x(T, F, y) = \left\langle \widetilde{G}_{\text{ML}}(F, y), \varphi(x) - \int_{\mathcal{X}} \tilde{q}_{\theta_{\text{ML},F}}(w)\varphi(w)\text{d}w \right\rangle, \tag{5.9}$$

where

$$\widetilde{G}_{\text{ML}}(F, y) := \left[\nabla^2 B(\theta_{\text{ML},F})\right]^{-1}\left(\varphi(y) - \mathbb{E}_F[\varphi(X)]\right)$$

(b) Assume $\mathbb{E}_F[W(X)]$ *exists,* $\int_{\mathcal{X}} \tilde{q}_{\theta_{\text{SM},F}}(x)\|\varphi(x)\|_2^2 \text{d}x < \infty$, *and* $\mathbb{E}_F\left[D_1(X)D_1(X)^\top\right]$ *is invertible. Then, for all* $x \in \mathcal{X}$,

$$\text{IF}_x(T, F, y) = \left\langle \widetilde{G}_{\text{SM}}(F, y), \varphi(x) - \int_{\mathcal{X}} \tilde{q}_{\theta_{\text{SM},F}}(w)\varphi(w)\text{d}w \right\rangle, \tag{5.10}$$

where

$$\widetilde{G}_{\text{SM}}(F, y) := \left[\mathbb{E}_F\left[D_1(X)D_1(X)^\top\right]\right]^{-1} \times \left\{W(y) - D_1(y)D_1(y)^\top \left[\mathbb{E}_F\left[D_1(X)D_1(X)^\top\right]\right]^{-1} \mathbb{E}_F[W(X)]\right\}.$$

The proof of Theorem 5.1 will be provided in Section 5.5.3.

From the proof, we can see $\widetilde{G}_{\mathrm{ML}}(F, y)$ is the influence function of the M-estimator $\theta_{\mathrm{ML},F}$ that satisfies

$$0 = \int_{\mathcal{X}} \Big(\nabla B(\theta_{\mathrm{ML},F}) - \varphi(x) \Big) \mathrm{d}F(x),$$

and $\widetilde{G}_{\mathrm{SM}}(F, y)$ is the influence function of the M-estimator $\theta_{\mathrm{SM},F}$ that satisfies

$$0 = \mathbb{E}_F \big[D_1(X) D_1(X)^\top \big] \theta_{\mathrm{SM},F} - \mathbb{E}_F[W(X)];$$

in other words, they are the influence functions of their respective natural parameters.

Comparing (5.9) and (5.10), we see that $\mathrm{IF}_x(T, F, y)$ depends on y *only* through the canonical statistics φ, but $\mathrm{IF}_x(S, F, y)$ depends on y through the first two derivatives of φ and the first derivative of $\log \mu$. This is the key difference between $\mathrm{IF}_x(T, F, y)$ and $\mathrm{IF}_x(S, F, y)$. In the examples we will see later, their differences will become more apparent.

From Theorem 5.1, we can obtain some properties of $\mathrm{IF}_x(T, F, y)$ and $\mathrm{IF}_x(S, F, y)$ which are given in the corollaries below.

Corollary 5.1.

(a) Under the assumptions in Theorem 5.1(a), $\mathrm{IF}_x(T, F, y) = 0$ *for all* $x \in \mathcal{X}$ *if and only if* F *and* y *satisfy*

$$\varphi(y) - \mathbb{E}_F[\varphi(X)] = 0. \tag{5.11}$$

(b) Under the assumptions in Theorem 5.1(b), $\mathrm{IF}_x(S, F, y) = 0$ *for all* $x \in \mathcal{X}$ *if and only if* F *and* y *satisfy*

$$W(y) - D_1(y) D_1(y)^\top \Big[\mathbb{E}_F \big[D_1(X) D_1(X)^\top \big] \Big]^{-1} \mathbb{E}_F[W(X)] = 0. \tag{5.12}$$

Corollary 5.2.

(a) Under the assumptions in Theorem 5.1(a), suppose F and y satisfy $\varphi(y) - \mathbb{E}_F[\varphi(X)] \neq 0$.

(i) If $\sup_{j=1,\cdots,m} \sup_{x\in\mathcal{X}}|\varphi_j(x)| < \infty$, then $M(T,F,y) < \infty$.

(ii) If there exists $j^* \in \{1,\cdots,m\}$ and $x_0 \in \overline{\mathcal{X}}$ such that $\lim_{x\to x_0}|\varphi_{j^*}(x)| = \infty$ so that $\sup_{x\in\mathcal{X}}|\varphi_j(x)| = \infty$, $\varphi_{j^*}(y) - \mathbb{E}_F[\varphi_{j^*}(X)] \neq 0$, and $\lim_{x\to x_0} \frac{\varphi_j(x)}{\varphi_{j^*}(x)} = 0$ for all $j \neq j^*$, then $M(T,F,y) = \infty$.

(b) Under the assumptions in Theorem 5.1(b), suppose F and y satisfy

$$W(y) - D_1(y)D_1(y)^\top \Big[\mathbb{E}_F\big[D_1(X)D_1(X)^\top\big]\Big]^{-1} \mathbb{E}_F[W(X)] \neq 0.$$

(i) If $\sup_{j=1,\cdots,m} \sup_{x\in\mathcal{X}}|\varphi_j(x)| < \infty$, then $M(S,F,y) < \infty$.

(ii) If there exists $j^* \in \{1,\cdots,m\}$ and $x_0 \in \overline{\mathcal{X}}$ such that $\lim_{x\to x_0}|\varphi_{j^*}(x)| = \infty$ so that $\sup_{x\in\mathcal{X}}|\varphi_j(x)| = \infty$, the j^*-th element of

$$W(y) - D_1(y)D_1(y)^\top \Big[\mathbb{E}_F\big[D_1(X)D_1(X)^\top\big]\Big]^{-1} \mathbb{E}_F[W(X)]$$

is nonzero, and $\lim_{x\to x_0} \frac{\varphi_j(x)}{\varphi_{j^*}(x)} = 0$ for all $j \neq j^*$, then $M(S,F,y) = \infty$.

These corollaries are obvious from Theorem 5.1, in particular, from (5.9) and (5.10), and their proofs are omitted here. In the rest of this section, we provide several examples to illustrate Theorem 5.1 and Corollaries 5.1 and 5.2.

Example 5.5 (Normal location-scale model). Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{Q}$ contain all pdfs of the form

$$\tilde{q}_\theta(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\omega)^2}{2\sigma^2}\right) \text{ for all } x \in \mathcal{X}, \qquad \theta := (\omega, \sigma^2) \in \Theta,$$

where $\Theta := \mathbb{R} \times (0, \infty)$. In this example, we have

$$\mu(x) = \frac{1}{\sqrt{2\pi}}, \qquad \varphi(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}, \qquad B(\theta) = \frac{\omega^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2,$$

and

$$D_1(x) = \begin{pmatrix} 1 \\ 2x \end{pmatrix}, \qquad D_2(x) = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \qquad (\log\mu)'(x) = 0, \qquad W(x) = \begin{pmatrix} 0 \\ -2 \end{pmatrix}.$$

Assume $m_1 := \mathbb{E}_F[X]$ and $m_2 := \mathbb{E}_F[X^2]$ both exist and $m_2 - m_1^2 > 0$. Then, using Theorem 5.1, we have

$$\mathrm{IF}_x(T, F, y) = \mathrm{IF}_x(S, F, y) = a_1 x^2 + a_2 x + a_3, \qquad \text{for all } x \in \mathcal{X},$$

where

$$\begin{aligned} a_1 &:= \frac{-m_1}{(m_2 - m_1^2)^2}(y - m_1) + \frac{1}{2(m_2 - m_1^2)^2}(y^2 - m_2), \\ a_2 &:= \frac{m_2 + m_1^2}{(m_2 - m_1^2)^2}(y - m_1) - \frac{m_1}{(m_2 - m_1^2)^2}(y^2 - m_2), \\ a_3 &:= \frac{m_1^3 - 2m_1 m_2}{(m_2 - m_1^2)^2}(y - m_1) + \frac{m_2}{2(m_2 - m_1^2)^2}(y^2 - m_2). \end{aligned}$$

Since, in this example, we have $\varphi_1(x) = x$ and $\varphi_2(x) = x^2$ for all $x \in \mathcal{X}$, and $\sup_{x\in\mathcal{X}} |\varphi_2(x)| = \infty$ with $\lim_{x\to\pm\infty} \varphi_2(x) = \infty$, and $\lim_{x\to\pm\infty} \frac{\varphi_1(x)}{\varphi_2(x)} = 0$, we have $M(T, F, y) = M(S, F, y) = \infty$, which verifies Corollary 5.2. ▶

Example 5.6 (Lognormal location model). Let $\mathcal{X} = (0, \infty)$ and $\mathcal{Q}$ contain all pdfs of the form

$$\tilde{q}_\theta(x) := \frac{1}{x\sqrt{2\pi}} \exp\left(-\frac{(\log x - \theta)^2}{2}\right) \text{ for all } x \in \mathcal{X}, \qquad \theta \in \mathbb{R}.$$

In this example, we have

$$\mu(x) = \frac{1}{x\sqrt{2\pi}} \exp\left(-\frac{(\log x)^2}{2}\right), \qquad \varphi(x) = \log x, \qquad B(\theta) = \frac{\theta^2}{2},$$

and

$$D_1(x) = \frac{1}{x}, \quad D_2(x) = -\frac{1}{x^2}, \quad (\log \mu)'(x) = -\frac{1}{x} - \frac{\log x}{x}, \quad W(x) = \frac{2}{x^2} + \frac{\log x}{x^2}.$$

Then, using Theorem 5.1, we obtain

$$\begin{aligned} \mathrm{IF}_x(T, F, y) &= \big(\log y - m_1\big)\big(\log x - m_1\big), \\ \mathrm{IF}_x(S, F, y) &= \frac{1}{m_3 y^2}\left(\log y - \frac{m_2}{m_3}\right)\left(\log x - 2 - \frac{m_2}{m_3}\right), \end{aligned}$$

for all $x \in \mathcal{X}$, where we assume $m_1 := \mathbb{E}_F[\log X]$, $m_2 := \mathbb{E}_F[X^{-2} \log X]$ and $m_3 := \mathbb{E}_F[X^{-2}]$ exist.

In particular, $\mathrm{IF}_x(T, F, y) = 0$ for all $x \in \mathcal{X}$ if and only if F and y satisfy $\log y = m_1$, which is exactly (5.11); and, $\mathrm{IF}_x(S, F, y) = 0$ for all $x \in \mathcal{X}$ if and only if F and y satisfy $\log y - \frac{m_2}{m_3} = 0$, which is exactly (5.12). These illustrate Corollary 5.1.

Also, note that, in this example, $\sup_{x \in \mathcal{X}} |\varphi(x)| = \sup_{x \in \mathcal{X}} |\log x| = \infty$. Supposing $\log y - \mathbb{E}_F[\log X] \neq 0$, we have $M(T, F, y) = \infty$, which illustrates Corollary 5.2(a); supposing $\log y - \frac{m_2}{m_3} \neq 0$, we have $M(S, F, y) = \infty$, which illustrates Corollary 5.2(b). ▶

Example 5.7 (Gamma rate model). Let $\mathcal{X} = (0, \infty)$ and $\mathcal{Q}$ contain all pdfs of the form

$$\tilde{q}_\theta(x) := \frac{x^{\alpha-1}}{\Gamma(\alpha)} \exp\Big(-\theta x + \alpha \log \theta\Big) \text{ for all } x \in \mathcal{X}, \qquad \theta \in (0, \infty).$$

We assume $\alpha > 1$ is known. In this example

$$\mu(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)}, \qquad \varphi(x) = -x, \qquad B(\theta) = -\alpha \log \theta,$$

and

$$D_1(x) = -1, \quad D_2(x) = 0, \quad (\log \mu)'(x) = \frac{\alpha - 1}{x}, \quad W(x) = \frac{\alpha - 1}{x}.$$

Then, using Theorem 5.1, we obtain

$$\begin{aligned}\mathrm{IF}_x(T, F, y) &= \frac{\alpha}{m_1^2}(y - m_1)(x - m_1),\\ \mathrm{IF}_x(S, F, y) &= (\alpha - 1)(m_2 - y^{-1})\left(x - \frac{\alpha}{\alpha - 1}\frac{1}{m_2}\right),\end{aligned}$$

for all $x \in \mathcal{X}$, where we assume $m_1 := \mathbb{E}_F[X]$ and $m_2 := \mathbb{E}_F[X^{-1}]$ exist.

In particular, $\mathrm{IF}_x(T, F, y) = 0$ for all $x \in \mathcal{X}$ if and only if F and y satisfy $y = m_1$, which is exactly (5.11); and, $\mathrm{IF}_x(S, F, y) = 0$ for all $x \in \mathcal{X}$ if and only if F and y satisfy $y^{-1} - m_2 = 0$, which is exactly (5.12). These illustrate Corollary 5.1.

Note that, in this example, $\sup_{x\in\mathcal{X}}|\varphi(x)| = \sup_{x\in\mathcal{X}}|-x| = \infty$. Supposing $\mathbb{E}_F[X] - y \neq 0$, we have $M(T, F, y) = \infty$, which illustrates Corollary 5.2(a); supposing $m_2 - y^{-1} \neq 0$, we have $M(S, F, y) = \infty$, which illustrates Corollary 5.2(b). ▶

5.4 Influence Function of (Log-)Density Projection in a Kernel Exponential Family

We now turn to the influence functions of the ML and SM (log-)density projections in a kernel exponential family $\mathcal{Q}_{\mathrm{ker}}$ (introduced in Chapter 2) in this section. Recall $\mathcal{Q}_{\mathrm{ker}}$ contains all pdfs of the form

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

We still let F be a distribution function over $\mathcal{X}$, and define the following maps

$$\begin{aligned} T_\lambda(F) &= \log q_{f^{(\lambda)}_{\mathrm{ML},F}}, && \text{and} && S_\rho(F) = \log q_{f^{(\rho)}_{\mathrm{SM},F}},\\ \widetilde{T}_\lambda(F) &= q_{f^{(\lambda)}_{\mathrm{ML},F}}, && \text{and} && \widetilde{S}_\rho(F) = q_{f^{(\rho)}_{\mathrm{SM},F}},\end{aligned}$$

where

$$f^{(\lambda)}_{\mathrm{ML},F} := \operatorname*{arg\,min}_{f\in\mathcal{F}}\left\{A(f) - \int_{\mathcal{X}} f(x)\mathrm{d}F(x) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}\right\}, \tag{5.13}$$

$$f_{\mathrm{SM},F}^{(\rho)} := \arg\min_{f\in\mathcal{F}}\left\{\frac{1}{2}\langle f, C_F f\rangle_{\mathcal{H}} - \langle f, z_F\rangle_{\mathcal{H}} + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2\right\}, \tag{5.14}$$

with $C_F := \int_{\mathcal{X}} \sum_{u=1}^{d} \partial_u k(x,\cdot) \otimes \partial_u k(x,\cdot) \mathrm{d}F(x)$ mapping from $\mathcal{H}$ to $\mathcal{H}$, and $z_F := -\int_{\mathcal{X}} \sum_{u=1}^{d} (\partial_u \log \mu(x) \partial_u k(x,\cdot) + \partial_u^2 k(x,\cdot)) \mathrm{d}F(x) \in \mathcal{H}$. If the distribution function F has a pdf p_0, the resulting C_F is (3.1) defined in Chapter 3; if $F = F_n$, the empirical distribution function of data $X_1, \cdots, X_n$, the resulting C_{F_n} is $\widehat{C}$ defined in (2.14) in Chapter 2.

Note that both objective functionals in (5.13) and (5.14) are strongly convex with constants $\lambda > 0$ and $\rho > 0$, respectively. It follows that each of $f_{\mathrm{ML},F}^{(\lambda)}$ and $f_{\mathrm{SM},F}^{(\rho)}$ exists and is unique (Corollary 11.17 in Bauschke and Combettes, 2011).

We then have the following results on the influence functions of the penalized ML and SM log-density projections onto $\mathcal{Q}_{\mathrm{ker}}$ evaluated at $x \in \mathcal{X}$.

Theorem 5.2 (Influence functions of ML and SM log-density projections in $\mathcal{Q}_{\mathrm{ker}}$ evaluated at $x \in \mathcal{X}$)**.**

(a) Under ***(A1)*** *-* ***(A4)*** *in Chapter 2, we have, for all* $x \in \mathcal{X}$*,*

$$\mathrm{IF}_x(T_\lambda, F, y) = \left\langle G_{\mathrm{ML}}(F,y), k(x,\cdot) - \int_{\mathcal{X}} k(w,\cdot) q_{f_{\mathrm{ML},F}^{(\lambda)}}(w)\mathrm{d}w \right\rangle_{\mathcal{H}}, \tag{5.15}$$

where

$$G_{\mathrm{ML}}(F,y) := \left\{ \mathbb{E}_{q_{f_{\mathrm{ML},F}^{(\lambda)}}}\left[\left(k(X,\cdot) - \Upsilon\right) \otimes \left(k(X,\cdot) - \Upsilon\right)\right] + \lambda I \right\}^{-1} \left(k(y,\cdot) - \int_{\mathcal{X}} k(x,\cdot)\mathrm{d}F(x) \right) \in \mathcal{H},$$

and $\Upsilon := \int_{\mathcal{X}} k(x,\cdot) q_{f_{\mathrm{ML},F}^{(\lambda)}}(x)\mathrm{d}x \in \mathcal{H}$*.*

(b) Under ***(A1)*** *-* ***(A4)*** *in Chapter 2 and* ***(B1)*** *-* ***(B5)*** *in Chapter 3, we have, for all* $x \in \mathcal{X}$*,*

$$\mathrm{IF}_x(S_\rho, F, y) = \left\langle G_{\mathrm{SM}}(F,y), k(x,\cdot) - \int_{\mathcal{X}} k(w,\cdot) q_{f_{\mathrm{SM},F}^{(\rho)}}(w)\mathrm{d}w \right\rangle_{\mathcal{H}}, \tag{5.16}$$

where

$$G_{\mathrm{SM}}(F, y) := (C_F + \rho I)^{-1}\Big(z_{\delta_y} - (C_{\delta_y} + \rho I)(C_F + \rho I)^{-1} z_F\Big) \in \mathcal{H}.$$

The proof of Theorem 5.2 can be found in Section 5.5.4.

Comparing Theorems 5.1 and 5.2, we see they share many similarities. In particular, $\mathrm{IF}_x(T_\lambda, F, y)$ depends on y through k *only*, but $\mathrm{IF}_x(S_\rho, F, y)$ depends on y through the first two partial derivatives of k and the first derivative of $\log \mu$. Moreover, $G_{\mathrm{ML}}(F, y)$ is the influence function of the M-estimator $f^{(\lambda)}_{\mathrm{ML},F}$ that satisfies

$$0 = \int_{\mathcal{X}} \Big(\nabla A(f^{(\lambda)}_{\mathrm{ML},F}) - k(x, \cdot) + \lambda f^{(\lambda)}_{\mathrm{ML},F}\Big) \mathrm{d}F(x),$$

and $G_{\mathrm{SM}}(F, y)$ is the influence function of the M-estimator $f^{(\rho)}_{\mathrm{SM},F}$ that satisfies

$$0 = C_F f^{(\rho)}_{\mathrm{SM},F} - z_F + \rho f^{(\rho)}_{\mathrm{ML},F}.$$

The covariance operator

$$\mathbb{E}_{q_{f^{(\lambda)}_{\mathrm{ML},F}}}\Big[\big(k(X, \cdot) - \Upsilon\big) \otimes \big(k(X, \cdot) - \Upsilon\big)\Big]$$

appearing in $G_{\mathrm{ML}}(F, y)$ plays the role of $\nabla^2 B(\theta_{\mathrm{ML},F})$ in $\widetilde{G}_{\mathrm{ML}}(F, y)$, where $\nabla^2 B(\theta_{\mathrm{ML},F})$ is the covariance matrix of φ under $\tilde{q}_{\theta_{\mathrm{ML},F}}$. The operator C_F appearing in $G_{\mathrm{SM}}(F, y)$ plays the role of $\mathbb{E}_F[D_1(X)D_1(X)^\top]$ in $\widetilde{G}_{\mathrm{SM}}(F, y)$, and z_F in $G_{\mathrm{SM}}(F, y)$ plays the role of $\mathbb{E}_F[W(X)]$ in $\widetilde{G}_{\mathrm{SM}}(F, y)$.

Even though Theorem 5.2 provides explicit expressions of $\mathrm{IF}_x(T_\lambda, F, y)$ and $\mathrm{IF}_x(S_\rho, F, y)$, they are hard to work with directly to compare the sensitivities of the penalized ML and SM density estimators. We will work with the sample influence function in the next chapter and compare their sensitivities numerically.

5.5 Proofs

5.5.1 Proof of Proposition 5.1

Proof of Proposition 5.1. Under the assumption in the proposition and with an application of the chain rule, we have

$$\begin{aligned}
\mathrm{IF}_x(\widetilde{T}, F, y) &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big(\widetilde{T}\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x)\Big)\Big|_{\varepsilon=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big(\exp\big(T((1-\varepsilon)F + \varepsilon\delta_y)(x)\big)\Big)\Big|_{\varepsilon=0} \\
&= \exp\big(T(F)(x)\big)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big(T\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x)\Big)\Big|_{\varepsilon=0} \\
&= \widetilde{T}(F)(x)\mathrm{IF}_x(T, F, y).
\end{aligned}$$

■

5.5.2 Proof of Proposition 5.2

Proof of Proposition 5.2. Let $\varepsilon \in (0, 1)$. By the definition of the KL-divergence, we have

$$\mathrm{KL}\Big(\widetilde{T}(F)\Big\|\widetilde{T}\big((1-\varepsilon)F + \varepsilon\delta_y\big)\Big) = -\int_{\mathcal{X}} \widetilde{T}(F)(x)\Big(T\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - T(F)(x)\Big)\mathrm{d}x.$$

Differentiating both sides wrt ε, interchanging differentiation and integral (which is allowed by the assumptions), and evaluating at $\varepsilon = 0$, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{KL}\Big(\widetilde{T}(F)\Big\|\widetilde{T}\big((1-\varepsilon)F + \varepsilon\delta_y\big)\Big)\Big|_{\varepsilon=0} \\
&= -\int_{\mathcal{X}} \widetilde{T}(F)(x)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big(T\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - T(F)(x)\Big)\Big|_{\varepsilon=0}\mathrm{d}x \\
&= -\int_{\mathcal{X}} \widetilde{T}(F)(x)\mathrm{IF}_x(T, F, y)\mathrm{d}x \\
&= -\int_{\mathcal{X}} \mathrm{IF}_x(\widetilde{T}, F, y)\mathrm{d}x,
\end{aligned}$$

where the last equality follows from Proposition 5.1. ■

5.5.3 Proof of Theorem 5.1

Proof of Theorem 5.1. For simplicity, we let $F_\varepsilon := (1-\varepsilon)F + \varepsilon\delta_y$ for $\varepsilon \in (0,1)$.

(a) First note

$$\frac{1}{\varepsilon}\Big(T(F_\varepsilon)(x) - T(F)(x)\Big) = \left\langle \frac{1}{\varepsilon}\big(\theta_{\mathrm{ML},F_\varepsilon} - \theta_{\mathrm{ML},F}\big), \varphi(x)\right\rangle - \frac{1}{\varepsilon}\Big(B(\theta_{\mathrm{ML},F_\varepsilon}) - B(\theta_{\mathrm{ML},F})\Big).$$

If we let $\varepsilon \to 0^+$ on both sides and use the chain rule, we have

$$\begin{aligned}
&\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\Big(T(F_\varepsilon)(x) - T(F)(x)\Big) \\
&= \left\langle \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\big(\theta_{\mathrm{ML},F_\varepsilon} - \theta_{\mathrm{ML},F}\big), \varphi(x)\right\rangle - \int_{\mathcal{X}} \tilde{q}_{\theta_{\mathrm{ML},F}}(w)\left\langle \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\big(\theta_{\mathrm{ML},F_\varepsilon} - \theta_{\mathrm{ML},F}\big), \varphi(w)\right\rangle \mathrm{d}w,
\end{aligned}$$

which exists if the limit $\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}(\theta_{\mathrm{ML},F_\varepsilon} - \theta_{\mathrm{ML},F})$ exists. The desired result follows if we can show

$$\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\big(\theta_{\mathrm{ML},F_\varepsilon} - \theta_{\mathrm{ML},F}\big) = \big[\nabla^2 B(\theta_{\mathrm{ML},F})\big]^{-1}\big(\varphi(y) - \mathbb{E}_F[\varphi(X)]\big). \tag{5.17}$$

Note the LHS of (5.17) is the influence function of the M-estimator defined by

$$0 = \int_{\mathcal{X}} \Big(\nabla B(\theta_{\mathrm{ML},F}) - \varphi(x)\Big)\mathrm{d}F(x).$$

Using the result in Example 5.3, we can see (5.17) follows, which completes the proof.

(b) Similar to (a), we have

$$\begin{aligned}
&\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\Big(S(F_\varepsilon)(x) - S(F)(x)\Big) \\
&= \left\langle \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\big(\theta_{\mathrm{SM},F_\varepsilon} - \theta_{\mathrm{SM},F}\big), \varphi(x)\right\rangle - \int_{\mathcal{X}} \tilde{q}_{\theta_{\mathrm{SM},F}}(w)\left\langle \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\big(\theta_{\mathrm{SM},F_\varepsilon} - \theta_{\mathrm{SM},F}\big), \varphi(w)\right\rangle \mathrm{d}w,
\end{aligned}$$

which exists if the limit $\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}(\theta_{\mathrm{SM},F_\varepsilon} - \theta_{\mathrm{SM},F})$ exists. The desired result follows if we can show

$$\lim_{\varepsilon\to 0^+} \big(\theta_{\mathrm{SM},F_\varepsilon} - \theta_{\mathrm{SM},F}\big) = \Big[\mathbb{E}_F\big[D_1(X)D_1(X)^\top\big]\Big]^{-1} \times$$

$$\left\{W(y) - D_1(y)D_1(y)^\top \left[\mathbb{E}_F\left[D_1(X)D_1(X)^\top\right]\right]^{-1} \mathbb{E}_F[W(X)]\right\}. \tag{5.18}$$

Note the LHS of (5.18) is the influence function of the M-estimator defined by

$$0 = \mathbb{E}_F\left[D_1(X)D_1(X)^\top\right]\theta_{\mathrm{SM},F} - \mathbb{E}_F[W(X)].$$

Using the result in Example 5.3, we can see (5.18) follows, which completes the proof.

■

5.5.4 Proof of Theorem 5.2

Proof of Theorem 5.2. We let $F_\varepsilon := (1-\varepsilon)F + \varepsilon\delta_y$ for some $\varepsilon \in (0,1)$ throughout this proof.

(a) Using the chain rule, we obtain

$$\mathrm{IF}_x(T_\lambda, F, y) = \left\langle \frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\Big|_{\varepsilon=0}, k(x,\,\cdot\,) - \int_{\mathcal{X}} k(w,\,\cdot\,) q_{f^{(\lambda)}_{\mathrm{ML},F}}(w)\mathrm{d}w \right\rangle_{\mathcal{H}}.$$

What remains to show is $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\big|_{\varepsilon=0} = G_{\mathrm{ML}}(F,y)$.

Observe that $f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}$ is an M-estimator and must satisfy

$$0 = \nabla A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}) - \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}F_\varepsilon(x) + \lambda f^{(\lambda)}_{\mathrm{ML},F_\varepsilon},$$

which is equivalent to the following

$$0 = \int_{\mathcal{X}} \mu(x)\exp\big(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\big)k(x,\,\cdot\,)\mathrm{d}x - \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}F_\varepsilon(x) + \lambda f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}.$$

Differentiating both sides of the preceding equation with respect to ε yields

$$\begin{aligned} 0 = \int_{\mathcal{X}} \mu(x)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left[\exp\big(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\big)\right]k(x,\,\cdot\,)\mathrm{d}x \\ - \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}(-F(x) + \delta_y(x)) + \lambda\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}. \end{aligned} \tag{5.19}$$

We next work out $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left[\exp\big(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\big)\right]$ part. By the chain rule, we have

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left[\exp\big(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\big)\right]$$

$$= \exp\big(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\big)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big[f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\Big]$$
$$= \exp\big(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\big)\left[\left\langle \frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}, k(x,\,\cdot\,) - \int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(v)k(v,\,\cdot\,)\mathrm{d}v\right\rangle_{\mathcal{H}}\right].$$

Plugging the preceding equation back to (5.19), we have

$$\begin{aligned}
0 = &\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(x)\left\langle \frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}, k(x,\,\cdot\,)\right\rangle_{\mathcal{H}} k(x,\,\cdot\,)\mathrm{d}x \\
&\quad - \left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(x)k(x,\,\cdot\,)\mathrm{d}x\right)\left\langle \frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}, \int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(v)k(v,\,\cdot\,)\mathrm{d}v\right\rangle_{\mathcal{H}} \\
&\qquad - \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}(-F(x) + \delta_y(x)) + \lambda\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon} \\
= &\left[\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(x)\big(k(x,\,\cdot\,)\otimes k(x,\,\cdot\,)\big)\mathrm{d}x\right]\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\right) \\
&\quad - \left[\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(x)k(x,\,\cdot\,)\mathrm{d}x\right)\otimes\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(v)k(v,\,\cdot\,)\mathrm{d}v\right)\right]\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\right) \\
&\qquad - \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}(-F(x) + \delta_y(x)) + \lambda\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon} \\
= &\left\{\left[\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(x)\big(k(x,\,\cdot\,)\otimes k(x,\,\cdot\,)\big)\mathrm{d}x\right]\right. \\
&\quad \left. - \left[\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(x)k(x,\,\cdot\,)\mathrm{d}x\right)\otimes\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(v)k(v,\,\cdot\,)\mathrm{d}v\right)\right] + \lambda I\right\}\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\Big|_{\varepsilon=0}\right) \\
&\qquad + \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}F(x) - k(y,\,\cdot\,).
\end{aligned}$$

Evaluating at $\varepsilon = 0$ and rearranging terms, we obtain

$$\begin{aligned}
&\left\{\left[\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F}}(x)\big(k(x,\,\cdot\,)\otimes k(x,\,\cdot\,)\big)\mathrm{d}x\right]\right. \\
&\quad \left. - \left[\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F}}(x)k(x,\,\cdot\,)\mathrm{d}x\right)\otimes\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F}}(v)k(v,\,\cdot\,)\mathrm{d}v\right)\right] + \lambda I\right\}\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\Big|_{\varepsilon=0}\right) \\
&\qquad = k(y,\,\cdot\,) - \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}F(x).
\end{aligned}$$

Note that

$$\left[\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F}}(x)\big(k(x,\,\cdot\,)\otimes k(x,\,\cdot\,)\big)\mathrm{d}x\right]$$

$$
\begin{aligned}
&\quad - \left[\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F}}(x)k(x,\,\cdot\,)\mathrm{d}x\right) \otimes \left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F}}(v)k(v,\,\cdot\,)\mathrm{d}v\right)\right] \\
&= \mathbb{E}_{q_{f^{(\lambda)}_{\mathrm{ML},F}}}\big[k(X,\,\cdot\,)\otimes k(X,\,\cdot\,)\big] - \Upsilon\otimes\Upsilon \\
&= \mathbb{E}_{q_{f^{(\lambda)}_{\mathrm{ML},F}}}\big[(k(X,\,\cdot\,)-\Upsilon)\otimes(k(X,\,\cdot\,)-\Upsilon)\big],
\end{aligned}
$$

which is the covariance operator we have discussed in Chapter 2, and is positive semi-definite. Then, with $\lambda > 0$, the operator

$$
\mathbb{E}_{q_{f^{(\lambda)}_{\mathrm{ML},F}}}\big[(k(X,\,\cdot\,)-\Upsilon)\otimes(k(X,\,\cdot\,)-\Upsilon)\big] + \lambda I
$$

is invertible, and $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\big|_{\varepsilon=0}$ is equal to $G_{\mathrm{ML}}(F,y)$ given in the theorem, which completes the proof.

(b) Similar to (a), using the chain rule, we obtain

$$
\mathrm{IF}_x(S_\rho,F,y) = \left\langle \frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon}\bigg|_{\varepsilon=0}, k(x,\,\cdot\,) - \int_{\mathcal{X}} k(w,\,\cdot\,) q_{f^{(\rho)}_{\mathrm{SM},F}}(w)\mathrm{d}w \right\rangle_{\mathcal{H}}.
$$

It remains to show $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon}\big|_{\varepsilon=0}$ is equal to $G_{\mathrm{SM}}(F,y)$.

Then, observe that $f^{\rho}_{\mathrm{SM},F_\varepsilon}$ is an M-estimator and must satisfy

$$
0 = C_{F_\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon} - z_{F_\varepsilon} + \rho f^{(\rho)}_{\mathrm{SM},F_\varepsilon},
$$

which is equivalent to

$$
0 = \big((1-\varepsilon)C_F + \varepsilon C_{\delta_y}\big) f^{(\rho)}_{\mathrm{SM},F_\varepsilon} - \big((1-\varepsilon)z_F + \varepsilon z_{\delta_y}\big) + \rho f^{(\rho)}_{\mathrm{SM},F_\varepsilon}.
$$

Now, differentiating both sides of the preceding equation with respect to ε yields

$$
0 = \big(C_{\delta_y} - C_F\big) f^{(\rho)}_{\mathrm{SM},F_\varepsilon} + \big((1-\varepsilon)C_F + \varepsilon C_{\delta_y}\big)\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon}\right) - \big(z_{\delta_y} - z_F\big) + \rho\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon}\right).
$$

Letting $\varepsilon \to 0^+$ on both sides, we obtain

$$
0 = \big(C_{\delta_y} - C_F\big) f^{(\rho)}_{\mathrm{SM},F} + C_F\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon}\bigg|_{\varepsilon=0}\right) - \big(z_{\delta_y} - z_F\big) + \rho\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon}\bigg|_{\varepsilon=0}\right).
$$

In addition, since $f_{\mathrm{SM},F}^{(\rho)}$ satisfies $C_F f_{\mathrm{SM},F}^{(\rho)} - z_F + \rho f_{\mathrm{SM},F}^{(\rho)} = 0$, we can use it to simplify the preceding equation as

$$0 = C_{\delta_y} f_{\mathrm{SM},F}^{(\rho)} + C_F \left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f_{\mathrm{SM},F_\varepsilon}^{(\rho)} \bigg|_{\varepsilon=0} \right) - z_{\delta_y} + \rho \left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f_{\mathrm{SM},F_\varepsilon}^{(\rho)} \bigg|_{\varepsilon=0} \right) + \rho f_{\mathrm{SM},F}^{(\rho)}. \tag{5.20}$$

Rearranging terms and using $f_{\mathrm{SM},F}^{(\rho)} = (C_F + \rho I)^{-1} z_F$ yield the desired result. ■

Chapter 6: Numerical Studies of the Sensitivities of Penalized ML and SM (Log-)Density Estimators in $\mathcal{Q}_{\text{ker}}$

As we have seen from Theorem 5.2 in Chapter 5 that, even though the expressions of the influence functions of the penalized ML and SM (log-)density estimators in $\mathcal{Q}_{\text{ker}}$ exist, they are hard to be directly used to compare the sensitivities of these (log-)density estimators. Instead, we are going to compare their sensitivities numerically, which will be the focus of Section 6.1. Since we have seen the penalized and early stopping SM density estimators are qualitatively very similar through numerical examples in Chapters 3 and 4, we only consider the penalized SM density estimator in this chapter. Since one of the most popular methods of selecting the penalty parameter in the penalized SM density estimation approach is the K-fold cross-validation, we turn to studying the sensitivity of the cross-validated penalized SM density estimator in Section 6.2. Since we have both penalized ML and SM density estimators, we discuss in Section 6.3 which density estimator we should use and how to use it.

6.1 Comparison of the Sensitivities of Penalized ML and SM Density Estimators

We study the sensitivities of the penalized ML and SM density estimators numerically in this section.

The tool we use is the sample influence function defined in Chapter 5. We discuss how to compute the sample influence function of a log-density estimator and that of a density estimator in Section 6.1.1. Since we can use either the sample influence function of the log-density estimator or that of the density estimator, we compare them in Section 6.1.2 and show that the former one is a better choice for us. The main comparison of the sensitivities of penalized ML and SM density estimators will be given in Section 6.1.3.

In order to achieve the desired goal, we still define the following maps as we have done in Section 5.4 in Chapter 5

$$T_\lambda(F) = \log q_{f^{(\lambda)}_{\mathrm{ML},F}}, \qquad \text{and} \qquad S_\rho(F) = \log q_{f^{(\rho)}_{\mathrm{SM},F}},$$
$$\widetilde{T}_\lambda(F) = q_{f^{(\lambda)}_{\mathrm{ML},F}}, \qquad \text{and} \qquad \widetilde{S}_\rho(F) = q_{f^{(\rho)}_{\mathrm{SM},F}},$$

where

$$f^{(\lambda)}_{\mathrm{ML},F} := \operatorname*{arg\,min}_{f\in\mathcal{F}} \left\{ A(f) - \int_{\mathcal{X}} f(x)\mathrm{d}F(x) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}} \right\}, \tag{6.1}$$
$$f^{(\rho)}_{\mathrm{SM},F} := \operatorname*{arg\,min}_{f\in\mathcal{F}} \left\{ \frac{1}{2}\langle f, C_F f\rangle_{\mathcal{H}} - \langle f, z_F\rangle_{\mathcal{H}} + \frac{\rho}{2}\|f\|^2_{\mathcal{H}} \right\}, \tag{6.2}$$

with $C_F := \int_{\mathcal{X}} \sum_{u=1}^{d} \partial_u k(x, \cdot) \otimes \partial_u k(x, \cdot)\mathrm{d}F(x)$ mapping from $\mathcal{H}$ to $\mathcal{H}$, and $z_F := -\int_{\mathcal{X}} \sum_{u=1}^{d} (\partial_u^2 k(x, \cdot) + \partial_u \log \mu(x) \partial_u k(x, \cdot))\mathrm{d}F(x) \in \mathcal{H}$.

In order to ensure the comparability of the penalized ML and SM density estimates, we minimize the objective functionals in (6.1) and (6.2) over the same finite-dimensional approximating subspace of $\mathcal{H}$ that is found by the procedures described in Chapter 4, and denote this finalized subspace by $\widetilde{\mathcal{H}}$.

6.1.1 Computation of the Sample Influence Function

We describe how to compute the sample influence function in practice.

We choose ε in the definition of the sample influence function to be $\varepsilon_0 := \frac{1}{n+1}$. Then, $\mathrm{SIF}_{x,\varepsilon_0}(T_\lambda, F_n, y)$ and $\mathrm{SIF}_{x,\varepsilon_0}(S_\rho, F_n, y)$ assess the sensitivity of the penalized ML and SM log-density estimators evaluated at x to the additional observation y, respectively; and, similarly, $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{T}_\lambda, F_n, y)$ and $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{S}_\rho, F_n, y)$ assess the sensitivity of the penalized ML and SM density estimators evaluated at x to the additional observation y, respectively.

Algorithm 6.1 describes how to compute $\mathrm{SIF}_{x,\varepsilon_0}(T_\lambda, F_n, y)$; if one would like to compute $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{T}_\lambda, F_n, y)$, simply do not take the logarithm in Steps 2 and 3.

Algorithm 6.1 Computation of $\mathrm{SIF}_{x,\varepsilon_0}(T_\lambda, F_n, y)$

Require:

- $X_1, \cdots, X_n$, data;
- $y \in \mathcal{X}$, contaminant;
- $\lambda > 0$, penalty parameter;
- $\widetilde{\mathcal{H}}$, the finite-dimensional approximating subspace over which we minimize the penalized NLL loss functional;
- $\{x_\ell\}_{\ell=1}^L \subset \mathcal{X}$, a set of evaluation points.

1: Compute $f^{(\lambda)}_{\mathrm{ML},F_n}$ and $f^{(\lambda)}_{\mathrm{ML},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}$ using Algorithm 4.3 in Chapter 4;
2: Compute $\log q_{f^{(\lambda)}_{\mathrm{ML},F_n}}(x_\ell)$ and $\log q_{f^{(\lambda)}_{\mathrm{ML},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}}(x_\ell)$ for all $\ell = 1, \cdots, L$;
3: Compute

$$\left(\log q_{f^{(\lambda)}_{\mathrm{ML},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}}(x_\ell) - \log q_{f^{(\lambda)}_{\mathrm{ML},F_n}}(x_\ell)\right) \times (n+1), \qquad \text{for all } \ell = 1, \cdots, L;$$

4: **return** the results from Step 3.

Similarly, Algorithm 6.2 describes how to compute $\mathrm{SIF}_{x,\varepsilon_0}(S_\rho, F_n, y)$; if one would like to compute $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{S}_\rho, F_n, y)$, do not take the logarithm in Steps 2 and 3.

Algorithm 6.2 Computation of $\mathrm{SIF}_{x,\varepsilon_0}(S_\rho, F_n, y)$

Require:

- $X_1, \cdots, X_n$, data;
- $y \in \mathcal{X}$, contaminant;
- $\rho > 0$, penalty parameter;
- $\widetilde{\mathcal{H}}$, the finite-dimensional approximating subspace over which we minimize the penalized SM loss functional;
- $\{x_\ell\}_{\ell=1}^L \subset \mathcal{X}$, a set of evaluation points.

1: Compute $f^{(\rho)}_{\mathrm{SM},F_n}$ and $f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}$ using Proposition 4.3 in Chapter 4;
2: Compute $\log q_{f^{(\rho)}_{\mathrm{SM},F_n}}(x_\ell)$ and $\log q_{f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}}(x_\ell)$ for all $\ell = 1, \cdots, L$;
3: Compute

$$\left(\log q_{f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}}(x_\ell) - \log q_{f^{(\rho)}_{\mathrm{SM},F_n}}(x_\ell)\right) \times (n+1), \qquad \text{for all } \ell = 1, \cdots, L;$$

4: **return** the results from Step 3.

6.1.2 Comparison of the Sample Influence Functions of Log-density and Density Estimators

Our goal of the current section is to show, between the sample influence function of the log-density estimator and that of the density estimator in $\mathcal{Q}_{\mathrm{ker}}$, the former is the better choice for us.

We first use numerical examples to demonstrate this. We still use the `waiting` variable in the Old Faithful Geyser dataset but remove the original isolated observation 108 therein. This is to avoid the interaction between 108 and the additional observation y. We use the same $\mathcal{X}$, k, μ, and $\widetilde{\mathcal{H}}$ as those in Chapter 4.

We focus on the penalized SM density estimator for now. Fix $\rho = e^{-11}$. Figure 6.1 shows the penalized SM (log-)density estimates with and without $y = 120$ and the corresponding sample influence functions, and Figure 6.2 shows the penalized SM (log-)density estimates with and without $y = 180$ and the corresponding sample

influence functions. Recall that we choose μ to be the pdf of the Gamma distribution with the shape and scale parameters to be 36 and 2, respectively, and $\mu(x) \to 0$ and $\log \mu(x) \to -\infty$ as $x \to \infty$. Then, comparing Panel [D] in Figure 6.1 and that in Figure 6.2, we see that, when y is large, the spike in the penalized SM density estimate may disappear due to this particular choice of μ. The sample influence function of the penalized SM density estimator may fail to capture the bump or the spike in the penalized SM density estimate when an isolated observation is present (see Panel [F] in Figure 6.2). In other words, understanding the sensitivity of the penalized SM density estimator via the sample influence function of the density estimator can be misleading.

However, the issue with the sample influence function of the density estimator described above does *not* occur to the sample influence function of the log-density estimator. Comparing Panel [C] in Figure 6.1 and that in Figure 6.2, we see the sample influence function of the penalized SM log-density estimator succeeds in capturing the spike at the isolated observation, no matter how far y is from the bulk of the data. Numerical examples of the penalized ML density estimator also confirm this; see Figures 6.3 and Figure 6.4 for evidence.

Analytically, note that

$$
\begin{aligned}
& S_\rho((1-\varepsilon_0)F_n + \varepsilon_0\delta_y)(x) - S_\rho(F_n)(x) \\
&= \big[\cancel{\log \mu(x)} + f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}(x) - A(f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y})\big] \\
&\qquad - \big[\cancel{\log \mu(x)} + f^{(\rho)}_{\mathrm{SM},F_n}(x) - A(f^{(\rho)}_{\mathrm{SM},F_n})\big] \\
&= \big[f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}(x) - A(f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y})\big] - \big[f^{(\rho)}_{\mathrm{SM},F_n}(x) - A(f^{(\rho)}_{\mathrm{SM},F_n})\big],
\end{aligned}
$$

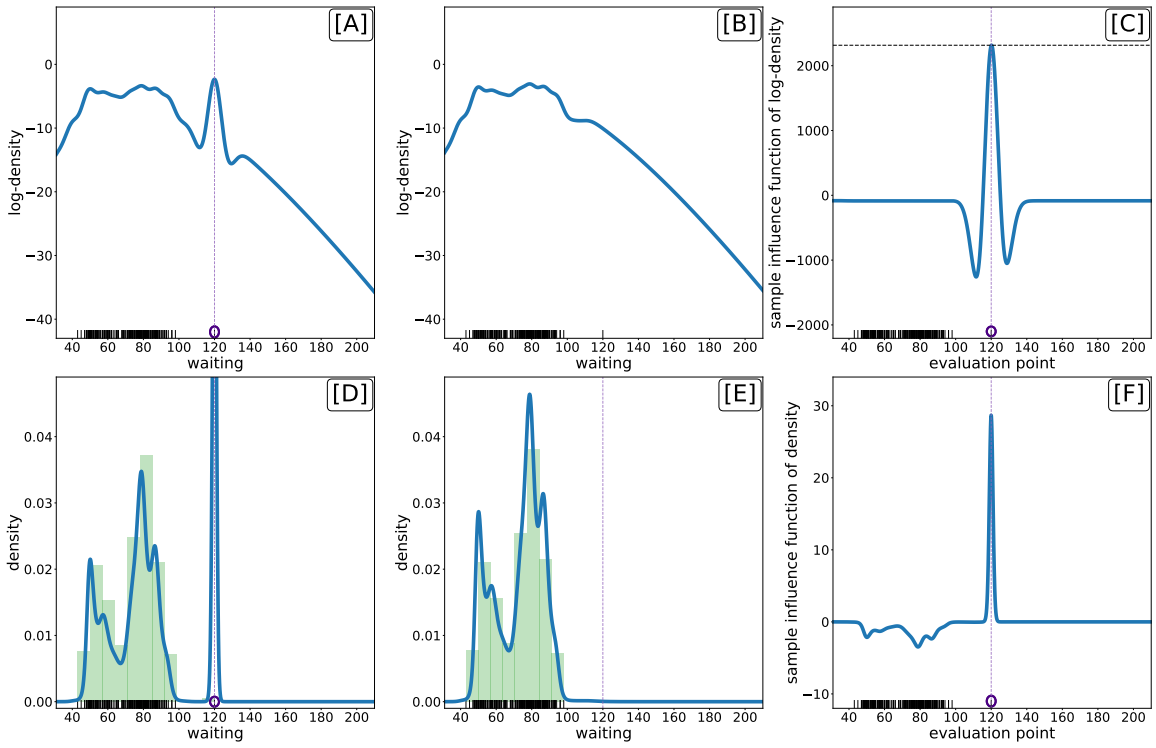

Figure 6.1: Fix $\rho = e^{-11}$. Panels [A] and [B] show the penalized SM log-density estimates with and without the additional observation $y = 120$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 120$. Panel [F] shows the sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 120$.

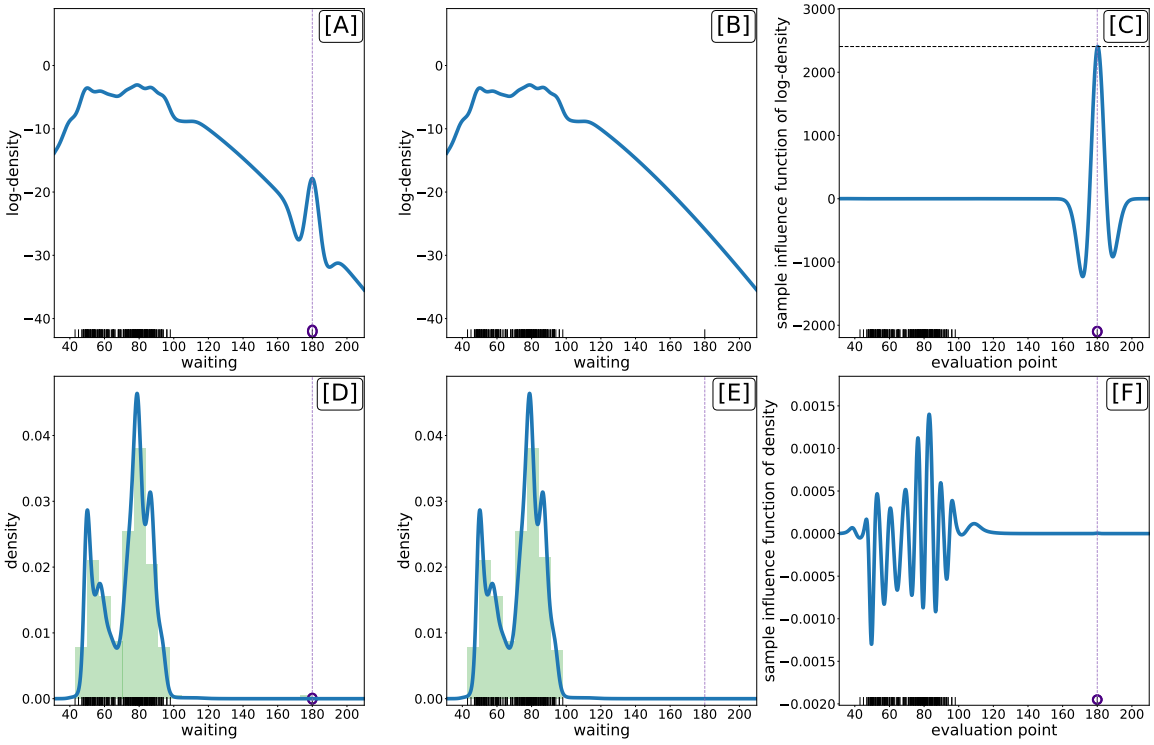

Figure 6.2: Fix $\rho = e^{-11}$. Panels [A] and [B] show the penalized SM log-density estimates with and without the additional observation $y = 180$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 180$. Panel [F] shows the resulting sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 180$.

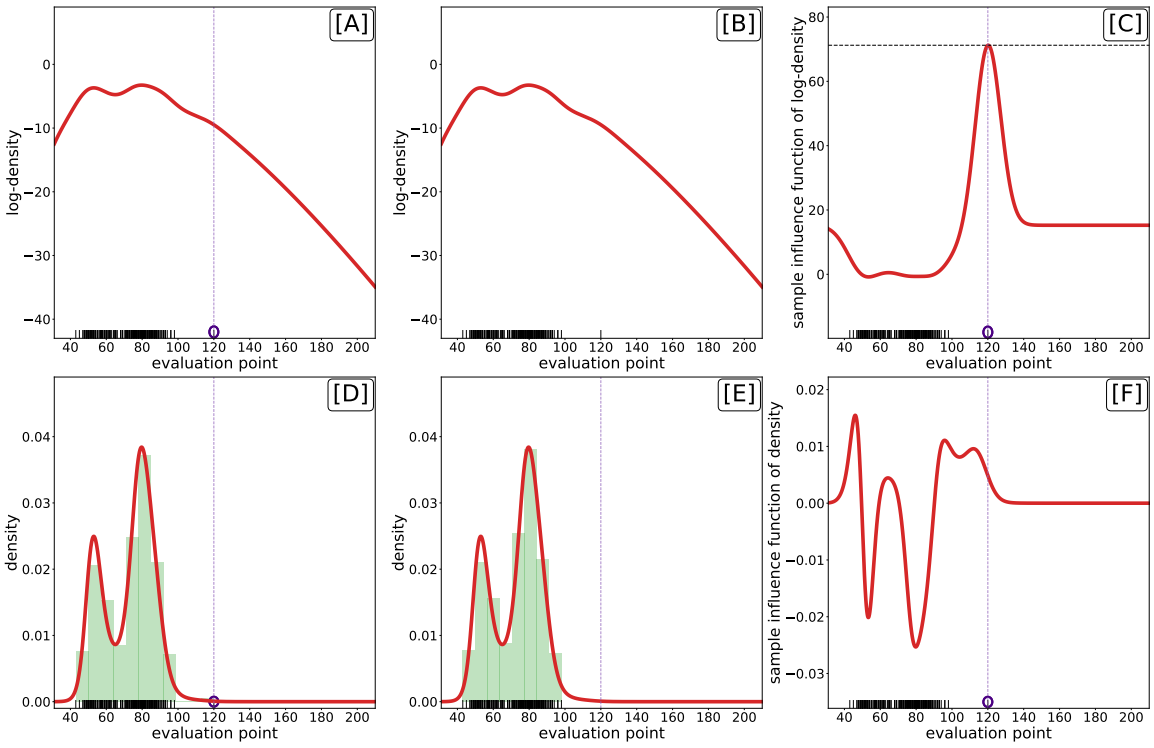

Figure 6.3: Fix $\lambda = e^{-15}$. Panels [A] and [B] show the penalized ML log-density estimates with and without the additional observation $y = 120$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 120$. Panel [F] shows the resulting sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 120$.

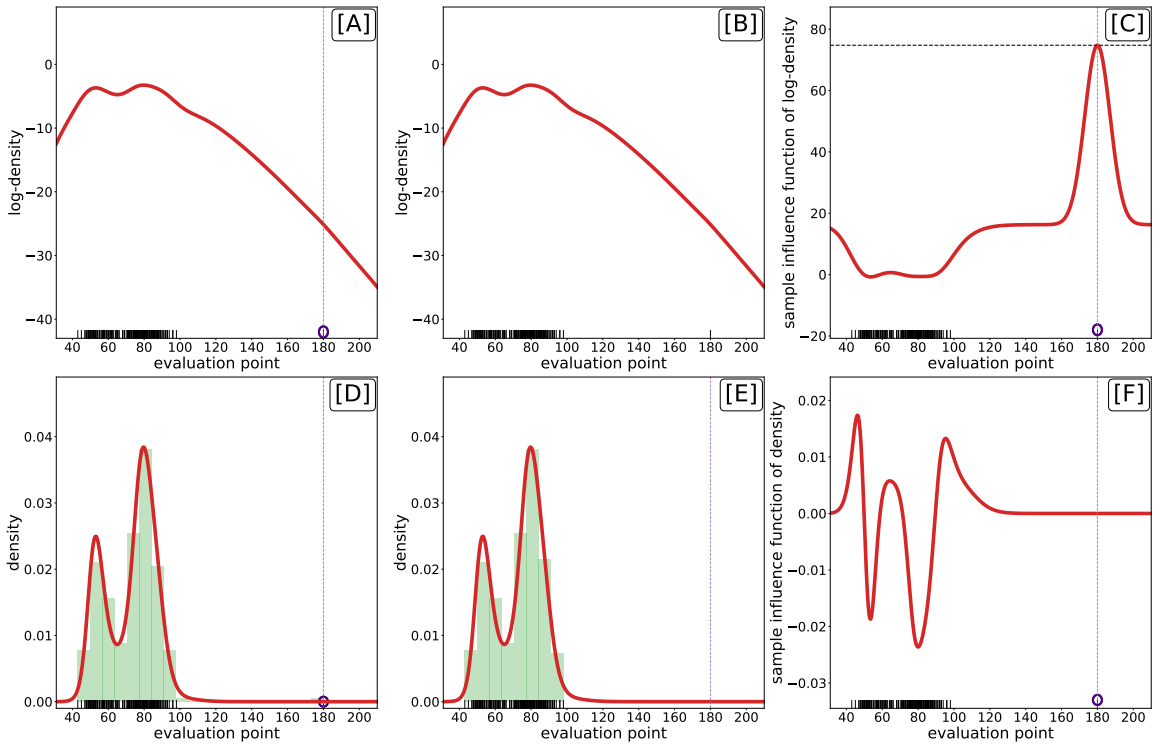

Figure 6.4: Fix $\lambda = e^{-15}$. Panels [A] and [B] show the penalized ML log-density estimates with and without the additional observation $y = 180$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 180$. Panel [F] shows the resulting sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 180$.

from which we see $\mathrm{SIF}_{x,\varepsilon_0}(S_\rho, F, y)$ does *not* depend on $\mu(x)$, but only on the natural parameter part. If we work with $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{S}_\rho, F, y)$, however, we have

$$\begin{aligned}&\widetilde{S}_\rho((1-\varepsilon_0)F_n + \varepsilon_0\delta_y)(x) - \widetilde{S}_\rho(F_n)(x)\\ =&\mu(x)\Big[\exp\big(f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}(x) - A(f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y})\big) - \exp\big(f^{(\rho)}_{\mathrm{SM},F_n}(x) - A(f^{(\rho)}_{\mathrm{SM},F_n})\big)\Big],\end{aligned}$$

and we cannot get rid of $\mu(x)$ in the front. The resulting $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{S}_\rho, F, y)$ inevitably depends on $\mu(x)$. By a similar approach as above, we can also see $\mathrm{SIF}_{x,\varepsilon_0}(T_\lambda, F, y)$ does *not* depend on $\mu(x)$ but $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{T}_\lambda, F, y)$ does.

Thus, from both numerical examples and analytic analysis, we see the sample influence function of the *log*-density estimator is the better choice than that of the density estimator. Hence, we will only consider the former in the sequel.

6.1.3 Comparison of the Sensitivities

Our goal of this section is to compare the sensitivities of the penalized ML and SM density estimators in $\mathcal{Q}_{\mathrm{ker}}$.

Let us still fix $\rho = e^{-11}$ and $y = 120$, and return to the first row of Figure 6.1 where we show $S_\rho((1-\varepsilon_0)F_n + \varepsilon_0\delta_y)$, $S_\rho(F_n)$, and the resulting sample influence function evaluated at different points. It is apparent that $y = 120$ has different effects on S_ρ at different evaluation points. The overall influence of y on S_ρ, $\widehat{M}_{\varepsilon_0}(S_\rho, F_n, y)$, is approximately equal to 2315.48, which is achieved roughly at 120.

Let us still fix $\rho = e^{-11}$ but vary y. The left panel in Figure 6.5 shows the overall influence of y on S_ρ against different choices of y. It is obvious that different locations of y have different overall influences on S_ρ. When y is below 40 or above 100 (low-density region), it has a larger overall influence on S_ρ; and when y is between 40 and 100 (high-density region), it has a smaller overall influence.

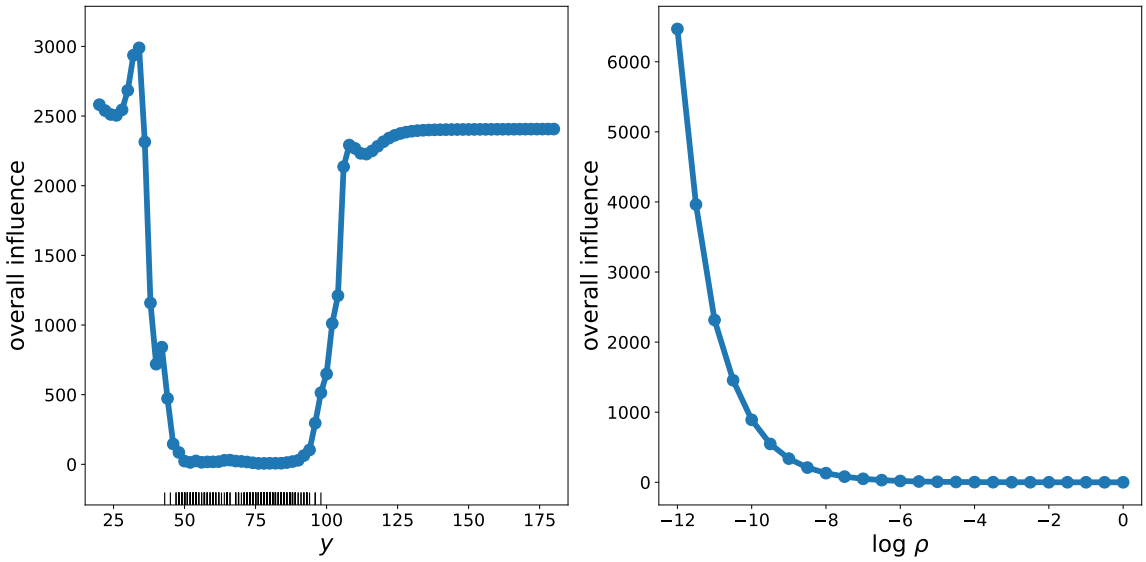

Figure 6.5: Left panel shows the overall influence versus different choices of y, where we fix $\rho = e^{-11}$ and the rugs indicate the location of the `waiting` data. Right panel shows the overall influence against different choices of ρ (shown in log scale), where we fix $y = 120$.

Now, fix $y = 120$ but vary the values of ρ. The right panel in Figure 6.5 shows the overall influence of $y = 120$ on S_ρ versus different choices of ρ. It is obvious that the overall influence of y keeps increasing as the value of ρ keeps decreasing.

Thus, both the location of y and the value of ρ impact the overall influence on log-density estimators. To exhibit the sensitivity of the penalized SM log-density estimators subject to these two factors, a natural idea is to plot the heat map of the overall influence against them, which is shown in Figure 6.6. However, recall our ultimate goal is to compare the sensitivities of both penalized ML and SM density estimators, and their respective penalty parameters, λ and ρ, are on different scales. Thus, plotting the penalty parameter on the horizontal axis is not conducive to comparison. Instead, we plot the RKHS norm of natural parameter under F_n, i.e.,

$\|f^{(\lambda)}_{\mathrm{ML},F_n}\|_{\mathcal{H}}$ and $\|f^{(\rho)}_{\mathrm{SM},F_n}\|_{\mathcal{H}}$, on the horizontal axis. The smaller the penalty parameter is, the larger the RKHS norm of the natural parameter is.

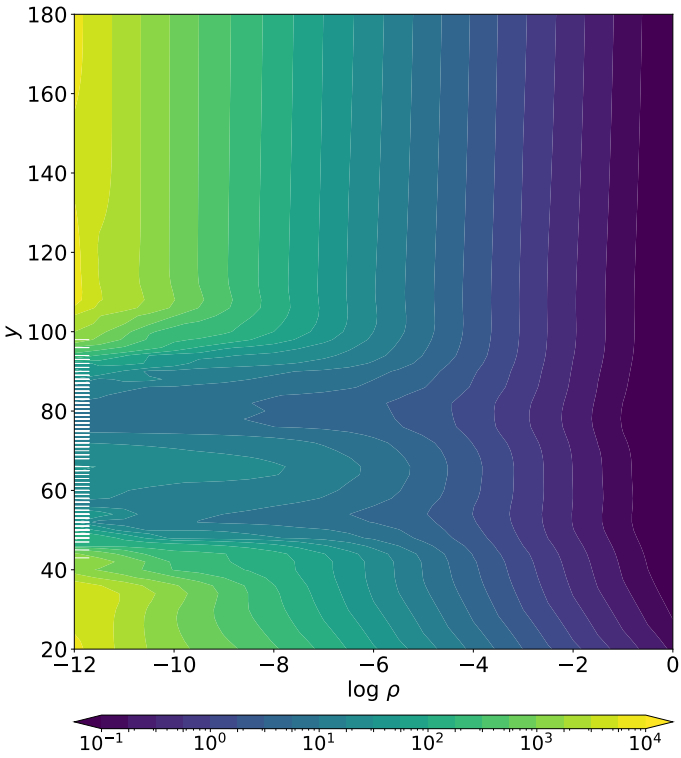

Figure 6.6: Heat map of the overall influence on the penalized SM log-density estimates against y and ρ (shown in log scale). White rugs indicate locations of the `waiting` data.

The resulting heat maps for the penalized ML and SM density estimates are shown in Figure 6.7, where we choose $y = 20, 22, \cdots, 180$, $\lambda = 0, e^{-15}, e^{-14.5}, \cdots, e^{0.5}, e^{1}$, and $\rho = e^{-12}, e^{-11.5}, \cdots, e^{0}$. If we look at each panel individually, findings are consistent as before: with a fixed value of the penalty parameter, y in the low-density region has a larger overall influence on log-density estimates than that in the high-density region; with a fixed y, a larger RKHS norm of the natural parameter (corresponding to a smaller penalty parameter value) implies a larger overall influence of y. In particular,

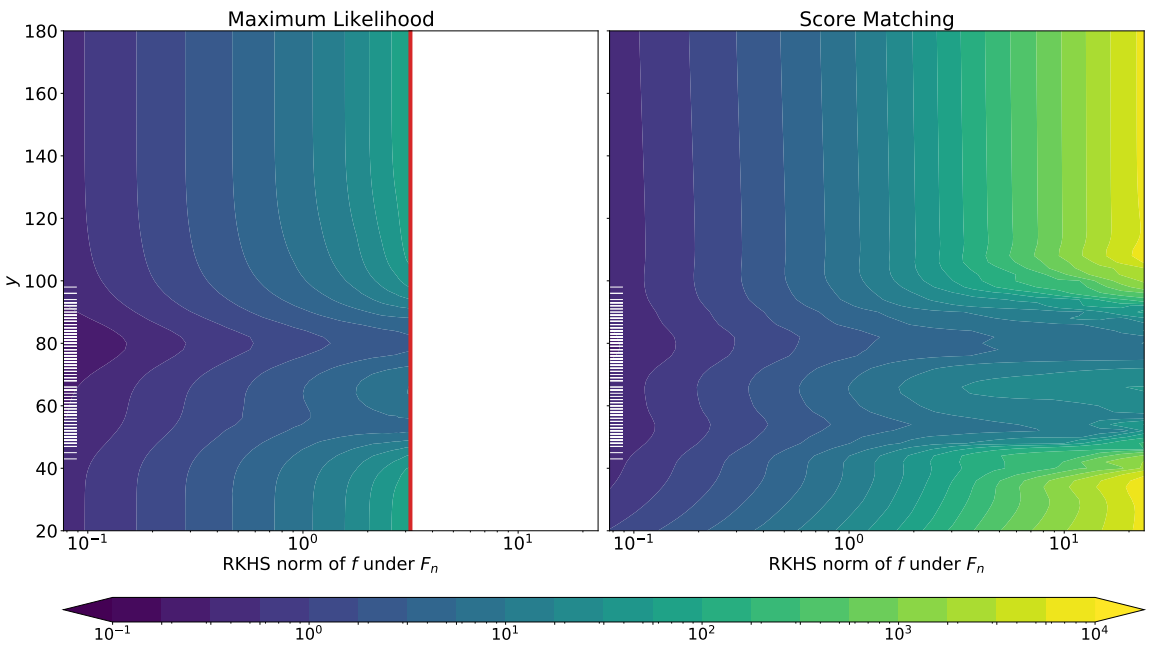

Figure 6.7: Heat maps of the overall influence on penalized ML (left) and SM (right) log-density estimates against y and the RKHS norm of the natural parameter under F_n (shown in log scale). Red vertical line in left panel indicates the case $\lambda = 0$. White rugs indicate locations of `waiting` data.

if we fix a y value, say y_0, the maximal possible overall influence of y_0 on the penalized ML density estimates for all values of $\lambda \geq 0$ is the intersection of $y = y_0$ and the red vertical line, which corresponds to the overall influence of y_0 on the unpenalized ML density estimate (i.e., $\lambda = 0$).

If we compare two panels in Figure 6.7, we observe that, for each choice of y, as we decrease the values of penalty parameters so that the RKHS norm of f under F_n keeps increasing, the overall influences on the penalized ML density estimates stop increasing at the red vertical line, but that on the penalized SM density estimates can continue increasing. In other words, for each choice of y, the overall influences on the penalized SM log-density estimates with sufficiently small values of $\rho > 0$ are larger than the overall influence on the *unpenalized* ML log-density estimate, implying that,

when there is a small amount of penalization, the penalized SM density estimator is more sensitive to an additional observation (not only to the isolated observation) than the unpenalized and penalized ML density estimator. Additional numerical examples also confirm our observations here.

6.2 The Sensitivity of K-fold Cross-validated Penalized SM Density Estimator

K-fold cross-validation (CV) is perhaps the most popular method to choose the penalty parameter in practice. In this section, we are going to investigate whether the K-fold cross-validated penalized SM density estimator is sensitive to the presence of an additional observation or not. The procedure of computing the overall influence of y on the K-fold cross-validated penalized SM density estimator is shown in Algorithm 6.3.

We still use the `waiting` variable in the Old Faithful Geyser dataset and choose the number of folds to be $K = 3, 5, 10$. For each value of K, we replicate the procedures outlined in Algorithm 6.3 for 30 times. The additional observations y we choose are the same as those in the preceding section, $y = 20, 22, \cdots, 180$. Figure 6.8 shows the results.

For each number of folds, similar to our earlier observations, when y is in the high-density region (between 40 and 100), the overall influences of y on cross-validated penalized SM density estimates tend to be small; and when y is in the low-density region (< 40 or > 100), the overall influences of y tend to be much larger.

In addition, we see as the number of folds increases, the cross-validated penalized SM log-density estimates become less sensitive in general. This suggests that when one uses the K-fold CV to select the penalty parameter, it is better to use a relatively

Algorithm 6.3 Computation of the overall influence of K-fold cross-validated penalized SM density estimator

Require:

- $X_1, \cdots, X_n$, data;
- $y \in \mathcal{X}$, contaminant;
- K, the number of folds in CV;
- $\{\rho_j\}_{j=1}^M$, a list of penalty parameter candidates;
- $\widetilde{\mathcal{H}}$, the finite-dimensional approximating subspace over which we minimize the penalized NLL loss functional;
- $\{x_\ell\}_{\ell=1}^L \subset \mathcal{X}$, a set of dense evaluation points in $\mathcal{X}$.

1: Compute the best density estimate using $X_1, \cdots, X_n$ with the penalty parameter selected by CV; denote the resulting log-density estimate by $S_{\mathrm{CV}}(F_n)$;
2: Compute the best density estimate using $X_1, \cdots, X_n$ and y with the penalty parameter selected by CV; denoted the resulting log-density estimate by $S_{\mathrm{CV}}(F_{n+1})$;
3: Compute

$$\mathrm{SIF}_{x_\ell, \varepsilon_0}(S_{\mathrm{CV}}, F_n, y) := \Big(S_{\mathrm{CV}}(F_{n+1})(x_\ell) - S_{\mathrm{CV}}(F_n)(x_\ell) \Big) \times (n+1),$$

for all $\ell = 1, \cdots, L$;
4: **return** $\max_{\ell=1,\cdots,L} \big| \mathrm{SIF}_{x_\ell, \varepsilon_0}(S_{\mathrm{CV}}, F_n, y) \big|$.

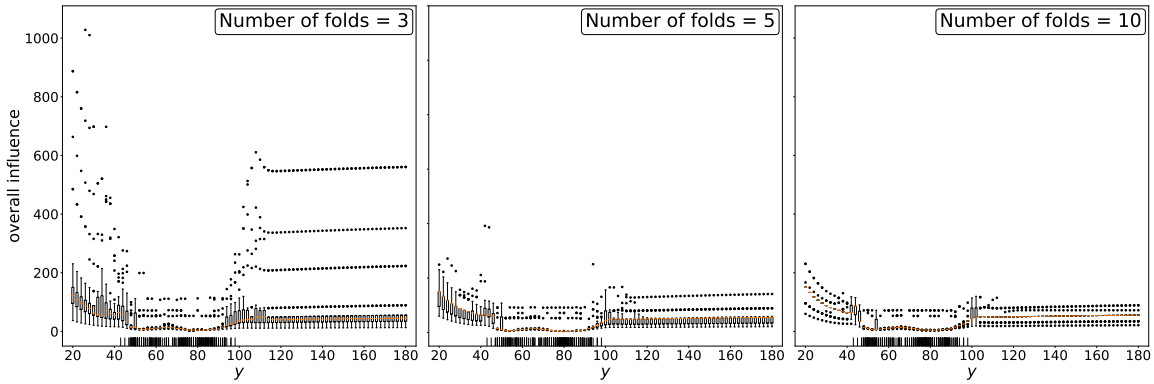

Figure 6.8: Overall influence of y on the K-fold cross-validated penalized SM density estimates against the values of y. We choose $K = 3$ (left panel), 5 (middle panel), and 10 (right panel).

large number of folds, e.g., $K = 5$ or 10, in order to obtain a density estimate that is *not* sensitive to the presence of an isolated observation.

6.3 Which One to Use: Penalized ML or Regularized SM Density Estimators?

Now, we can compute penalized ML and regularized SM density estimators within $\mathcal{Q}_{\mathrm{ker}}$. A natural question arises: which density estimator should one use? Regularized SM density estimators are easy to compute, and can be obtained by solving a linear system or adding and multiplying certain matrices, but are sensitive to the presence of isolated observations, especially when the amount of regularization is small. Penalized ML density estimator, on the other hand, is hard to compute as one has to handle A and its derivative, but is *not* very sensitive to the presence of isolated observations. We recommend using regularized SM density estimators, primarily due to their computational advantage. When using them, one needs to impose an appropriate amount of regularization to ensure the resulting density estimate is not just a spike at the isolated observation; otherwise, such a density estimate can be of no use

and be misleading. One can use the K-fold cross-validation to select the appropriate amount of regularization as we did in this chapter and the earlier ones.

Chapter 7: Summary and Future Directions

7.1 Summary

In this dissertation, we focus on the density estimation problem in an exponential family induced by a RKHS, $\mathcal{Q}_{\text{ker}}$. We proposed a new early stopping SM density estimator obtained via minimizing the (unpenalized) SM loss functional by the gradient descent algorithm and terminating early. We studied its statistical properties and compared it with the penalized SM density estimator in the literature and showed their similarities and differences.

We also compared these two kinds of regularized SM density estimators with the penalized ML density estimator. Via numerical examples, we observed that the regularized SM density estimators are very sensitive to the presence of an isolated observation, especially when there is a small amount of regularization, but the penalized ML density estimator does *not*. We attempted to explain why this happens.

In order to understand this phenomenon, we extended the classic notion of the influence function to allow its input to be a function-valued statistical functional. Using this extended influence function, we studied the sensitivity of regularized SM and ML density estimators in both finite-dimensional and kernel exponential families.

As we have mentioned at the end of the preceding chapter, we recommend to use the regularized SM density estimators, mainly due to its computational advantages.

But when using them, one needs to make sure an appropriate amount of regularization has been imposed.

7.2 Future Directions

Numerical examples shown in this dissertation are restricted to $d = 1$. We conjecture that regularized SM density estimators in higher dimensions ($d \geq 2$) are also very sensitive to the presence of an isolated observation when the regularization is small, and need numerical examples to confirm this.

Several versions of generalized SM loss functionals have been proposed in the literature in recent years. Parry, Dawid, and Lauritzen (2012) proposed a version that does not only involve the first two derivatives of a log-density function but also the higher-order derivatives. Yu, Drton, and Shojaie (2020) proposed the following generalized H-divergence between p_0 and q

$$\int_{\mathcal{X}} p_0(x) \left\| \nabla \log p_0(x) \odot \sqrt{h(x)} - \nabla \log q(x) \odot \sqrt{h(x)} \right\|_2^2 \mathrm{d}x, \tag{7.1}$$

where $\odot$ denotes element-wise multiplication of two vectors, $h : \mathcal{X} \to [0, \infty)^d$, and the square root is taken element-wise. Under certain regularity conditions, one can apply integration by parts and obtain a loss functional from (7.1) that depends on q only. It is interesting to apply these generalized SM loss functionals to the density estimation problem in $\mathcal{Q}_{\text{ker}}$ and see whether the resulting density estimators are sensitive to the presence of an isolated observation. We conjecture, with an appropriate choice of h in (7.1), the resulting density estimators may *not* be very sensitive to isolated observations.

As we have discussed in Chapter 1, there has been an increasing interest in log-concave density estimation recently. The dominant approach is to minimize the NLL loss functional over the class of log-concave density functions over $\mathbb{R}^d$, resulting in

the ML log-concave density estimator. This approach leads to a non-differentiable objective function to optimize and involves approximations of the normalizing constant and its sub-gradient, which is computationally challenging. In addition, as is shown in Figure 7.1, the ML log-concave density estimates contain ridges, and are only supported over the convex hull of data with all boundary points being the discontinuous points of the density estimate, which are undesirable qualitative features and may cause series issues in statistical applications.

Since the SM loss functional does not involve the normalizing constant and involves the first two partial derivatives of the log-density function, it is interesting to estimate a log-concave density function by minimizing the SM loss functional and compare the resulting SM log-concave density estimator with the ML log-concave density estimator. We conjecture the issues with the ML log-concave density estimator described above can be solved by using the SM loss functional.

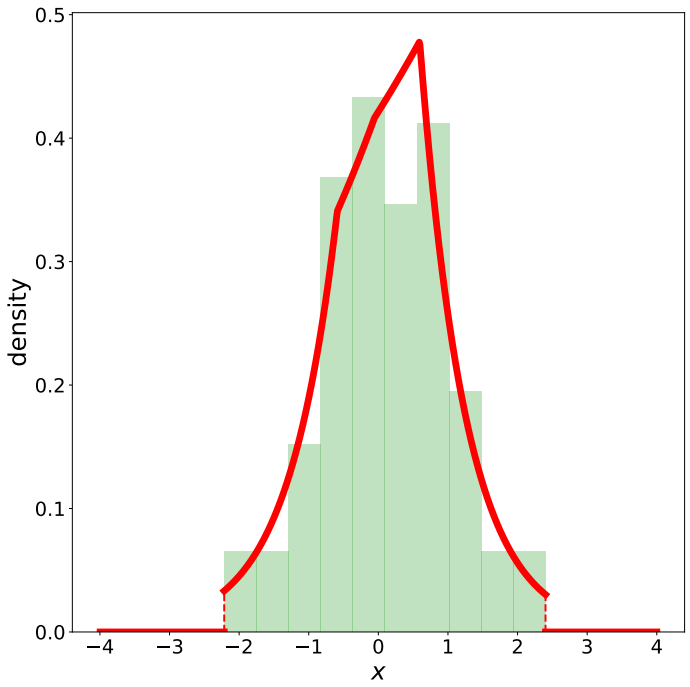

Figure 7.1: ML log-concave density estimate with 100 random samples from the standard normal distribution, where the density estimate is computed using the `R` package `logcondens` (Dümbgen and Rufibach, 2010). Histogram with the bin width chosen by the Freedman-Diaconis rule is shown in green.

Appendix A: Math Background

A.1 Fréchet Differentiability and Derivative

We provide details on the Fréchet differentiability and derivative. Throughout this section, we let $\mathcal{H}$ be a real Hilbert space, $J : \mathcal{H} \to \mathbb{R}$ be a map, $\mathcal{B}(\mathcal{H}, \mathbb{R})$ denote the collection of all bounded linear operators from $\mathcal{H}$ to $\mathbb{R}$, and, similarly, $\mathcal{B}(\mathcal{H}, \mathcal{H})$ denote the collection of all bounded linear operators from $\mathcal{H}$ to itself. All materials of this section come from Section 2.6 in Bauschke and Combettes (2011) and Section 5.1 in Denkowski, Migórski, and Papageorgiou (2013).

Definition A.1 (Fréchet differentiability and derivative)**.** The map J is said to be *(first-order) Fréchet differentiable* at $f \in \mathcal{H}$ if there exists an operator $\mathrm{D}J(f) \in \mathcal{B}(\mathcal{H}, \mathbb{R})$ such that

$$\lim_{\substack{\|g\|_{\mathcal{H}} \to 0 \\ g \neq 0}} \frac{\big|J(f+g) - J(f) - \mathrm{D}J(f)(g)\big|}{\|g\|_{\mathcal{H}}} = 0, \tag{A.1}$$

and the operator $\mathrm{D}J(f)$ is called the *(first-order) Fréchet derivative*. The map J is said to be *(first-order) Fréchet differentiable on* $\mathcal{H}$ if it is Fréchet differentiable at all $f \in \mathcal{H}$.

Proposition A.1. *Suppose J is Fréchet differentiable at $f \in \mathcal{H}$ and the Fréchet derivative $\mathrm{D}J(f)$ exists. Then, $\mathrm{D}J(f)$ is unique.*

Remark A.1. If J is Fréchet differentiable at $f \in \mathcal{H}$, we then can write

$$J(f+g) = J(f) + \mathrm{D}J(f)(g) + o(\|g\|_{\mathcal{H}}), \tag{A.2}$$

for all $g \in \mathcal{H}$ in a small neighborhood of the origin, where $o(\|g\|_{\mathcal{H}})$ denotes $\frac{o(\|g\|_{\mathcal{H}})}{\|g\|_{\mathcal{H}}} \to 0$ as $\|g\|_{\mathcal{H}} \to 0$. Thus, from (A.2), we see $J(f) + \mathrm{D}J(f)(g)$ provides the best linear approximation of J in a small neighborhood of f, which is the similar interpretation of the derivative of a real-valued function of a single variable. ▶

Frechét derivative shares many properties of the derivative of a real-valued function of a single variable. The following proposition lists two properties we use in studying the Frechét differentiability and deriving the Frechét derivative of the log-partition functional A in Chapter 2.

Proposition A.2.

(a) Suppose $J_1, J_2 : \mathcal{H} \to \mathbb{R}$ are Frechét differentiable at $f \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$ are arbitrary. Then, $\alpha J_1 + \beta J_2$ is also Frechét differentiable at $f \in \mathcal{H}$, and

$$\mathrm{D}(\alpha J_1 + \beta J_2)(f) = \alpha \mathrm{D}J_1(f) + \beta \mathrm{D}J_2(f).$$

(b) (Chain rule) Suppose $J_1 : \mathcal{H} \to \mathbb{R}$ is Frechét differentiable at $f \in \mathcal{H}$ and $J_2 : \mathbb{R} \to \mathbb{R}$ is differentiable at $J_1(f)$. Then, $J_2 \circ J_1 : \mathcal{H} \to \mathbb{R}$ is Frechét differentiable at $f \in \mathcal{H}$, and

$$\mathrm{D}(J_2 \circ J_1)(f) = J_2'(J_1(f))\mathrm{D}J_1(f). \tag{A.3}$$

Definition A.2 (Fréchet gradient)**.** Suppose $J : \mathcal{H} \to \mathbb{R}$ is Frechét differentiable at $f \in \mathcal{H}$. Since $\mathrm{D}J(f)$ is a bounded linear map from $\mathcal{H}$ to $\mathbb{R}$, the Riesz-Fréchet representation theorem (Fact 2.24 in Bauschke and Combettes, 2011) guarantees there exists a unique element $\nabla J(f) \in \mathcal{H}$ such that, for any $g \in \mathcal{H}$,

$$\mathrm{D}J(f)(g) = \langle g, \nabla J(f) \rangle_{\mathcal{H}}, \tag{A.4}$$

and $\nabla J(f)$ is called the *Fréchet gradient* of J at f. If J is Fréchet differentiable on $\mathcal{H}$, the *Fréchet gradient operator* is defined to be $\nabla J : \mathcal{H} \to \mathcal{H}, f \mapsto \nabla J(f)$.

Remark A.2. Note that $\mathrm{D}J(f)$ is a bounded linear map from $\mathcal{H}$ to $\mathbb{R}$, and belongs to the dual space of $\mathcal{H}$, denoted by $\mathcal{H}^*$. Since we have $\mathrm{D}J(f)(g) = \langle \nabla J(f), g \rangle_{\mathcal{H}}$, the Riesz-Fréchet representation theorem implies that $\|\mathrm{D}J(f)\|_{\mathcal{H}^*} = \|\nabla J(f)\|_{\mathcal{H}}$, where $\|\cdot\|_{\mathcal{H}^*}$ denotes the norm of the dual space $\mathcal{H}^*$. ▶

We now extend Definition A.1 to higher orders.

Definition A.3 (Higher-order Fréchet differentiability and derivatives)**.** Higher-order Fréchet differentiability and derivatives are defined inductively.

In particular, the map J is said to be *twice Fréchet differentiable at* $f \in \mathcal{H}$ if J itself is Fréchet differentiable at $f \in \mathcal{H}$ and the map $\mathrm{D}J(f) : \mathcal{H} \to \mathbb{R}$ is also Fréchet differentiable at $f \in \mathcal{H}$. The *second Fréchet derivative* of J at $f \in \mathcal{H}$, denoted by $\mathrm{D}^2 J(f)$, is an operator from $\mathcal{H}$ to $\mathcal{B}(\mathcal{H}, \mathbb{R})$, that satisfies

$$\lim_{\substack{\|g\|_{\mathcal{H}} \to 0 \\ g \neq 0}} \frac{\left\| \mathrm{D}J(f+g) - \mathrm{D}J(f) - \mathrm{D}^2 J(f)(g) \right\|_{\mathcal{H}^*}}{\|g\|_{\mathcal{H}}} = 0, \tag{A.5}$$

where $\|\cdot\|_{\mathcal{H}^*}$ denotes the norm of the dual space of $\mathcal{H}$. The map J is said to be *twice Fréchet differentiable on* $\mathcal{H}$ if it is twice Fréchet differentiable at all $f \in \mathcal{H}$.

Suppose J is twice Fréchet differentiable on $\mathcal{H}$. The *second-order Fréchet gradient operator*, denoted by $\nabla^2 J$, is a bounded linear operator that maps from $\mathcal{H}$ to $\mathcal{B}(\mathcal{H}, \mathcal{H})$ and satisfies

$$\mathrm{D}^2 J(f)(g)(h) = \langle h, \nabla^2 J(f)(g) \rangle_{\mathcal{H}}, \qquad \text{for all } f, g, h \in \mathcal{H}.$$

In other words, $\nabla^2 J \in \mathcal{B}(\mathcal{H}, \mathcal{B}(\mathcal{H}, \mathcal{H}))$ and $\nabla^2 J(f) \in \mathcal{B}(\mathcal{H}, \mathcal{H})$ for all $f \in \mathcal{H}$.

A.2 Bochner Integral

In this section, we present the definition of the Bochner integral, which is the extension of the Lebesgue integral of real-valued functions to the integral of functions taking values in a Banach space. We also present some properties of the Bochner integral that we have used in the dissertation (in particular, in Chapter 2 and 3). All materials of this section come from Appendix A.5.3 in Steinwart and Christmann (2008) and Section 3.10 Denkowski, Migórski, and Papageorgiou (2013).

Throughout this section, let $\mathcal{E}$ be a Banach space whose norm is denoted by $\|\cdot\|_{\mathcal{E}}$, and $(\mathcal{X}, \Sigma, \mu)$ be a σ-finite measure space (note that this μ differs from the one in the definition of finite-dimensional and kernel exponential families in Chapter 2). We first define the simple function (Definition A.4) and the measurable function (Definition A.5) in the Banach space setting and then define the Bochner μ-integral (Definition A.6).

Definition A.4 ($\mathcal{E}$-valued simple function)**.** A function $s : \mathcal{X} \to \mathcal{E}$ is said to be an *$\mathcal{E}$-valued simple function* if there exist $e_1, \cdots, e_n \in \mathcal{E}$ and $A_1, \cdots, A_n \in \Sigma$ such that

$$s(x) = \sum_{i=1}^{n} \mathbb{1}_{A_i}(x) e_i, \qquad \text{for all } x \in \mathcal{X},$$

where $\mathbb{1}_A$ is the indicator function of the set A, and is equal to 1 if $x \in A$ and to 0, otherwise.

Definition A.5 ($\mathcal{E}$-valued measurable function)**.** A function $f : \mathcal{X} \to \mathcal{E}$ is said to be an *$\mathcal{E}$-valued measurable function* if there exists a sequence of $\mathcal{E}$-valued simple functions, $\{s_n\}_{n \in \mathbb{N}}$, such that

$$\lim_{n \to \infty} \|f(x) - s_n(x)\|_{\mathcal{E}} = 0 \tag{A.6}$$

holds for all $x \in \mathcal{X}$.

Definition A.6 (Bochner μ-integral)**.** An $\mathcal{E}$-valued measurable function $f : \mathcal{X} \to \mathcal{E}$ is said to be *Bochner μ-integrable* if there exists a sequence of $\mathcal{E}$-valued simple functions, $\{s_n\}_{n\in\mathbb{N}}$, such that

$$\lim_{n\to\infty} \int_{\mathcal{X}} \|s_n(x) - f(x)\|_{\mathcal{E}} \, \mathrm{d}\mu(x) = 0. \tag{A.7}$$

In this case, the limit

$$\int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) := \lim_{n\to\infty} \int_{\mathcal{X}} s_n(x)\mathrm{d}\mu(x)$$

exists and is called the *Bochner μ-integral* of f.

A criterion to check the Bochner μ-integrability is the following.

Proposition A.3. *A measurable function $f : \mathcal{X} \to \mathcal{E}$ is Bochner μ-integrable if and only if $\int_{\mathcal{X}} \|f(x)\|_{\mathcal{E}} \mathrm{d}\mu(x) < \infty$.*

Finally, we look at some properties of Bochner μ-integral we use.

Proposition A.4. *The Bochner μ-integral defined above has the following properties:*

(a) The Bochner μ-integral is linear.

(b) If $f : \mathcal{X} \to \mathcal{E}$ is Bochner μ-integrable, we have

$$\left\| \int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) \right\|_{\mathcal{E}} \leq \int_{\mathcal{X}} \|f(x)\|_{\mathcal{E}} \mathrm{d}\mu(x).$$

(c) Suppose $\mathcal{E}'$ is another Banach space. If $S : \mathcal{E} \to \mathcal{E}'$ is a bounded linear operator and $f : \mathcal{X} \to \mathcal{E}$ is Bochner μ-integrable, then $S \circ f : \mathcal{X} \to \mathcal{E}'$ is also Bochner μ-integrable. In this case, the integral commutes with S, that is,

$$S\left(\int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) \right) = \int_{\mathcal{X}} (S \circ f)(x) \, \mathrm{d}\mu(x).$$

A.3 Partial Derivative of a Kernel Function

In this section, we discuss the partial derivatives of the kernel function associated with a RKHS and their reproducing property. We follow the development in Section 4.3 in Steinwart and Christmann (2008) and the paper by Zhou (2008). Throughout this section, we let $\mathcal{X} \subseteq \mathbb{R}^d$ be an open set and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

We first consider a real-valued function $f : \mathcal{X} \to \mathbb{R}$. The function f is said to be *m-times continuously differentiable* if, for all $\alpha := (\alpha_1, \cdots, \alpha_d)^\top \in \mathbb{N}_0^d$ with $|\alpha| := \sum_{i=1}^d \alpha_i \leq m$ and all $x \in \mathcal{X}$,

$$\partial^\alpha f(x) = \frac{\partial^{|\alpha|}}{\partial u_1^{\alpha_1} \cdots \partial u_d^{\alpha_d}} f(u)\bigg|_{u=x},$$

exists, where $u := (u_1, \cdots, u_d)^\top \in \mathcal{X}$.

We then define the m-times continuous differentiability of the kernel function k.

Definition A.7 (m-times continuous differentiability of a kernel function)**.** Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function and $m \in \mathbb{N}$. We say k is *m-times continuously differentiable* if $\partial^{\alpha,\alpha} k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ exists and is continuous for all $\alpha := (\alpha_1, \cdots, \alpha_d) \in \mathbb{N}_0^d$ with $|\alpha| := \sum_{i=1}^d \alpha_i \leq m$, where

$$\partial^{\alpha,\alpha} k(x, y) = \frac{\partial^{2|\alpha|}}{\partial u_1^{\alpha_1} \cdots \partial u_d^{\alpha_d} \partial v_1^{\alpha_1} \cdots \partial v_d^{\alpha_d}} k(u, v)\bigg|_{u=x, v=y}, \qquad \text{for all } x, y \in \mathcal{X}.$$

The partial derivative of k is an element in $\mathcal{H}$ and has the reproducing property as k does, as the following proposition states.

Proposition A.5 (Partial derivatives of kernels and its reproducing property)**.** *Let $\mathcal{H}$ be a RKHS with the kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and assume k is m-times continuously differentiable on $\mathcal{X}$. Then,*

(a) we have

$$\partial^\alpha k(x, \cdot) = \frac{\partial^{|\alpha|}}{\partial u_1^{\alpha_1} \cdots \partial u_d^{\alpha_d}} k(u, \cdot)\bigg|_{u=x} \in \mathcal{H} \tag{A.8}$$

where $u := (u_1, \cdots, u_d)^\top \in \mathcal{X}$, *and*

(b) every $f \in \mathcal{H}$ *is* m*-times continuously differentiable, and, for all* $\alpha \in \mathbb{N}_0^d$ *with* $|\alpha| \leq m$ *and all* $x \in \mathcal{X}$, *the partial derivative reproducing property holds, i.e.,*

$$\partial^\alpha f(x) = \langle \partial^\alpha k(x, \cdot), f \rangle_{\mathcal{H}}, \qquad \text{for all } x \in \mathcal{X}. \tag{A.9}$$

In particular, we have $\partial^{\alpha,\alpha} k(x, y) = \langle \partial^\alpha k(x, \cdot), \partial^\alpha k(y, \cdot) \rangle_{\mathcal{H}}$ *for all* $x, y \in \mathcal{X}$.

A.4 Some Theories on Bounded Linear Operators

In our development in Chapters 2 and 3, we have used some theories on the bounded linear operators in a Hilbert space. In this section, we aim to give a very brief overview of these theories. More information can be found in, for example, Chapters 13 and 16 in Royden and Fitzpatrick (2018) and Chapter VI in Reed and Simon (2012).

Throughout this section, we let $\mathcal{H}$ be a real Hilbert space with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and the norm $\| \cdot \|_{\mathcal{H}}$, and assume $\mathcal{H}$ is separable, meaning that it contains a dense countable subset. Due to the separability of $\mathcal{H}$, it admits a countable orthonormal basis.

Definition A.8 (Linear operator)**.** An operator $C : \mathcal{H} \to \mathcal{H}$ is said to be *linear* if, for any $f, g \in \mathcal{H}$ and any $\alpha, \beta \in \mathbb{R}$, we have

$$C(\alpha f + \beta g) = \alpha C f + \beta C g.$$

Definition A.9 (Bounded operator)**.** A linear operator $C : \mathcal{H} \to \mathcal{H}$ is said to be *bounded* if there exists a constant $M \geq 0$ such that

$$\|Cf\|_{\mathcal{H}} \leq M \cdot \|f\|_{\mathcal{H}}, \qquad \text{for all } f \in \mathcal{H}.$$

The infimum of all such M is called the *operator norm* of C and is denoted by $\|C\|$.

Definition A.10 (Adjoint and self-adjoint operators)**.** Let $C : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator. Then, the *adjoint* of C is the bounded linear operator $C^* : \mathcal{H} \to \mathcal{H}$ satisfying

$$\langle Cf, g\rangle_{\mathcal{H}} = \langle f, C^* g\rangle_{\mathcal{H}}, \qquad \text{for all } f, g \in \mathcal{H}.$$

The adjoint C^* exists and is unique.

The operator C is said to be *self-adjoint*, if $C = C^*$, that is, $\langle Cf, g\rangle_{\mathcal{H}} = \langle f, Cg\rangle_{\mathcal{H}}$ for all $f, g \in \mathcal{H}$.

Definition A.11 (Positive semidefinite and definite operators)**.** Let $C : \mathcal{H} \to \mathcal{H}$ be a self-adjoint bounded linear operator. Then, C is said to be *positive semidefinite* if, for all $f \in \mathcal{H}$, $\langle f, Cf\rangle_{\mathcal{H}} \geq 0$, and to be *positive definite* if, for all $f \in \mathcal{H}\backslash\{0\}$, $\langle f, Cf\rangle_{\mathcal{H}} > 0$.

Definition A.12 (Compact operator)**.** A bounded linear operator $C : \mathcal{H} \to \mathcal{H}$ is said to be *compact*, if, for every bounded sequence $\{f_n\}_{n\in\mathbb{N}}$ in $\mathcal{H}$, $\{Cf_n\}_{n\in\mathbb{N}}$ has a subsequence that converges in $\mathcal{H}$ (with respect to the norm $\|\cdot\|_{\mathcal{H}}$).

Other equivalent ways of defining a compact operator can be found, for example, in Section 16.5 in Royden and Fitzpatrick (2018) and Section VI.5 in Reed and Simon (2012).

Definition A.13 (Finite rank operator)**.** A bounded linear operator $C : \mathcal{H} \to \mathcal{H}$ is said to be of *finite rank* if its range is finite-dimensional. That is, every element in range(C) can be written as $Cf = \sum_{i=1}^{m} \alpha_i g_i$, for some $f \in \mathcal{H}$, some fixed family $\{g_i\}_{i=1}^{m} \subset \mathcal{H}$, some $\alpha_1, \cdots, \alpha_m \in \mathbb{R}$, and $m < \infty$.

The relationship between the compact operator and the finite rank operator is given in the following proposition.

Proposition A.6. *Every finite rank operator is compact. In addition, every compact operator on $\mathcal{H}$ is the norm limit of a sequence of finite rank operators.*

In addition, we have the following characterization when the identity operator is compact.

Proposition A.7. *The identity operator $I : \mathcal{H} \to \mathcal{H}$ is compact if and only if $\mathcal{H}$ is finite-dimensional.*

We also have the following result characterizing the relationship between the invertibility and the compactness of a bounded linear operator.

Proposition A.8. *Let $\mathcal{H}$ be infinite-dimensional and $C : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator. If C is compact, it is* not *invertible.*

In the development in Chapter 3, we have repeatedly used the following Hilbert-Schmidt theorem, which is an extension of the eigen-decomposition of a real symmetric matrix.

Theorem A.1 (Hilbert-Schmidt theorem)**.** *Let $C : \mathcal{H} \to \mathcal{H}$ be a self-adjoint compact operator. There exists an orthonormal basis $\{\psi_\nu\}_{\nu=1}^R$ for $\overline{\text{range}(C)}$, together with a monotonically non-increasing sequence of nonzero real numbers $\{\xi_\nu\}_{\nu=1}^R$, such that $C\psi_\nu = \xi_\nu\psi_\nu$ for all $\nu = 1, \cdots, R$. In addition, the following identity holds*

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

If C is of finite rank, then $R < \infty$ and R is the rank of C; if C is not of finite rank, $R = \infty$ and $\lim_{\nu\to\infty} \xi_\nu = 0$.

In Theorem A.1, the functions $\psi_1, \cdots, \psi_R$ are called the *eigenfunctions* of C, and ξ_ν is called the *eigenvalue* of C associated with the eigenfunction ψ_ν, for all $\nu = 1, \cdots, R$.

Using Theorem A.1, it is easy to see that, if C is positive semidefinite, all its eigenvalues are positive.

We next introduce two special classes of bounded linear operators, the trace class and the Hilbert-Schmidt class, that are of great importance in proving various properties of penalized and early stopping SM density estimators in Chapter 2 and 3.

We first introduce the trace of a bounded linear operator, based on which we define the trace class.

Definition A.14 (Trace)**.** Let $\{\psi_\nu\}_{\nu=1}^{\infty}$ be an orthonormal basis of $\mathcal{H}$. For any positive definite bounded linear operator $C : \mathcal{H} \to \mathcal{H}$, the *trace* of C is defined to be

$$\text{trace}(C) := \sum_{\nu=1}^{\infty} \langle \psi_\nu, C\psi_\nu \rangle_{\mathcal{H}},$$

where the sum is independent of the choice of the orthonormal basis.

Definition A.15 (Trace class)**.** A bounded linear operator $C : \mathcal{H} \to \mathcal{H}$ is said to be of *trace class* if $\text{trace}(|C|) < \infty$.

Then, we have the following properties of operators that are of trace class.

Proposition A.9. *Let $C : \mathcal{H} \to \mathcal{H}$ be of trace class. Then, C is compact, and its operator norm, $\|C\|$, and its trace,* $\text{trace}(C)$*, are related by* $\|C\| \leq \text{trace}(C)$.

We now turn to the Hilbert-Schmidt operator.

Definition A.16 (Hilbert-Schmidt operator)**.** Let $\{\psi_\nu\}_{\nu=1}^{\infty}$ be an orthonormal basis of $\mathcal{H}$. A bounded linear operator $C : \mathcal{H} \to \mathcal{H}$ is said to be *Hilbert-Schmidt (HS)* if

$$\sum_{\nu=1}^{\infty} \|C\psi_\nu\|_{\mathcal{H}}^2 < \infty.$$

Still let $\{\psi_\nu\}_{\nu=1}^{\infty}$ be an orthonormal basis of $\mathcal{H}$. Given two HS operators $C_1, C_2 : \mathcal{H} \to \mathcal{H}$, define the *HS inner product* between them to be

$$\langle C_1, C_2 \rangle_{\mathrm{HS}} = \sum_{\nu=1}^{\infty} \langle C_1\psi_\nu, C_2\psi_\nu \rangle_{\mathcal{H}},$$

and the *HS norm* to be

$$\|C_1\|_{\mathrm{HS}} := \sqrt{\langle C_1, C_1 \rangle_{\mathrm{HS}}} = \sqrt{\mathrm{trace}(C_1^* C_1)} = \left(\sum_{\nu=1}^{\infty} \|C_1\psi_\nu\|_{\mathcal{H}}^2 \right)^{1/2}.$$

We then have the following properties of HS operators.

Proposition A.10 (Properties of HS operators)**.** *The following properties of the HS operators hold:*

(a) Let $C : \mathcal{H} \to \mathcal{H}$ be a HS operator. Then, its operator norm $\|C\|$, its trace $\mathrm{trace}(C)$, and its HS norm $\|C\|_{\mathrm{HS}}$ are related by $\|C\| \leq \|C\|_{\mathrm{HS}} \leq \mathrm{trace}(C)$.

(b) The class of all HS operators with the inner product $\langle \cdot, \cdot \rangle_{\mathrm{HS}}$ forms a Hilbert space.

(c) Every HS operator is compact.

Bibliography

Axelrod, Brian et al. (July 2019). "A Polynomial Time Algorithm for Log-Concave Maximum Likelihood via Locally Exponential Families". In: arXiv: 1907.08306 [cs.DS].

Azzalini, A and A W Bowman (1990). "A look at some data on the old faithful geyser". In: *J. R. Stat. Soc. Ser. C Appl. Stat.* 39.3, p. 357.

Bagnoli, Mark and Ted Bergstrom (2005). "Log-Concave Probability and Its Applications". In: *Economic Theory* 26.2, pp. 445–469. URL: http://www.jstor.org/stable/25055959.

Baker, Charles R. (1973). "Joint Measures and Cross-Covariance Operators". In: *Transactions of the American Mathematical Society* 186, pp. 273–289. (Visited on 06/13/2022).

Balabdaoui, Fadoua, Kaspar Rufibach, and Jon A Wellner (June 2009). "Limit Distribution Theory for Maximum Likelihood Estimation of a Log-Concave Density". en. In: *Ann. Stat.* 37.3, pp. 1299–1331.

Barndorff-Nielsen, O (May 2014). *Information and Exponential Families: In Statistical Theory*. en. John Wiley & Sons.

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

Bauschke, Heinz H. and Patrick L. Combettes (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. 1st. Springer Publishing Company. ISBN: 9781441994660.

Bertinet, A. and Thomas C. Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers.

Birge, Lucien (1989). "The Grenander Estimator: A Nonasymptotic Approach". In: *The Annals of Statistics* 17.4, pp. 1532–1549.

Birgé, Lucien (June 1997). "Estimation of unimodal densities without smoothness assumptions". In: *Ann. Stat.* 25.3, pp. 970–981.

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.

Brown, Lawrence D (1986). *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. en. IMS.

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

Canu, Stéphane and Alex Smola (Mar. 2006). "Kernel methods and the exponential family". In: *Neurocomputing* 69.7, pp. 714–720.

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.

Casella, George and Roger L Berger (2002). *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA.

Chen, Wenyu, Rahul Mazumder, and Richard J Samworth (May 2021). "A new computational framework for log-concave density estimation". In: arXiv: 2105.11387 [stat.CO].

Chen, Yining and Richard J Samworth (2013). "Smoothed Log-concave Maximum Likelihood Estimation with Applications". In: *Stat. Sin.* 23.3, pp. 1373–1398.

Cook, R Dennis (Feb. 1977). "Detection of Influential Observation in Linear Regression". In: *Technometrics* 19.1, pp. 15–18.

Cook, R Dennis and Sanford Weisberg (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

Cule, M, R Samworth, and M Stewart (2010). "Maximum likelihood estimation of a multi-dimensional log-concave density". In: *Journal of the Royal Statistical Society* 72, pp. 545–607.

Dai, Bo et al. (Nov. 2018). "Kernel Exponential Family Estimation via Doubly Dual Embedding". In: arXiv: 1811.02228 [cs.LG].

Darmois, Georges (1935). "Sur les lois de probabilitéa estimation exhaustive". In: *CR Acad. Sci. Paris* 260.1265, p. 85.

Debruyne, Michiel, Mia Hubert, and Johan A K Suykens (2008). "Model selection in kernel based regression using the influence function". In: *Journal of machine learning research* 9, pp. 2377–2400.

Denkowski, Zdzislaw, Stanislaw Migórski, and Nikolaos S Papageorgiou (Dec. 2013). *An Introduction to Nonlinear Analysis: Theory*. en. Springer Science & Business Media.

Diestel, Joseph and John Jerry Uhl (June 1977). *Vector Measures*. en. American Mathematical Soc.

Doss, Charles R and Jon A Wellner (Apr. 2016). "Global Rates of Convergence of the MLEs of Log-concave and s-concave Densities". en. In: *Ann. Stat.* 44.3, pp. 954–981.

Dümbgen, Lutz, Andre Huesler, and Kaspar Rufibach (July 2007). "Active Set and EM Algorithms for Log-Concave Densities Based on Complete and Censored Data". In: arXiv: 0707.4643 [stat.ME].

Dümbgen, Lutz and Kaspar Rufibach (Feb. 2009). "Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency". en. In: *Bernoulli* 15.1, pp. 40–68.

— (2010). "logcondens: Computations related to univariate log-concave density estimation". In: *Journal of Statistical Software, to appear*.

Dümbgen, Lutz, Richard Samworth, and Dominic Schuhmacher (2011). "Approximation by Log-concave Distributions, with Applications to Regression". In: *Ann. Stat.* 39.2, pp. 702–730.

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

Fisher, Ronald A (1922). "On the mathematical foundations of theoretical statistics". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222.594-604, pp. 309–368.

Freedman, David and Persi Diaconis (Dec. 1981). "On the histogram as a density estimator:L 2 theory". en. In: *Z. Wahrscheinlichkeitstheorie verw Gebiete* 57.4, pp. 453–476.

Fukumizu, Kenji (2005). "Infinite dimensional exponential families by reproducing kernel Hilbert spaces". In: *2nd International Symposium on Information Geometry and its Applications*, pp. 324–333.

Fukumizu, Kenji, Francis R Bach, and Arthur Gretton (2007). "Statistical Consistency of Kernel Canonical Correlation Analysis". In: *J. Mach. Learn. Res.* 8.2.

Fukumizu, Kenji, Francis R Bach, and Michael I Jordan (2004). "Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces". In: *J. Mach. Learn. Res.* 5.Jan, pp. 73–99.

— (Aug. 2009). "Kernel dimension reduction in regression". en. In: *aos* 37.4, pp. 1871–1905.

Fukumizu, Kenji et al. (2007). "Kernel measures of conditional dependence". In: *Adv. Neural Inf. Process. Syst.* 20.

Fukunaga, K and L Hostetler (Jan. 1975). "The estimation of the gradient of a density function, with applications in pattern recognition". In: *IEEE Trans. Inf. Theory* 21.1, pp. 32–40.

Good, I J and R A Gaskins (1971). "Nonparametric Roughness Penalties for Probability Densities". In: *Biometrika* 58.2, pp. 255–277.

Grenander, Ulf (1956). "On the theory of mortality measurement: part ii". In: *Scand. Actuar. J.* 1956.2, pp. 125–153.

Gretton, Arthur et al. (2005). "Kernel Methods for Measuring Independence". In: *Journal of Machine Learning Research* 6.70, pp. 2075–2129. URL: http://jmlr.org/papers/v6/gretton05a.html.

Gretton, Arthur et al. (2012). "A Kernel Two-Sample Test". In: *J. Mach. Learn. Res.* 13.25, pp. 723–773.

Groeneboom, P (1984). *Estimating a Monotone Density*. en. Centrum voor Wiskunde en Informatica.

Gu, Chong (June 1993). "Smoothing Spline Density Estimation: A Dimensionless Automatic Algorithm". In: *J. Am. Stat. Assoc.* 88.422, pp. 495–504.

Gu, Chong and Chunfu Qiu (1993). "Smoothing Spline Density Estimation: Theory". In: *Ann. Stat.* 21.1, pp. 217–234.

Hampel, Frank R (1968). "Contributions to the theory of robust estimation". PhD thesis. University of California.

— (June 1974). "The Influence Curve and its Role in Robust Estimation". In: *J. Am. Stat. Assoc.* 69.346, pp. 383–393.

Hampel, Frank R et al. (1986). *Robust Statistics: The Approach Based on Influence Functions*. en. John Wiley & Sons.

Huber, Peter J. (1964). "Robust Estimation of a Location Parameter". In: *The Annals of Mathematical Statistics* 35.1, pp. 73 –101. DOI: 10.1214/aoms/1177703732. URL: https://doi.org/10.1214/aoms/1177703732.

Huber, Peter J (1967). "The behavior of maximum likelihood estimates under non-standard conditions". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1, pp. 221–233.

Hyvärinen, Aapo (2005). "Estimation of Non-Normalized Statistical Models by Score Matching". In: *J. Mach. Learn. Res.* 6.Apr, pp. 695–709.

Ibragimov, I A (Jan. 1956). "On the Composition of Unimodal Distributions". In: *Theory Probab. Appl.* 1.2, pp. 255–260.

Izenman, Alan J (Mar. 2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. en. Springer Science & Business Media.

Jegelka, Stefanie et al. (2009). "Generalized Clustering via Kernel Embeddings". In: *KI 2009: Advances in Artificial Intelligence*. Springer Berlin Heidelberg, pp. 144–152.

Kim, Arlene K H and Richard J Samworth (Dec. 2016). "Global rates of convergence in log-concave density estimation". en. In: *aos* 44.6, pp. 2756–2779.

Kimeldorf, George and Grace Wahba (Jan. 1971). "Some results on Tchebycheffian spline functions". In: *J. Math. Anal. Appl.* 33.1, pp. 82–95.

Koenker, Roger and Ivan Mizera (Mar. 2007). "Density Estimation by Total Variation Regularization". In: *Advances in Statistical Modeling and Inference*. Vol. 3. Series in Biostatistics. World Scientific, pp. 613–633.

Koh, Pang Wei and Percy Liang (2017). "Understanding Black-box Predictions via Influence Functions". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1885–1894.

Koopman, B O (1936). "On Distributions Admitting a Sufficient Statistic". In: *Trans. Amer. Math. Soc.* 39.3, pp. 399–409.

Le, Song (2008). *Learning Via Hilbert Space Embedding of Distributions*. en. University of Sydney.

Leonard, Tom (Jan. 1978). "Density Estimation, Stochastic Processes and Prior Information". In: *J. R. Stat. Soc. Series B Stat. Methodol.* 40.2, pp. 113–132.

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.

McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models, Second Edition*. Monographs on Statistics and Applied Probability Series. Chapman & Hall. ISBN: 9780412317606.

Muandet, Krikamol et al. (2016). "Kernel mean shrinkage estimators". In: *J. Mach. Learn. Res.* 17.1, pp. 1656–1696.

Muandet, Krikamol et al. (2017). "Kernel Mean Embedding of Distributions: A Review and Beyond". In: *Foundations and Trends® in Machine Learning* 10.1-2, pp. 1–141.

Pal, Jayanta Kumar, Michael Woodroofe, and Mary Meyer (2007). "Estimating a Polya Frequency Function$_2$". In: *Lect. Notes Monogr. Ser.* 54, pp. 239–249.

Parry, Matthew, A Philip Dawid, and Steffen Lauritzen (Feb. 2012). "Proper local scoring rules". en. In: *Ann. Stat.* 40.1, pp. 561–592.

Pitman, E J G (Dec. 1936). "Sufficient statistics and intrinsic accuracy". In: *Math. Proc. Cambridge Philos. Soc.* 32.4, pp. 567–579.

Raj, Anant et al. (Oct. 2020). "Model-specific Data Subsampling with Influence Functions". In: arXiv: 2010.10218 [cs.LG].

Rao, B L S Prakasa (1969). "Estimation of a Unimodal Density". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 31.1, pp. 23–36.

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.

Rathke, Fabian and Christoph Schnörr (Dec. 2019). "Fast multivariate log-concave density estimation". In: *Comput. Stat. Data Anal.* 140, pp. 41–58.

Reed, Michael and Barry Simon (Dec. 2012). *Methods of Modern Mathematical Physics: Functional Analysis*. en. Elsevier. ISBN: 9780125850506.

Ripley, Brian D. (1996). *Pattern Recognition and Neural Networks*. en. Cambridge University Press. DOI: 10.1017/CBO9780511812651.

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

Rufibach, Kaspar (July 2007). "Computing maximum likelihood estimators of a log-concave density function". In: *J. Stat. Comput. Simul.* 77.7, pp. 561–574.

Samworth, Richard J (Nov. 2018). "Recent Progress in Log-Concave Density Estimation". en. In: *Stat. Sci.* 33.4, pp. 493–509.

Sardy, Sylvain and Paul Tseng (Jan. 2010). "Density Estimation by Total Variation Penalized Likelihood Driven by the Sparsity ℓ_1 Information Criterion: Total variation density estimation". In: *Scand. Stat. Theory Appl.* 37.2, pp. 321–337.

Schölkopf, Bernhard, Ralf Herbrich, and Alex J Smola (2001). "A Generalized Representer Theorem". In: *Computational Learning Theory*. Springer Berlin Heidelberg, pp. 416–426.

Schölkopf, Bernhard, Alexander Smola, and Klaus-Robert Müller (July 1998). "Nonlinear Component Analysis as a Kernel Eigenvalue Problem". In: *Neural Comput.* 10.5, pp. 1299–1319.

Silverman, B W (Sept. 1982). "On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method". en. In: *Ann. Stat.* 10.3, pp. 795–810.

— (1986). *Density Estimation for Statistics and Data Analysis*. Boston, MA: Chapman and Hall. ISBN: 978-0412246203.

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.

Smola, Alex et al. (2007). "A Hilbert Space Embedding for Distributions". In: *Algorithmic Learning Theory*. Springer Berlin Heidelberg, pp. 13–31.

Song, Le, Kenji Fukumizu, and Arthur Gretton (2013). *Kernel Embeddings of Conditional Distributions: A Unified Kernel Framework for Nonparametric Inference in Graphical Models*.

Song, Le, Arthur Gretton, and Carlos Guestrin (2010). "Nonparametric Tree Graphical Models". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, pp. 765–772.

Song, Le et al. (Jan. 2014). "Nonparametric Latent Tree Graphical Models: Inference, Estimation, and Structure Learning". In: arXiv: 1401.3940 [stat.ML].

Sriperumbudur, Fukumizu, and others (2011). "Universality, Characteristic Kernels and RKHS Embedding of Measures". In: *Journal of Machine Learning Research* 12.70, pp. 2389–2410.

Sriperumbudur et al. (2010). "Hilbert space embeddings and metrics on probability measures". In: *Journal of Machine Learning Research* 11, pp. 1517–1561.

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.

Steinwart, Ingo and Andreas Christmann (Sept. 2008). *Support Vector Machines*. en. Springer Science & Business Media.

Sun, Siqi, Mladen Kolar, and Jinbo Xu (2015). "Learning structured densities via infinite dimensional exponential families". In: *Advances in Neural Information*

Processing Systems. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., pp. 2287 –2295.

Sutherland, Dougal J et al. (May 2017). "Efficient and principled score estimation with Nyström kernel exponential families". In: arXiv: 1705.08360 [stat.ML].

Ting, Daniel and Eric Brochu (2018). "Optimal Subsampling with Influence Functions". In: *Adv. Neural Inf. Process. Syst.* 31.

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

Wainwright, Martin J and Michael I Jordan (2008). "Graphical Models, Exponential Families, and Variational Inference". In: *Foundations and Trends® in Machine Learning* 1.1–2, pp. 1–305.

Walther, Guenther (June 2002). "Detecting the Presence of Mixing with Multiscale Maximum Likelihood". In: *J. Am. Stat. Assoc.* 97.458, pp. 508–513.

Wasserman, Larry (Sept. 2006). *All of Nonparametric Statistics*. en. Springer Science & Business Media.

Wegman, Edward J (1970a). "Maximum Likelihood Estimation of a Unimodal Density Function". In: *Ann. Math. Stat.* 41.2, pp. 457–471.

— (1970b). "Maximum Likelihood Estimation of a Unimodal Density, II". In: *Ann. Math. Stat.* 41.6, pp. 2169–2174.

White, Halbert (1982). "Maximum Likelihood Estimation of Misspecified Models". In: *Econometrica* 50.1, pp. 1–25.

Williams, Christopher K I and Matthias Seeger (2001). "Using the Nyström Method to Speed Up Kernel Machines". In: *Advances in Neural Information Processing*

Systems 13. Ed. by T K Leen, T G Dietterich, and V Tresp. MIT Press, pp. 682–688.

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

Yu, Shiqing, Mathias Drton, and Ali Shojaie (Sept. 2020). "Generalized Score Matching for General Domains". In: arXiv: 2009.11428 [stat.ME].

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

Zhou, Ding-Xuan (Oct. 2008). "Derivative reproducing properties for kernel methods in learning theory". In: *J. Comput. Appl. Math.* 220.1, pp. 456–463.