Journal of Machine Learning Research 23 (2022) 1-38 Submitted 3/21, Revised 3/22; Published 7/22

Estimating Density Models with Truncation Boundaries using Score Matching

Song Liu song.liu@bristol.ac.uk
University of Bristol

Takafumi Kanamori kanamori@c.titech.ac.jp
Tokyo Institute of Technology,
RIKEN AIP

Daniel J. Williams daniel.williams@bristol.ac.uk
University of Bristol

Editor: Qiang Liu

Abstract

Truncated densities are probability density functions defined on truncated domains. They share the same parametric form with their non-truncated counterparts up to a normalizing constant. Since the computation of their normalizing constants is usually infeasible, Maximum Likelihood Estimation cannot be easily applied to estimate truncated density models. Score Matching (SM) is a powerful tool for fitting parameters using only unnormalized models. However, it cannot be directly applied here as boundary conditions that derive a tractable SM objective are not satisfied by truncated densities. This paper studies parameter estimation for truncated probability densities using SM. The estimator minimizes a weighted Fisher divergence. The weight function is simply the shortest distance from a data point to the domain's boundary. We show this choice of weight function naturally arises from minimizing the Stein discrepancy and upper bounding the finite-sample estimation error. We demonstrate the usefulness of our method via numerical experiments and a study on the Chicago crime data set. We also show that the proposed density estimation can correct the outlier-trimming bias caused by aggressive outlier detection methods.

Keywords: score matching, truncated density estimation, unnormalized density models, stein operator, non-asymptotic analysis

1. Introduction

In many applications, we cannot observe a problem's "full picture". Instead, our observation window is limited, so we can only observe a truncated data set. For example, a police department can only monitor crimes within their city's boundary, although crimes do not automatically stop at an artificial border. Similarly, we can only observe some geolocation tracking data up to the coverage of mobile signals. Data sets like these are skewed representations of actual activities due to truncation. In many cases, these truncation boundaries can be very complex. For example, the boundary of the city of Chicago is a complex polygon (see Figure 1) which cannot be easily approximated by a bounding box or circle.

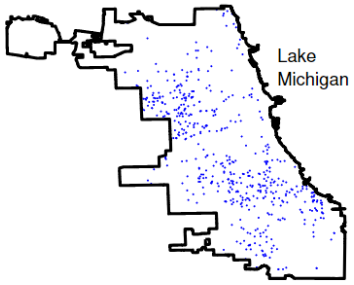


Figure 1: Boundary of Chicago, where blue dots are locations of homicides in 2008.

The key challenge of estimating parameters in truncated densities is that the normalizing constant is not computationally tractable. The normalizing constant ensures that the integration of a density function equals one over its input domain. As the normalization takes place in an irregular bounded domain in $\mathbb{R}^d$, the integration does not have a closed-form in general. This integration creates a computational issue since the classic Maximum Likelihood Estimation (MLE) requires the evaluation of such a normalizing constant. Although the normalizing constant in the likelihood function can be approximated using Monte Carlo methods (Geyer, 1994), it is hard to guarantee the approximation accuracy when the data is in high dimensional space, and the truncation domain is complex.

Recent years have seen a new class of estimators called Score Matching (SM) (Hyvärinen, 2005, 2007; Lyu, 2009) rise in popularity. They estimate parameters by minimizing the Fisher-Hyvärinen divergence (Lyu, 2009). This divergence is defined using the difference between the gradients of log model density and log data density. The gradients are taken with respect to the *input variable*, and the normalizing constant is eliminated and not involved in the estimation procedure. Thus SM is a natural candidate for estimating truncated density models.

However, the original SM cannot work for estimating these truncated models, as the regularity condition used to derive the tractable objective function is not satisfied. Hyvärinen (2007) and Yu et al. (2019) proposed *generalized SM* to handle the distributions on the non-negative orthant $\mathbb{R}^d_+$. In generalized SM, a weight function is introduced to satisfy the boundary condition required for deriving the tractable objective function. Promising results have been observed on high-dimensional non-negative graphical model structure estimation (Yu et al., 2019). When the density is truncated in a dimension-wise manner with a lower and upper bound, this problem is known as a doubly truncated distribution estimation (Turnbull, 1976; Moreira and de Uña-Álvarez, 2012). Though some works have tackled the problem of estimating truncated multivariate Gaussian densities (Daskalakis et al., 2018, 2019), little work has been done for estimating a wide range of density models with complicated truncation domains. Note that if the truncation domain is simple, it is possible to apply the change of variable rule to our data set so that the truncated density estimation becomes a non-truncated (or doubly truncated) density estimation problem in the new domain. This technique has been widely used in modelling distributions on a sphere (Mardia and Jupp, 2009). However, this technique is not applicable when the truncation domain is complex, such as the one in the Chicago Crime data set.

This paper proposes a novel estimator for truncated density models using generalized SM. Our choice of the weight function is the shortest distance from the data point to the truncation boundary. We show that this choice naturally arises from minimizing a Stein Discrepancy and lowering the finite sample estimation error upperbound. We can efficiently compute this distance for many complicated domains.

We demonstrate the usefulness of our method via numerical experiments and an application to the Chicago crime data set. Finally, we apply this estimator to correct the outlier trimming bias caused by One-class Support Vector Machines.

2. Problem Formulation

Denote a probability density function parameterized by $\boldsymbol{\theta}$ over the domain $V \subset \mathbb{R}^d$ as $p_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{x} \in V$. Without loss of generality, we can write

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) := \frac{\bar{p}_{\boldsymbol{\theta}}(\boldsymbol{x})}{Z_V(\boldsymbol{\theta})}, Z_V(\boldsymbol{\theta}) := \int_V \bar{p}_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

where $\bar{p}_{\boldsymbol{\theta}}$ is an unnormalized model and $Z_V(\boldsymbol{\theta})$ is the normalizing constant so that $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is integrated to 1 over its domain V. The domain $V \subset \mathbb{R}^d$ may be a complicated bounded domain e.g., a polytope. In such cases, $Z_V(\boldsymbol{\theta})$ may not have a closed-form and cannot be efficiently evaluated. For example, let $\bar{p}_{\boldsymbol{\theta}}$ be a Gaussian mixture model restricted in a generic polygon, then $Z_V(\boldsymbol{\theta})$ does not have a closed-form expression.

Our task is classic statistical model estimation: Suppose that V is given, we want to estimate the parameter $\boldsymbol{\theta}$ in $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ using $X_q = \{\boldsymbol{x}_i\}_{i=1}^n$, which is a set of observed i.i.d. samples from a truncated data generating distribution Q with an unknown probability density function $q(\boldsymbol{x}), \boldsymbol{x} \in V$.

The challenge comes from the fact that we need to obtain the estimator of $\boldsymbol{\theta}$ using only the unnormalized density model $\bar{p}_{\boldsymbol{\theta}}(\boldsymbol{x})$: It is not straightforward to calculate $Z_V(\boldsymbol{\theta})$ for a complicated V. Therefore, MLE, which requires evaluating the normalizing constant, cannot be performed easily. A popular tool for estimating unnormalized densities is Score Matching (SM) proposed by Hyvärinen (2005). We now introduce SM and its extension (Yu et al., 2019), then explain why it cannot be readily used for estimating complicated truncated densities.

Notation: The finite set $\{1, \ldots, d\}$ for a positive integer d is denoted by $[d]$. Given a vector $\boldsymbol{x}$, let x_k denote the k-th element of $\boldsymbol{x}$. Let $\langle \cdot, \cdot \rangle$ be the standard inner product, and the Euclidean norm of the vector $\boldsymbol{a}$ is denoted as $\|\boldsymbol{a}\| = \sqrt{\langle \boldsymbol{a}, \boldsymbol{a} \rangle}$. The ℓ^p-norm of $\boldsymbol{x}$ is denoted by $\|\boldsymbol{x}\|_p$. Thus, $\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_2$ holds. Let ∂_k for $k \in [d]$ be the partial differential operator $\frac{\partial}{\partial x_k}$ and $\nabla_{\boldsymbol{x}}$ be $(\partial_1, \ldots, \partial_d)$ to the function $f(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^d$. Similarly, the gradient operator with respect to the parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^r$ of the statistical model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is denoted by $\nabla_{\boldsymbol{\theta}} = (\frac{\partial}{\partial \theta_1}, \ldots, \frac{\partial}{\partial \theta_r})$. $\mathbb{E}_q[f(\boldsymbol{x})]$ stands for the expectation of $f(\boldsymbol{x})$ with respect to the probability density $q(\boldsymbol{x})$. To reduce the clutter, we sometimes shorten $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $q(\boldsymbol{x})$ as $p_{\boldsymbol{\theta}}$ and q wherever such abbreviations do not lead to confusion. In addition, the log-likelihood $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is expressed by $\ell_{\boldsymbol{\theta}}(\boldsymbol{x})$. $\|f\|_{L^2(Q)} := \left(\mathbb{E}_q\left[f(\boldsymbol{x})^2\right]\right)^{1/2}$ where Q denotes a distribution with density function $q(\boldsymbol{x})$. Let $\boldsymbol{e}_i, i = 1, \ldots, d$ be the unit vector along the i-th axis in $\mathbb{R}^d$, i.e., $\boldsymbol{e}_1 = (1, 0, \ldots, 0)$, etc. We use $\overline{V}$ to represent the closure of a bounded open set V.

3. Score Matching and Its Generalization

In this section, we introduce classic SM (Hyvärinen, 2005) and one of its variants (Yu et al., 2019).

Definition 1 *The Fisher-Hyvärinen (FH) divergence (Lyu, 2009) between q and $p_{\boldsymbol{\theta}}$ is defined as*

$$\mathrm{FH}(q, p_{\boldsymbol{\theta}}) := \mathbb{E}_q[\|\nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log q(\boldsymbol{x})\|^2].$$

Suppose $q(\boldsymbol{x}) > 0$ on $\mathbb{R}^d$. Then, the FH divergence is non-negative and it vanishes if and only if $p_{\boldsymbol{\theta}}(\boldsymbol{x}) = q(\boldsymbol{x})$ almost surely. When a density model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is defined on $V = \mathbb{R}^d$, SM finds an estimate of $\boldsymbol{\theta}$ by minimizing $\mathrm{FH}(q, p_{\boldsymbol{\theta}})$ over the parameter space $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^r$, i.e.,

$$\begin{aligned}\boldsymbol{\theta}_{\mathrm{SM}} &:= \operatorname*{argmin}_{\boldsymbol{\theta}} \mathrm{FH}(q, p_{\boldsymbol{\theta}}) \\ &= \operatorname*{argmin}_{\boldsymbol{\theta}} \mathbb{E}_q\left[\|\nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}\|^2\right] - 2\mathbb{E}_q\left[\langle \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}, \nabla_{\boldsymbol{x}} \log q\rangle\right] + \mathrm{C}, \end{aligned} \tag{1}$$

where C is a constant independent of $\boldsymbol{\theta}$. The key advantage of SM is that the normalizing constant $Z_V(\boldsymbol{\theta})$ is not required when evaluating (1) as $\nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log \bar{p}_{\boldsymbol{\theta}}(\boldsymbol{x})$. Thus, SM is widely used for estimating "unnormalizable" statistical models.

Unfortunately, (1) is not tractable as we cannot directly evaluate the second term of (1) without access to $\nabla_{\boldsymbol{x}} \log q$. Using the integration by parts rule, however, we find that the equality,

$$\mathbb{E}_q[\langle \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}, \nabla_{\boldsymbol{x}} \log q\rangle] = -\sum_{k=1}^{d} \mathbb{E}_q[\partial_k^2 \log p_{\boldsymbol{\theta}}]$$

holds under the smoothness condition of $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $\log q(\boldsymbol{x})$ w.r.t. $\boldsymbol{x}$ and the boundary condition

$$\lim_{|x_k| \to \infty} q(\boldsymbol{x}) \partial_k \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = 0, \forall k \in [d]. \tag{2}$$

Many density functions defined on $\mathbb{R}^d$, such as multivariate Gaussian or Gaussian mixture, satisfy these conditions. See the study by Hyvärinen (2005, 2007) for details. Thus, FH-divergence in (1) can be re-written as

$$\mathrm{FH}(q, p_{\boldsymbol{\theta}}) = \mathbb{E}_q[\|\nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}\|^2] + 2\sum_{k=1}^{d} \mathbb{E}_q[\partial_k^2 \log p_{\boldsymbol{\theta}}] + C \tag{3}$$

where the objective only relies on q through the expectations which can be approximated by the empirical mean over the observed samples X_q.

When $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is defined on the *truncated* subset in $\mathbb{R}^d$ such as the *non-negative orthant*,

$$\mathbb{R}^d_+ := \{\boldsymbol{x} \in \mathbb{R}^d | x_k \geq 0, \forall k \in [d]\},$$

the boundary condition required to apply the integration by parts rule (i.e., (2)) no longer holds for many density functions such as Gaussian or Gaussian mixtures. To estimate

parameters of density functions on the non-negative orthant, Hyvärinen (2007); Yu et al. (2019) introduced generalized SM:

$$\begin{aligned}\boldsymbol{\theta}_{\mathrm{GSM}} &:= \operatorname*{argmin}_{\boldsymbol{\theta}} \mathrm{FH}_{\boldsymbol{g}}(q, p_{\boldsymbol{\theta}}),\\ \mathrm{FH}_{\boldsymbol{g}}(q, p_{\boldsymbol{\theta}}) &:= \mathbb{E}_q[\|\boldsymbol{g}^{1/2} \circ \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}} - \boldsymbol{g}^{1/2} \circ \nabla_{\boldsymbol{x}} \log q\|^2],\\ &= \sum_{k=1}^{d} \mathbb{E}_q[g_k \cdot (\partial_k \log p_{\boldsymbol{\theta}})^2] - 2\sum_{k=1}^{d} \mathbb{E}_q[g_k \cdot (\partial_k \log p_{\boldsymbol{\theta}})(\partial_k \log q)] + C, \qquad (4)\end{aligned}$$

where $\boldsymbol{g}(\boldsymbol{x}) := (g_1(\boldsymbol{x}), \ldots, g_d(\boldsymbol{x})) \in \mathbb{R}^d$ is a non-negative valued, continuously differentiable function with $\boldsymbol{g}(\mathbf{0}) = \mathbf{0}$. $\boldsymbol{g}^{1/2}$ is the element-wise square root operation applied on $\boldsymbol{g}$ and $\circ$ is the element-wise product. Examples of $\boldsymbol{g}$ include $g_k(\boldsymbol{x}) = x_k$ or $g_k(\boldsymbol{x}) = \max(x_k, 1), \forall k \in [d]$ given $\boldsymbol{x} \in \mathbb{R}^d_+$.

We can derive the tractable objective function from (4) with the help of a lemma, which appeared in the proof of Theorem 3 in Yu et al. (2019). Here we restate it using our symbols for convenience.

Lemma 1 *Suppose that* $\log q(\boldsymbol{x})$ *and* $\boldsymbol{g}(\boldsymbol{x})$ *are continuously differentiable almost everywhere (a.e.) on* $\mathbb{R}^d_+$ *and* $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ *is twice continuously differentiable with respect to* $\boldsymbol{x}$ *on* $\mathbb{R}^d_+$. *Furthermore, we assume the boundary condition,*

$$\lim_{|x_k| \to 0+, |x_k| \to \infty} g_k(\boldsymbol{x}) q(\boldsymbol{x}) \partial_k \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = 0$$

for $k \in [d]$. *Then,*

$$\sum_{i=1}^{d} \mathbb{E}_q[g_k \cdot (\partial_k \log p_{\boldsymbol{\theta}})(\partial_k \log q)] = -\sum_{k=1}^{d} \mathbb{E}_q[\partial_k (g_k \partial_k \log p_{\boldsymbol{\theta}})].$$

Readers can find the proof in Section A.1 in (Yu et al., 2019) which uses dimension-wise integration by parts. Using Lemma 1, the generalized SM objective (4) has a tractable expression,

$$\mathrm{FH}_{\boldsymbol{g}}(q, p_{\boldsymbol{\theta}}) = \sum_{k=1}^{d} \mathbb{E}_q[g_k \cdot (\partial_k \log p_{\boldsymbol{\theta}})^2] + 2\sum_{k=1}^{d} \mathbb{E}_q[\partial_k (g_k \partial_k \log p_{\boldsymbol{\theta}})] + C \qquad (5)$$

where C is a constant independent of $\boldsymbol{\theta}$. One can replace the expectation with the empirical mean over X_q to obtain an unbiased estimator of the above objective function. It is straightforward to modify the generalized SM formulation to work for doubly truncated distributions. For example, $g_k(\boldsymbol{x}) = \min\{x_k - a_k, b_k - x_k\}$ can be used to estimate truncated densities on the product space $\prod_{k=1}^{d}(a_k, b_k)$.

It is worth pointing out that Lemma 1 is a specification of the divergence theorem such as Green's or Stokes' theorem, which usually deals with a bounded domain. Recently, Mardia et al. (2016) studied an SM objective based on Stokes' theorem for estimating densities on a Riemannian manifold.

We intend to extend the generalized SM to a generic truncated domain. We now show the validity of the objective function (4) when considering a generic domain V. For two differantiable probability densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ with respect to the Lebesgue measure $\mu(\cdot)$ on $V \subset \mathbb{R}^d$, the following lemma holds:

Lemma 2 *Suppose that V is a* ***connected open subset*** *in $\mathbb{R}^d$ and that $g_k(\boldsymbol{x}) > 0$ and $q(\boldsymbol{x}) > 0, \forall \boldsymbol{x} \in V$. Then, $\mathrm{FH}_{\boldsymbol{g}}(q, p) = 0$ if and only if $p = q$ for the probability density functions p and q.*

Proof First, it is easy to see that $p = q \implies \mathrm{FH}_{\boldsymbol{g}}(p, q) = 0$. Second, as $\mathrm{FH}_{\boldsymbol{g}}(q, p) = 0$, $g_k(\boldsymbol{x})q(\boldsymbol{x})\|\nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})\|^2 = 0$ should hold a.e. with respect to μ on V. Thus, we have $\nabla_{\boldsymbol{x}}(\log q(\boldsymbol{x}) - \log p(\boldsymbol{x})) = \mathbf{0}$ on V. As V is connected, $\log q(\boldsymbol{x}) - \log p(\boldsymbol{x})$ should be a constant independent of $\boldsymbol{x}$ on V. Hence, $p(\boldsymbol{x})$ is proportional to $q(\boldsymbol{x})$ on V. As both are normalized probability density functions, we have $p = q$. ■

The following theorem states that the minimizer of $\mathrm{FH}_{\boldsymbol{g}}(q, p_{\boldsymbol{\theta}})$ is unique and is the true parameter under a mild identifiability condition.

Theorem 1 *Suppose all assumptions stated in Lemma 2 holds. Let $\mathcal{P} = \{p_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^r\}$ be a set of parametric statistical models on the connected open subset V in $\mathbb{R}^d$. If the model $p_{\boldsymbol{\theta}}$ is identifiable in the sense that $\mu(\{\boldsymbol{x} \,|\, p_{\boldsymbol{\theta}}(\boldsymbol{x}) \neq p_{\boldsymbol{\theta}'}(\boldsymbol{x})\}) > 0$ holds for $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$, and $q = p_{\boldsymbol{\theta}_0} \in \mathcal{P}$, then $\operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \mathrm{FH}_{\boldsymbol{g}}(q, p_{\boldsymbol{\theta}})$ is unique and is $\boldsymbol{\theta}_0$.*

Proof Given the assumption on q, we have $\mathrm{FH}_{\boldsymbol{g}}(q, p_{\boldsymbol{\theta}_0}) = 0$. Thus, $\boldsymbol{\theta}_0$ is a minimizer of $\mathrm{FH}_{\boldsymbol{g}}(q, p_{\boldsymbol{\theta}})$. Moreover, Lemma 2 guarantees that $\mathrm{FH}_{\boldsymbol{g}}(p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}'}) > 0$ holds for $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$. We can prove this by the contradiction. Suppose $\mathrm{FH}_{\boldsymbol{g}}(p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}'}) = 0$ for $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$. Thus, Lemma 2 states $p_{\boldsymbol{\theta}}(\boldsymbol{x}) = p_{\boldsymbol{\theta}'}(\boldsymbol{x})$ must hold on V. This contradicts $\mu(\{\boldsymbol{x} \,|\, p_{\boldsymbol{\theta}}(\boldsymbol{x}) \neq p_{\boldsymbol{\theta}'}(\boldsymbol{x})\}) > 0$. Hence, for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, $\mathrm{FH}_{\boldsymbol{g}}(p_{\boldsymbol{\theta}_0}, p_{\boldsymbol{\theta}}) = \mathrm{FH}_{\boldsymbol{g}}(q, p_{\boldsymbol{\theta}}) > 0$. This leads to the conclusion that the minimizer $\boldsymbol{\theta}_0$ is unique. ■

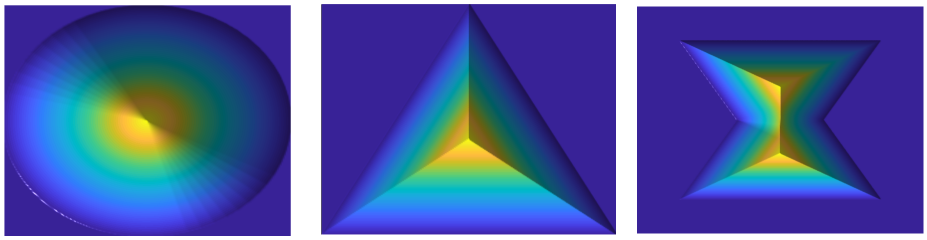
Remark 1 *When the domain is not connected, $\mathrm{FH}_{\boldsymbol{g}}(q, p) = 0$ implies that p is proportional to q on each connected region. However, for common statistical models, the proportional relationship on each connected region implies that the probability densities are the same. For instance, let us consider the truncated Gaussian model $p_{\boldsymbol{\theta}}$ on V that is the disjoint union of V_1 and V_2, where $\boldsymbol{\theta}$ is the mean parameter of the Gaussian distribution $\bar{p}_{\boldsymbol{\theta}}$ on $\mathbb{R}^d$. If $p_{\boldsymbol{\theta}}/p_{\boldsymbol{\theta}_0}$ is a constant function on a small open subset in V, we see that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ holds. Hence, Theorem 1 holds for such models even when the domain V is not connected.*

When using generalized SM on a generic domain V, the more significant challenge is selecting a weight function $\boldsymbol{g}$. First, we explain an intuitive way to construct $\boldsymbol{g}$. Then we show this intuitive choice is theoretically sound and empirically effective.

Denote the boundary of V as ∂V. We can design a weight function g_k taking 0 at ∂V, then hopefully an analogue to Lemma 1 would hold, giving a tractable form of the estimator. For example, we can consider a $\mathrm{dist}(\boldsymbol{x}, \boldsymbol{z})$, which is a distance function between $\boldsymbol{x}$ and $\boldsymbol{z}$, and let the weight function g_k be

$$g_k = g_0 := \min_{\boldsymbol{z} \in \partial V} \mathrm{dist}(\boldsymbol{x}, \boldsymbol{z}), \quad \forall k \in [d],\ \boldsymbol{x} \in \overline{V}. \tag{6}$$

i.e., the distance from $\boldsymbol{x}$ to the truncation boundary ∂V. See Figure 2 for examples of g_0 defined on some simple bounded domains V. This weight function is intuitive and can be easily computed for many complex truncation boundaries (See Section 7 for details).



(a) (b) (c)

Figure 2: Examples of $g_0(\boldsymbol{x})$ for (a) a circular boundary (b) a triangular boundary and (c) a polygon boundary. Here $\text{dist}(\cdot, \cdot)$ is the Euclidean distance function.

Since $g_0(\boldsymbol{x}) > 0, \forall \boldsymbol{x} \in V$ by construction, Theorem 1 guarantees the minimizer of the generalized SM objective is the true parameter as long as V is a connected open domain and $p_{\boldsymbol{\theta}}$ is correctly specified.

However, there are a few major concerns:

1. It is unclear whether letting $g_k = g_0$ would entail a tractable objective function. Lemma 1 is only derived for the non-negative orthant and it assumes $g_k(\boldsymbol{x})$ to be continuous differentiable a.e.. However, the distance weight g_0 is not necessarily differentiable a.e. on V.

2. The efficacy of the distance weight g_0 is unclear. The weight function g_k can be any function that satisfies the boundary condition ($\boldsymbol{g}$ taking $\mathbf{0}$ at ∂V) and is positive and differentiable on V. It is unclear why using g_0 as the weight function would yield good statistical estimation performance.

The following sections address these two concerns from theoretical and empirical perspectives. For simplicity, we refer to generalized SM with $g_k = g_0$ as truncated SM, or *TruncSM* for short.

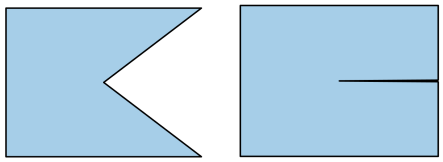
4. Tractable Truncated Score Matching Objective

Theorem 1 states that when V is a connected open domain, TruncSM is a valid density estimation criterion. Now we will show when V is a special kind of connected and open domain, the TruncSM objective is computationally tractable.

The key step of deriving a tractable SM objective is to show that the distance weight g_0 is *weakly differentiable*, so an analogue to Lemma 1 can be proven. Let us formally define the notion of a weakly differentiable function via Sobolev-Hilbert space:

Definition 2 (Sobolev-Hilbert space) *: Let $L^2(V)$ be the L^2 space on $V \subset \mathbb{R}^d$ endowed with the Lebesgue measure. Then, $H^1(V)$ is the Sobolev-Hilbert space space defined by*

$$H^1(V) = \left\{ f \in L^2(V) \,\middle|\, \|f\|^2_{L^2(V)} + \sum_k \|D_k f\|^2_{L^2(V)} < \infty \right\},$$



(a) Lipschitz domain (b) Crack domain

Figure 3: Lipschitz domain vs. Crack domain

where D_k is the weak derivative corresponding to ∂_k and $\|f\|_{L^2(V)} = \sqrt{\int_V |f(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}}$.

In this paper, we focus on a special type of open and connected domain, called *Lipschitz domain*. In engineering applications, most domains are Lipschitz domains. Intuitively speaking, a Lipschitz domain $V \subset \mathbb{R}^d$ is a bounded connected open domain whose local boundary is a level set of some Lipschitz function. However, it does not include domains with "cracks" (see Figure 3). All polytopes are Lipschitz domains.

Definition 3 (Lipschitz Domain) *Let V be an open and bounded domain in $\mathbb{R}^d$. We say V is a **Lipschitz domain** if for any $\boldsymbol{x} \in \partial V$, there exists an $r > 0$ and a Lipschitz function $f(x_1, \ldots, x_{d-1})$ such that $V \cap B(\boldsymbol{x}, r)$ is expressed by $\{\boldsymbol{z} \in B(\boldsymbol{x}, r) \,|z_d > f(z_1, \ldots, z_{d-1})\}$ upon a transformation of the coordinate system if necessary. $B(\boldsymbol{x}, r)$ denotes a d-dimensional ball centered at $\boldsymbol{x}$ with radius r.*

The full definition of the weak derivative and more details about Lipschitz domains and Sobolev-Hilbert spaces can be found in Section 7 of (Atkinson and Han, 2005).

We prove the following lemma which states that g_0 defined on a *Lipschitz domain* is in $H^1(V)$, thus is *weakly differentiable*. The proof can be found in Appendix B.

Lemma 3 *Suppose V is a Lipschitz domain, then $g_0 \in H^1(V)$.*

The classic Green's and Stokes' theorem that proved Lemma 1 could not be applied to derive a tractable objective anymore as functions in $H^1(V)$ are *not* differentiable in a classic sense. However, there exists an extension of Green's theorem of weakly differentiable functions (Proposition 7.6.1 in Atkinson and Han, 2005).

Lemma 4 (Extended Green's Theorem) *For the Lipschitz domain $V \subset \mathbb{R}^d$, suppose $f_1, f_2 \in H^1(V)$,*

$$\int_V f_1(\boldsymbol{x}) \partial_k f_2(\boldsymbol{x}) d\boldsymbol{x} = \int_{\partial V} f_1(\boldsymbol{x}) f_2(\boldsymbol{x}) \nu_k(\boldsymbol{x}) ds - \int_V f_2(\boldsymbol{x}) \partial_k f_1(\boldsymbol{x}) d\boldsymbol{x}, \quad \forall k \in [d],$$

where $(\nu_1, \ldots, \nu_d)$ is the unit outward normal vector on ∂V and ds is the surface element on ∂V.

Now we apply Lemma 3 and 4 to obtain a tractable form of TruncSM objective:

Theorem 2 *Assume $V \subset \mathbb{R}^d$ is a Lipschitz domain. Suppose $q, \partial_k \log p_{\boldsymbol{\theta}} \in H^1(V)$ and that for any $\boldsymbol{z} \in \partial V$ it holds that*

$$\lim_{\boldsymbol{x} \to \boldsymbol{z}} q(\boldsymbol{x}) g_0(\boldsymbol{x}) \partial_k \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \nu_k(\boldsymbol{z}) = 0, \forall k \in [d],$$

where $\boldsymbol{x} \to \boldsymbol{z}$ takes any point sequence converging to $\boldsymbol{z} \in \partial V$ into account. Then, we have

$$\sum_{i=1}^{d} \mathbb{E}_q[g_0 \cdot (\partial_k \log p_{\boldsymbol{\theta}})(\partial_k \log q)] = -\sum_{k=1}^{d} \mathbb{E}_q[\partial_k (g_0 \partial_k \log p_{\boldsymbol{\theta}})].$$

The proof of the above theorem can be found in Appendix C. This indicates that TruncSM indeed has a tractable objective function.

$$\begin{aligned}
\boldsymbol{\theta}_{\mathrm{TSM}} &:= \operatorname*{argmin}_{\boldsymbol{\theta}} \mathrm{FH}_{g_0}(q, p_{\boldsymbol{\theta}}), \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{k=1}^{d} \mathbb{E}_q[g_0 \cdot (\partial_k \log p_{\boldsymbol{\theta}})^2] + 2 \sum_{k=1}^{d} \mathbb{E}_q[\partial_k (g_0 \partial_k \log p_{\boldsymbol{\theta}})] \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{k=1}^{d} \mathbb{E}_q[g_0 \cdot (\partial_k \log p_{\boldsymbol{\theta}})^2] + 2 \sum_{k=1}^{d} \mathbb{E}_q[g_0 \partial_k^2 \log p_{\boldsymbol{\theta}}] + 2 \sum_{k=1}^{d} \mathbb{E}_q[\partial_k g_0 \cdot \partial_k \log p_{\boldsymbol{\theta}}],
\end{aligned} \tag{7}$$

where each expectation can be approximated using samples from the data set $X_q \sim Q$.

Although (7) is similar to the generalized SM objective (5) (it replaces g_k with g_0), this result cannot be taken for granted. It highlights an important restriction of TruncSM: V needs to be a Lipschitz domain. This constraint is required to ensure the weak differentiability of g_0, and is connected with the Lipschitzness of g_0. This result also suggests that we may be able to bypass this constraint when using different weight functions in generalized SM. The study along this line can be interesting for future work.

Now we turn our focus to the efficacy of TruncSM. In the next section, we show TruncSM is a Minimum Stein Discrepancy Estimator (Barp et al., 2019).

5. Truncated Score Matching as Minimum Stein Discrepancy Estimator

Maximum Stein Discrepancies are a family of discrepancies that measure the differences between two distributions (Chwialkowski et al., 2016; Liu et al., 2016). For simplicity, we assume that q and $\ell_{\boldsymbol{\theta}}$ are smooth in this section.

Definition 4 *Given an $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^d$, a **Stein operator** is defined as*

$$T_p \boldsymbol{f} := \sum_{k=1}^{d} \left\{ (\partial_k \log p) \cdot f_k + \partial_k f_k \right\},$$

*where $\boldsymbol{f} = (f_1, \ldots, f_d)$, p is a probability density function and $\mathcal{S}_p := \{\boldsymbol{f} \,|\, \mathbb{E}_p[T_p \boldsymbol{f}] = 0\}$ is called a **Stein class** of p. The Maximum Stein Discrepancy between two densities q and p (Chwialkowski et al., 2016; Liu et al., 2016) is defined as $\max_{\boldsymbol{f} \in \mathcal{S}_p} \mathbb{E}_q [T_p \boldsymbol{f}]$.*

By constructing different Stein classes $\mathcal{S}_p$, we obtain different Maximum Stein Discrepancies.

Since our task is estimating a parametric density model $p_{\boldsymbol{\theta}}$, we can consider a density estimator that minimizes this discrepancy, i.e., $\operatorname{argmin}_{\boldsymbol{\theta}} \max_{\boldsymbol{f} \in \mathcal{S}_{p_{\boldsymbol{\theta}}}} \mathbb{E}_q \left[T_{p_{\boldsymbol{\theta}}} \boldsymbol{f} \right]$. This estimator has been shown to be effective, robust, and closely related to SM. It is called Minimum Stein Discrepancy Estimator (Barp et al., 2019).

When using the above estimator, the key issue is constructing a Stein class S_{p_θ} which is expressive enough to capture subtle differences between q and $p_{\boldsymbol{\theta}}$. In the following theorem, we construct a Stein class with $\boldsymbol{f}$ that is the product of a smooth function and a Lipschitz function. We show the Maximum Stein Discrepancy using this specific Stein class becomes FH divergence weighted by the distance weight g_0.

First, we introduce a Lemma whose proof can be found in Appendix D.1.

Lemma 5 *Let* $\mathrm{Lip}_0^L(V)$ *be the set of all functions* $f : \overline{V} \to \mathbb{R}$

- *that are L-Lipschitz continuous with respect to* $\mathrm{dist}(\cdot, \cdot)$
- *satisfies the property that* $f(\boldsymbol{x}) = 0, \forall \boldsymbol{x} \in \partial V$

then $\max_{g_k \in \mathrm{Lip}_0^L,\ \forall k \in [d]} \mathrm{FH}_{\boldsymbol{g}}(q, p_{\boldsymbol{\theta}}) = L \cdot \mathrm{FH}_{g_0}(q, p_{\boldsymbol{\theta}})$.

Theorem 3 *Let* $\mathcal{F}$ *be a function class such that* $\forall \boldsymbol{f} \in \mathcal{F}$, $\boldsymbol{f} = \boldsymbol{h} \circ \boldsymbol{g}^{1/2}$ *and* $\mathbb{E}_q \|\boldsymbol{h}(\boldsymbol{x})\|^2 \leq 1$, *where* $\boldsymbol{h} : \overline{V} \to \mathbb{R}^d$ *is a smooth function and* $\forall k, g_k \in \mathrm{Lip}_0^L(V)$ *then*

- $\mathcal{F}$ *is a Stein class of a smooth density* q *defined on* V.
- *If* $\partial_k \ell_{\boldsymbol{\theta}}$ *is smooth with respect to* $\boldsymbol{\theta}$ *for all* k, *we have* $\max_{\boldsymbol{f} \in \mathcal{F}} \mathbb{E}_q \left[T_{p_{\boldsymbol{\theta}}} \boldsymbol{f} \right] = \sqrt{L \cdot \mathrm{FH}_{g_0}(q, p_{\boldsymbol{\theta}})}$.

The proof can be found in Appendix D.2 and it is partly based on the proof of Theorem 2 in Barp et al. (2019). Theorem 3 shows that the TruncSM objective is a Maximum Stein Discrepancy thus $\boldsymbol{\theta}_{\mathrm{TSM}}$ is a Minimum Stein Discrepancy Estimator.

Similarly, we can show how a "capped" distance weight arises by choosing a slightly different family for g_k in Theorem 3. We find this capped weight function can be also effective in many tasks (see Section 6.3 and 8.2).

Corollary 1 *Let us define* $\overline{\mathrm{Lip}}_0^L(V) := \{ f \in \mathrm{Lip}_0^L(V) | f(\boldsymbol{x}) \leq 1 \}$ *and a new function class* $\overline{\mathcal{F}}$ *such that for all* $\boldsymbol{f} \in \overline{\mathcal{F}}$, $\boldsymbol{f} = \boldsymbol{h} \circ \boldsymbol{g}^{1/2}$, $\mathbb{E}_q \|\boldsymbol{h}(\boldsymbol{x})\|^2 \leq 1$, *and* $g_k \in \overline{\mathrm{Lip}}_0^L(V)$, *then*

$$\max_{\boldsymbol{f} \in \overline{\mathcal{F}}} \mathbb{E}_q \left[T_{p_{\boldsymbol{\theta}}} \boldsymbol{f} \right] = \sqrt{\mathrm{FH}_{\bar{g}_L}(q, p_{\boldsymbol{\theta}})},$$

where $\bar{g}_L = \min(1, L \cdot g_0)$.

The proof can be found in Appendix D.3.

Yu et al. (2019) proposed to use a weight function $g_k = \min(1, x_k), k = 1, \ldots, d$ for estimating density functions defined on the non-negative orthant. This weight function is the special case of $\bar{g}_L$ when V is the entire non-negative orthant. We refer to $\bar{g}_L$ as a "capped weight function". The capped weight function is discussed in the recent paper by Yu et al. (2021) in the context of generalized SM for the unbounded V.

Although Theorem 3 and Corollary 1 show that g_0 and $\bar{g}_L$ are natural in the sense that they both give rise to Stein divergences, these choices may not be necessarily efficient in statistical inference tasks. In Section 6 and 8, we show that our choices of weight functions also lead to good statistical inference performance.

In the following section, we investigate the statistical guarantee of TruncSM through finite sample estimation error analysis.

6. Finite Sample Statistical Guarantee of Generalized Score Matching and Optimality of Truncated Score Matching

To discuss the efficiency of the TruncSM estimator, we first establish a finite sample estimation error bound for generalized SM. Using this result, we show that the TruncSM estimator is a good statistical estimator.

There have been studies (Yu et al. 2019; Barp et al. 2019) on the *asymptotic accuracy* of generalized SM. However, it is hard to study the impact of weight functions from the asymptotic variance due to its complicated expression.

We now establish an estimation error bound for generalized SM using a generic weight function $\boldsymbol{g}$. For the convenience of discussion in this section, let us rewrite the generalized SM objective function as

$$M(\boldsymbol{\theta}) = \mathbb{E}_q[m_{\boldsymbol{\theta}}(\boldsymbol{x})], \text{ where } m_{\boldsymbol{\theta}}(\boldsymbol{x}) := \sum_{k=1}^{d} \left\{ [\underbrace{(\partial_k \ell_{\boldsymbol{\theta}})^2 + 2\partial_k^2 \ell_{\boldsymbol{\theta}}}_{=:A_k(\boldsymbol{x};\boldsymbol{\theta})}] g_k + \underbrace{2\partial_k \ell_{\boldsymbol{\theta}}}_{=:B_k(\boldsymbol{x};\boldsymbol{\theta})} \partial_k g_k \right\}, \quad (8)$$

where we abbreviated $\log p_{\boldsymbol{\theta}}$ as $\ell_{\boldsymbol{\theta}}$. It holds that $\mathrm{FH}_{\boldsymbol{g}}(q, p_{\boldsymbol{\theta}}) = M(\boldsymbol{\theta}) + C$ in which C is independent of $\boldsymbol{\theta}$. Given finite samples $X_q \sim Q$, $M(\boldsymbol{\theta})$ is approximated by the empirical mean:

$$\hat{M}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} m_{\boldsymbol{\theta}}(\boldsymbol{x}_i).$$

Then $\hat{\boldsymbol{\theta}}$, the minimizer of $\hat{M}(\boldsymbol{\theta})$ s.t. $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^r$ is an M-estimator (van der Vaart, 2000) with the estimation function $m_{\boldsymbol{\theta}}(\boldsymbol{x})$. If $\ell_{\boldsymbol{\theta}}$ is the log-likelihood of an exponential family distribution, i.e., $\ell_{\boldsymbol{\theta}} := \boldsymbol{\theta}^\top \boldsymbol{t}(\boldsymbol{x}) - \log Z(\boldsymbol{\theta})$, where $Z(\boldsymbol{\theta})$ is the normalizing constant,

$$\hat{M}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{d} \left\{ \left[\boldsymbol{\theta}^\top \partial_k \boldsymbol{t}(\boldsymbol{x}_i) \partial_k \boldsymbol{t}(\boldsymbol{x}_i)^\top \boldsymbol{\theta} + 2\boldsymbol{\theta}^\top \partial_k^2 \boldsymbol{t}(\boldsymbol{x}_i) \right] g_k(\boldsymbol{x}_i) + 2\boldsymbol{\theta}^\top \partial_k \boldsymbol{t}(\boldsymbol{x}_i) \partial_k g_k \right\},$$

which is convex and quadratic with respect to $\boldsymbol{\theta}$. Thus, a unique $\hat{\boldsymbol{\theta}}$ can be easily obtained as long as $\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{d} g_k(\boldsymbol{x}_i) \partial_k \boldsymbol{t}(\boldsymbol{x}_i) \partial_k \boldsymbol{t}(\boldsymbol{x}_i)^\top \in \mathbb{R}^{r \times r}$ is invertible. However, our theorem below does not assume that $p_{\boldsymbol{\theta}}$ has a specific form.

6.1 Non-Asymptotic Error Bound

We first establish the non-asymptotic error bound of the generalized SM procedure (8). Let $\boldsymbol{\theta}^* \in \Theta$ be the minimizer of $M(\boldsymbol{\theta})$ over Θ. We assume that the optimal parameter $\boldsymbol{\theta}^*$ is well-separated from other neighbouring parameters in terms of the population objective values:

Assumption 1 *Assume that there exists $\alpha > 1$ such that*

$$\inf_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\delta} M(\boldsymbol{\theta}) - M(\boldsymbol{\theta}^*) \geq C_{\boldsymbol{g}}\delta^{\alpha} \tag{9}$$

holds for any small $\delta > 0$. Here, $C_{\boldsymbol{g}}$ is a positive constant that depends on the weight function $\boldsymbol{g}$ such that $C_{a\boldsymbol{g}} = aC_{\boldsymbol{g}}$ for any positive constant a.

Although we mainly focus on the dependency between the convergence rate and $\boldsymbol{g}$, the constant $C_{\boldsymbol{g}}$ can also depend on the dimensionality of the parameter space r, as we demonstrate in the following example.

Example 1 *Let us consider the exponential family*

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp\left\{\sum_{k=1}^{r} t_k(\boldsymbol{x})\theta_k - \phi(\boldsymbol{\theta})\right\}, \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^r \tag{10}$$

such that Θ is bounded, i.e., $\|\boldsymbol{\theta}\| < R$.

Suppose $q = p_{\boldsymbol{\theta}^*}$, guaranteed by Theorem 1 under mild conditions. Then, some calculation yields that $M(\boldsymbol{\theta})$ is a convex quadratic function,

$$M(\boldsymbol{\theta}) - M(\boldsymbol{\theta}^*) = (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\top}\left(\sum_{k=1}^{d} T_k\right)(\boldsymbol{\theta} - \boldsymbol{\theta}^*),$$

where $(T_k)_{ij} = (\mathbb{E}_{\boldsymbol{\theta}^*}[g_k(\boldsymbol{x})\partial_k t_i(\boldsymbol{x})\partial_k t_j(\boldsymbol{x})])_{ij} \in \mathbb{R}^{r\times r}$. Note that T_k is positive semidefinite. We assume that $\frac{1}{d}\sum_{k=1}^{d} T_k \succeq \bar{T} \succ O$ holds for a positive definite matrix $\bar{T} \in \mathbb{R}^{r\times r}$, where the inequalities are defined in the sense of positive definiteness. A mild assumption is that the eigenvalues of $\bar{T}$ does not depend on d. Let $\lambda_1 \geq \cdots \geq \lambda_r > 0$ be eigenvalues of $\bar{T}$, then, we have

$$M(\boldsymbol{\theta}) - M(\boldsymbol{\theta}^*) \geq d\lambda_r\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2. \tag{11}$$

Hence, $C_{\boldsymbol{g}} = d\lambda_r$ and $\alpha = 2$ meet Assumption 1. One can also confirm that $C_{a\boldsymbol{g}} = aC_{\boldsymbol{g}}$ for any $a > 0$.

Assumption 2 *The sequence $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}^*$ in probability.*

The consistency of M-estimator has been well-studied in the statistical literature, thus we do not discuss this in detail. Here we are only interested in the rate that governs the convergence of $\hat{\boldsymbol{\theta}}$ and its implication on choosing g_k. A set of sufficient conditions for proving the consistency of $\hat{\boldsymbol{\theta}}$ is discussed in Theorem 5.7 of van der Vaart (2000).

We also make continuity assumptions on $m_{\boldsymbol{\theta}}(\boldsymbol{x})$. This is needed to ensure the upper-boundedness of the covering number when proving the convergence rate.

Assumption 3 *For the function A_k and B_k in $m_{\boldsymbol{\theta}}(\boldsymbol{x})$, there exists $\dot{A}_k, \dot{B}_k, \forall k$ such that*

$$\begin{aligned} |A_k(\boldsymbol{x}, \boldsymbol{\theta}_1) - A_k(\boldsymbol{x}, \boldsymbol{\theta}_2)| &\leq \dot{A}_k(\boldsymbol{x})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|, \\ |B_k(\boldsymbol{x}, \boldsymbol{\theta}_1) - B_k(\boldsymbol{x}, \boldsymbol{\theta}_2)| &\leq \dot{B}_k(\boldsymbol{x})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|. \end{aligned} \tag{12}$$

If $A_k(\boldsymbol{x}, \boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$ for all k, due to Taylor expansion and Schwarz inequality, we see $|A_k(\boldsymbol{x}, \boldsymbol{\theta}_1) - A_k(\boldsymbol{x}, \boldsymbol{\theta}_2)| \leq \sup_{\boldsymbol{\theta} \in \Theta} \|\nabla_{\boldsymbol{\theta}} A_k(\boldsymbol{x}, \boldsymbol{\theta})\| \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$. Applying the same argument to $B_k(\boldsymbol{x}, \boldsymbol{\theta})$, we can see that both $\dot{A}_k := \sup_{\boldsymbol{\theta} \in \Theta} \|\nabla_{\boldsymbol{\theta}} A_k(\boldsymbol{x}, \boldsymbol{\theta})\|_2$ and $\dot{B}_k := \sup_{\boldsymbol{\theta} \in \Theta} \|\nabla_{\boldsymbol{\theta}} B_k(\boldsymbol{x}, \boldsymbol{\theta})\|_2$ would satisfy Assumption 3.

Let us define a function $\Gamma(\boldsymbol{g}; A, B)$ using $\dot{A}_k$ and $\dot{B}_k$:

$$\Gamma(\boldsymbol{g}; A, B) := \sum_{k=1}^{d} \left\{ (\mathbb{E}_q[\dot{A}_k^4] \mathbb{E}_q[g_k^4])^{1/4} + (\mathbb{E}_q[\dot{B}_k^4] \mathbb{E}_q[(\partial_k g_k)^4])^{1/4} \right\}, \tag{13}$$

then we have the following non-asymptotic estimation error bound of generalized SM:

Theorem 4 *Suppose that Assumptions 1, 2 and 3 hold and that $g_k, \partial_k g_k$, $\dot{A}$ and $\dot{B}$ have the fourth order moment under the population distribution q. Then for $\delta < C K_\alpha \cdot \frac{\sqrt{r}}{2^{\alpha-1}} \frac{\Gamma(\boldsymbol{g}; A, B)}{C_{\boldsymbol{g}}}$ we have*

$$\Pr\left[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \leq \left(C K_\alpha \cdot \frac{\Gamma(\boldsymbol{g}; A, B)}{\delta C_{\boldsymbol{g}}} \cdot \sqrt{\frac{r}{n}} \right)^{1/(\alpha-1)} \right] \geq 1 - \delta, \tag{14}$$

where C is a universal constant and $K_\alpha = \frac{2^{2\alpha}}{2^{\alpha-1}-1}$.

The proof can be found in Appendix E. In the proof, we use the convergence analysis of the M-estimator according to Section 5.8 in van der Vaart (2000).

Example 2 *Let us analyse the estimation error for the exponential family. Following the model definition* (10)*, we can see that $\alpha = 2$ and $\Gamma_{\boldsymbol{g}} = O(d)$[1]. The analysis in Example 1 leads to $C_{\boldsymbol{g}} = d\lambda_r$. Thus, we have $\frac{\Gamma_{\boldsymbol{g}}}{C_{\boldsymbol{g}}} \simeq \frac{1}{\lambda_r}$ and*

$$\Pr\left(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \frac{C'}{\delta \lambda_r} \sqrt{\frac{r}{n}} \right) \geq 1 - \delta,$$

where C' is a constant independent of d, r and δ. In the above bound, the probability δ and the sample size n are variables and other parameters such as r and λ_r are regarded as a fixed constant.

Theorem 4 shows how the choice of weight $\boldsymbol{g}$ affects the convergence rate of $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$. The only term involving $\boldsymbol{g}$ on the RHS of (14) is $\frac{\Gamma_{\boldsymbol{g}}}{C_{\boldsymbol{g}}}$. Naturally, we would like to choose a $\boldsymbol{g}$ such that $\frac{\Gamma_{\boldsymbol{g}}}{C_{\boldsymbol{g}}}$ is minimized. Some details will be discussed in Section 6.2 and Appendix F.

6.2 Choice of Weight Function

When our model $p_{\boldsymbol{\theta}}$ is not correctly-specified, $\boldsymbol{\theta}^*$ depends on the weight function $\boldsymbol{g}$. For simplicity, we assume the model is correctly specified and is identifiable, thus Theorem 1 ensures that $\boldsymbol{\theta}^*$ is unique and $p_{\boldsymbol{\theta}^*} = q$ under mild conditions.

We cannot minimize $\frac{\Gamma_{\boldsymbol{g}}}{C_{\boldsymbol{g}}}$ with respect to $\boldsymbol{g}$ analytically, as we do not know $q(\boldsymbol{x})$ and the closed form expressions of $\dot{A}_k$ or $\dot{B}_k$. Numerical minimization of $\frac{\Gamma_{\boldsymbol{g}}}{C_{\boldsymbol{g}}}$ can also be cumbersome.

1. For simplicity, let us shorten $\Gamma(\boldsymbol{g}; A, B)$ as $\Gamma_{\boldsymbol{g}}$ from now on.

Instead, we present a lightweight selection procedure, which eventually gives rise to the distance weight function g_0 by controlling an upperbound of $\frac{\Gamma_{\boldsymbol{g}}}{C_{\boldsymbol{g}}}$.

Suppose there exists a constant $\Gamma_{\mathcal{G}} \geq \sup_{\boldsymbol{g}\in\mathcal{G}} \Gamma_{\boldsymbol{g}}$, where $\mathcal{G}$ is a function family from which $\boldsymbol{g}$ is chosen. We can obtain an upperbound $\frac{\Gamma_{\mathcal{G}}}{C_{\boldsymbol{g}}} \geq \frac{\Gamma_{\boldsymbol{g}}}{C_{\boldsymbol{g}}}$. The choice of $\mathcal{G}$ is important as $\Gamma_{\mathcal{G}}$ may not exist or be very large for some $\mathcal{G}$. However, (13) suggests that as long as $|\partial_k g_k|$ and $|g_k|$ are upper bounded and $\dot{A}_k$ and $\dot{B}_k$ are well behaved, $\Gamma_{\mathcal{G}}$ should exist. Let us consider

$$\mathcal{G} := \{\boldsymbol{g} : \overline{V} \to \mathbb{R}^d | g_k \in \mathrm{Lip}_0^1(V), \forall k \in [d]\}.$$

We can see $\forall \boldsymbol{g} \in \mathcal{G}, |\partial_k g_k| \leq 1$ due to the property of Lipschitz function and $|g_k|$ is bounded as it is defined over a bounded domain.

Note that we have used Lipschitz functions to construct a Stein class in Section 5, but here, Lipschitz functions emerge from a different context: Suppressing the upperbound of $\Gamma_{\boldsymbol{g}}$.

After fixing $\mathcal{G}$, we seek to maximize the denominator $C_{\boldsymbol{g}}$ by choosing an appropriate $\boldsymbol{g} \in \mathcal{G}$. Let us introduce the following proposition:

Proposition 1 *Given the $\mathcal{G}$ defined above, suppose $q = p_{\boldsymbol{\theta}^*}$, then*

$$\inf_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\delta} M_{g_0}(\boldsymbol{\theta}) - M_{g_0}(\boldsymbol{\theta}^*) \geq \sup_{\boldsymbol{g}\in\mathcal{G}} \inf_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\delta} M_{\boldsymbol{g}}(\boldsymbol{\theta}) - M_{\boldsymbol{g}}(\boldsymbol{\theta}^*),$$

where $M_{\boldsymbol{g}}(\boldsymbol{\theta})$ is $M(\boldsymbol{\theta})$ using $\boldsymbol{g}$ as weight function.

Proof

$$\begin{aligned}
&\inf_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\delta} M_{g_0}(\boldsymbol{\theta}) - M_{g_0}(\boldsymbol{\theta}^*) \\
=&\inf_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\delta} \mathrm{FH}_{g_0}(\boldsymbol{\theta}) - \mathrm{FH}_{g_0}(\boldsymbol{\theta}^*) \\
=&\inf_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\delta} \sup_{\boldsymbol{g}\in\mathcal{G}} \mathrm{FH}_{\boldsymbol{g}}(\boldsymbol{\theta}) - \mathrm{FH}_{\boldsymbol{g}}(\boldsymbol{\theta}^*) \\
=&\inf_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\delta} \sup_{\boldsymbol{g}\in\mathcal{G}} M_{\boldsymbol{g}}(\boldsymbol{\theta}) - M_{\boldsymbol{g}}(\boldsymbol{\theta}^*) \geq \sup_{\boldsymbol{g}\in\mathcal{G}} \inf_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\delta} M_{\boldsymbol{g}}(\boldsymbol{\theta}) - M_{\boldsymbol{g}}(\boldsymbol{\theta}^*).
\end{aligned}$$

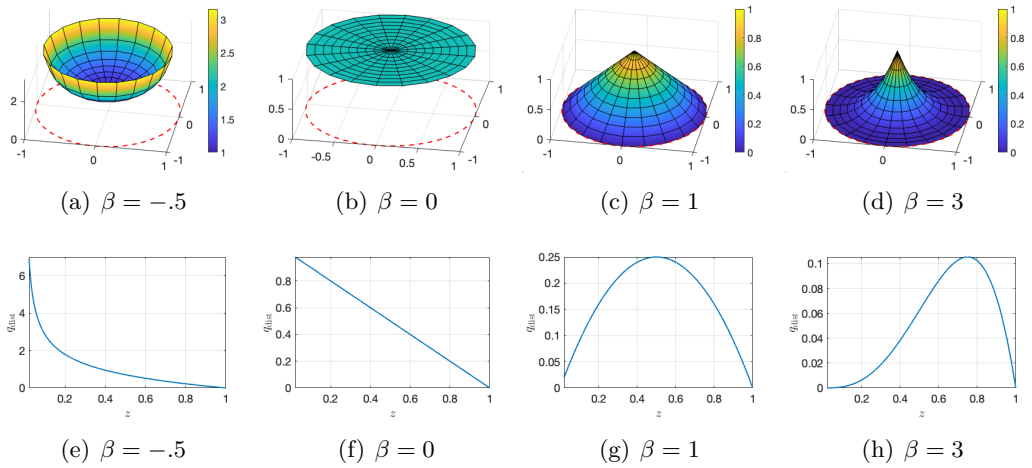
The second equality is due to Lemma 5 and $\mathrm{FH}_{\boldsymbol{g}}(\boldsymbol{\theta}^*) \equiv 0, \forall \boldsymbol{g}$. The inequality is due to the max-min inequality. ■

Proposition 1 shows, when using the distance weight g_0 as the weight function, Assumption 1 should hold for a constant $C_{g_0} = \sup_{\boldsymbol{g}\in\mathcal{G}} C_{\boldsymbol{g}}$. Since $\boldsymbol{g}_0 = (g_0, \ldots, g_0)$ is in $\mathcal{G}$ (see the proof of Lemma 3), g_0 maximizes $C_{\boldsymbol{g}}$ for all $\boldsymbol{g} \in \mathcal{G}$.

As $\Gamma_{\mathcal{G}}$ does not depend on any individual $\boldsymbol{g}$, g_0 minimizes the upper bound $\frac{\Gamma_{\mathcal{G}}}{C_{\boldsymbol{g}}}$.

Note that we can reach a similar conclusion by replacing $\mathcal{G}$ with $\overline{\mathcal{G}} := \{\boldsymbol{g} \in \mathbb{R}^d | g_k \in \overline{\mathrm{Lip}_0^1}(V), \forall k \in [d]\}$ and the minimizer of $\frac{\Gamma_{\mathcal{G}}}{C_{\boldsymbol{g}}}$ would change from distance g_0 to the capped distance $\bar{g}_L$.

Here we do *not* claim that g_0 or its capped counterpart $\bar{g}_L$ is *the best* weight function for estimating truncated densities using generalized SM. (13) and (14) suggest that there should be other data-driven choices of $\boldsymbol{g}$ that yield better error bounds. However, g_0 should be an adequate choice without using any information on $q, p_{\boldsymbol{\theta}}$ and V, judging from our analysis. Finding a tractable data-driven $\boldsymbol{g}$ that minimizes the estimation error bound is an interesting future work (See Section 9 for more information).



(a) $\beta = -.5$ (b) $\beta = 0$ (c) $\beta = 1$ (d) $\beta = 3$

(e) $\beta = -.5$ (f) $\beta = 0$ (g) $\beta = 1$ (h) $\beta = 3$

Figure 4: Top row: unnormalized $q_\beta(\boldsymbol{x})$, where ∂V in Example 3 is illustrated using a red dashed line. Bottom row: unnormalized $p_{\text{dist}}(z)$ in Example 3.

6.3 Case Study: Truncated Density in a Unit Ball

In this section, we consider a case where q has a parametric form, V is a unit ball. Although this is a particular setting, we can see how the relationship between q, V, and different choices of L in $\bar{g}_L$ would influence the error bound. In what follows, we consider the capped distance weight $\bar{g}_L(\boldsymbol{x}) = \min\{Lg_0(\boldsymbol{x}), 1\}$ which has been introduced in Corollary 1 where the distance weight g_0 is defined using the euclidean distance.

Let us define $p_{\text{dist}}(z)$ as the probability density of $Z = g_0(X)$ for $X \sim q$. For simplicity, we assume that there exist positive constants b and b' such that $bz^\beta \leq p_{\text{dist}}(z) \leq b'z^\beta$ holds for $0 < z \leq c$, where c and β are constants. Note that β should be greater than -1 since the integral of $p_{\text{dist}}(z)$ on the interval $(0, c)$ should be bounded.

Example 3 *Let V be the unit open ℓ^2 ball in $\mathbb{R}^d$, i.e., $V = \{\boldsymbol{x} \in \mathbb{R}^2 | \|\boldsymbol{x}\| < 1\}$. We can consider a distribution $q_\beta(\boldsymbol{x}) \propto (1 - \|\boldsymbol{x}\|)^\beta$ on V, where $\beta > -1$. Then, we have $p_{\text{dist}}(z) \propto (1-z)^{d-1}z^\beta$, $0 < z \leq 1$. For a small z, we have $bz^\beta \leq p_{\text{dist}}(z) \leq b'z^\beta$.*

See Figure 4 for illustrations of unnormalized q_β and $p_{\text{dist}}(z; \beta)$. It can be seen that:

- If $q(\boldsymbol{x})$ converges to a positive constant as $\boldsymbol{x} \to \partial V$, $\beta = 0$.
- If $q(\boldsymbol{x})$ tends to zero, as $\boldsymbol{x} \to \partial V$, $\beta > 0$.
- If $q(\boldsymbol{x})$ tends to infinity, as $\boldsymbol{x} \to \partial V$, $\beta < 0$.

For $\beta > -1$ and $L \geq \max\{1, 1/c\}$, where c is independent of L, we have the bounds for $\frac{\Gamma_{\bar{g}_L}}{C_{\bar{g}_L}}$:

$$C_A\left(1 - \frac{C_{b',\beta}}{L^{\beta+1}}\right)^{1/4} + C_B c_0 L^{(1-\beta)/2} \leq \frac{\Gamma_{\bar{g}_L}}{C_{\bar{g}_L}} \leq C_A\left(1 - \frac{c_{b,\beta}}{L^{\beta+1}}\right)^{1/4} + C_B c_1 L^{(3-\beta)/4}, \quad (15)$$

where $c_0, c_1, c_{b,\beta}, C_{b',\beta}, C_A, C_B$ are positive constants independent of L. The first term of the lower and upper bounds is positive for $L \geq 1$. The detailed derivation of this bound is found in Appendix G. Although we assume that p_{dist} takes a specific form in this example, the derivation mostly concerns the behavior of $p_{\mathrm{dist}}(z), z < c$. Thus a slight modification of (15) should cover a different pair of p_{dist} and q_β that has the same behavior near the boundary.

First, let us consider the condition $\beta > 3$ in which case $q(\boldsymbol{x})$ rapidly goes to zero as $\boldsymbol{x}$ converges to a point on the boundary of V. The upper and lower bounds in (15) converge to C_A as L tends to infinity, meaning that a large L does guarantee a reasonable accuracy. When $\beta > 3$, $q(\boldsymbol{x})$ is almost zero around the boundary (see Figure 4). By setting L to a large value, TruncSM with $\bar{g}_L$ is essentially the classic SM which is well-suited for non-truncated density estimation.

On the other hand, when q_β goes to zero slowly ($0 < \beta \leq 1$) or converges to a constant at the boundary ($\beta \leq 0$), a larger L leads to a larger lower bound, which is undesirable. In particular, when $\beta < 0$, increasing L will increase both upper and lowerbound rapidly. This trend implies a smaller L would yield better performance when β is small.

Our analysis can also be visually validated via Figure 4: When β is large, p_{dist} is low around the boundary, thus a steep $\bar{g}_L$ (i.e., large L) near the boundary would not blow up $\Gamma_{\boldsymbol{g}}$ (which depends on $\mathrm{E}_q[(\partial_k g_k)^4] \leq L^4$). However, as we reduce β, p_{dist} takes higher values near the boundary. A steep $\bar{g}_L$ near the boundary would lead to a large $\Gamma_{\boldsymbol{g}}$, hence a larger estimation error.

For the truncated distribution on the bounded domain, such as the truncated Gaussian (or Gaussian mixture) model, the probability density $p_{\mathrm{dist}}(z)$ is greater than a positive constant near the boundary. The above analysis shows that the capped distance function works efficiently for such truncated probability models. We will empirically validate this analysis in Section 8.2.

7. Computation of Distance Weight g_0

Compared to other weight function choices, distance weight g_0 (and capped distance $\bar{g}_L$) have a significant advantage: the computation of g_0 and its gradient can be efficiently carried out for a properly defined V.

For example, if V and ∂V are expressed by $V = \{\boldsymbol{x} \in \mathbb{R}^d | u(\boldsymbol{x}) < 0\}$ and $\partial V = \{\boldsymbol{x} \in \mathbb{R}^d | u(\boldsymbol{x}) = 0\}$ using a function $u : V \to \mathbb{R}$, evaluating $g_0(\boldsymbol{x})$ can be turned into an optimization problem: $g_0(\boldsymbol{x}) = \min_{\boldsymbol{z}}\{\mathrm{dist}(\boldsymbol{x}, \boldsymbol{z}) \,|\, u(\boldsymbol{z}) = 0\}$. In addition, if Euclidean distance is considered, the gradient of $g_0(\boldsymbol{x})$ is simply given by $\nabla_{\boldsymbol{x}} g_0(\boldsymbol{x}) = (\boldsymbol{x} - \tilde{\boldsymbol{x}})/\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|$, where $\tilde{\boldsymbol{x}}$ is the minimizer of $\min_{\boldsymbol{z} \in \partial V} \|\boldsymbol{x} - \boldsymbol{z}\|$.

We only need to evaluate g_0 and $\partial_k g_0$ *exactly once* for all $\boldsymbol{x} \in X_q$ *before* estimating $\hat{\boldsymbol{\theta}}$, since g_0 is model agnostic. This can be advantageous when $\bar{p}_{\boldsymbol{\theta}}$ is a sophisticated model and the optimization for $\boldsymbol{\theta}$ is time-consuming. In contrast, if one uses Monte Carlo methods (e.g., Kannan et al. 1997) to approximate the normalizing constant $Z_V(\boldsymbol{\theta})$ in the likelihood gradient, They must update the estimate of $Z_V(\boldsymbol{\theta})$ throughout the entire gradient descent procedure. Approximating $Z_V(\boldsymbol{\theta})$ using Markov Chain Monte Carlo (MCMC) (Robert and Casella, 2013) can also get less efficient both in terms of statistical accuracy and computation as dimensionality increases. Moreover, some algorithms designed for estimating the truncated multivariate Gaussian likelihood gradient, such as the ones proposed by Daskalakis et al.

(2018, 2019), do not require exhaustive sampling. They need to evaluate a membership oracle (if $\boldsymbol{x} \in V$ or not) for auxiliary samples freshly drawn at each iteration. Membership evaluation can be more efficient than computing $g_0(\boldsymbol{x})$ for each $\boldsymbol{x}$. However, TruncSM only once evaluates g_0 for samples in the data set X_q. Thus it is a fixed computation cost that does not grow with the number of gradient descent iterations.

There may be a $U \subset V$, for all $\boldsymbol{x} \in U$, the corresponding $\tilde{\boldsymbol{x}}$ is not unique. If so, g_0 will be non-differentiable on U. However, Lemma 3 implies that the area in V, where g_0 is non-differentiable, has a measure zero. Thus U also has a measure of zero. Therefore, we do not need to worry about such a case in practice. In what follows, we show efficient analytical methods for computing the distance function defined over *unit ball*, *unit cube*, *convex polytopes* and *polygons*.

- For the Unit Ball, $V = \{\boldsymbol{x} \in \mathbb{R}^d | \|\boldsymbol{x}\| < 1\}$, the distance function and its gradient:

$$g_0(\boldsymbol{x}) = 1 - \|\boldsymbol{x}\|, \nabla_{\boldsymbol{x}} g_0(\boldsymbol{x}) = \frac{-\boldsymbol{x}}{\|\boldsymbol{x}\|}.$$

- For the Unit Cube, $V = \{\boldsymbol{x} \in \mathbb{R}^d | \|\boldsymbol{x}\|_\infty < 1\}$, the distance function and its gradient:

$$g_0(\boldsymbol{x}) = 1 - \|\boldsymbol{x}\|_\infty, \ \nabla_{\boldsymbol{x}} g_0(\boldsymbol{x}) = -\boldsymbol{e}_j, \ j = \operatorname*{argmax}_k |x_k|.$$

- For the Convex Polytope $V = \{\boldsymbol{x} \in \mathbb{R}^d | \langle \boldsymbol{a}_t, \boldsymbol{x} \rangle + b_t < 0, t = 1, \ldots, T\}$, the distance function is given by

$$g_0(\boldsymbol{x}) = \min_{\boldsymbol{z}} \{\|\boldsymbol{x} - \boldsymbol{z}\| \mid \max_{t \in [T]} \{\langle \boldsymbol{a}_t, \boldsymbol{z} \rangle + b_t\} = 0\} = \min_{t \in [T]} \frac{|\langle \boldsymbol{a}_t, \boldsymbol{x} \rangle + b_t|}{\|\boldsymbol{a}_t\|} \tag{16}$$

  for $\boldsymbol{x} \in V$. The envelope theorem (Milgrom and Segal, 2002) yields $\nabla_{\boldsymbol{x}} g_0(\boldsymbol{x}) = -\boldsymbol{a}_{t^*}/\|\boldsymbol{a}_{t^*}\|$, where $t^* \in [T]$ is the minimizer of (16). Briec (1997) studied another representation of the minimum distance problem for the convex polyhedral using the extreme points. Note that if V is not a convex set, (16) cannot be used.

- For the Convex Polytope $V = \{\boldsymbol{x} \in \mathbb{R}^d | \langle \boldsymbol{a}_t, \boldsymbol{x} \rangle + b_t < 0, t = 1, \ldots, T\}$, let us compute the ℓ^1-based distance function $g_0(\boldsymbol{x}) = \min_{\boldsymbol{z} \in \partial V} \|\boldsymbol{x} - \boldsymbol{z}\|_1$ for $\boldsymbol{x} \in V$. It holds that $g_0(\boldsymbol{x}) = \max\{|\alpha| \mid \boldsymbol{x} + \alpha \boldsymbol{e}_i \in V, \ \forall i\}$ since V is convex. The condition $\boldsymbol{x} + \alpha \boldsymbol{e}_i \in V, \ \forall i$ is expressed by $\alpha \langle \boldsymbol{a}_t, \boldsymbol{e}_i \rangle \leq -\langle \boldsymbol{a}_t, \boldsymbol{x} \rangle - b_t$ for all t and i. Hence, we have

$$g_0(\boldsymbol{x}) = \min_{i, t \text{ s.t. } \langle \boldsymbol{a}_t, \boldsymbol{e}_i \rangle \neq 0} \left| \frac{\langle \boldsymbol{a}_t, \boldsymbol{x} \rangle + b_t}{\langle \boldsymbol{a}_t, \boldsymbol{e}_i \rangle} \right| = \min_{t \in [T]} \frac{|\langle \boldsymbol{a}_t, \boldsymbol{x} \rangle + b_t|}{\|\boldsymbol{a}_t\|_\infty}. \tag{17}$$

  The envelope theorem (Milgrom and Segal, 2002) yields $\nabla_{\boldsymbol{x}} g_0(\boldsymbol{x}) = -\boldsymbol{a}_{t^*}/\|\boldsymbol{a}_{t^*}\|_\infty$ for $\boldsymbol{x} \in V$, where $t^* \in [T]$ is the minimizer of the last minimization in (17).

- For the Polygon V in $\mathbb{R}^2$ surrounded by the points $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_T \in \mathbb{R}^2$, the boundary is given by $\partial V = \cup_{t=1}^T \{\alpha \boldsymbol{p}_t + (1-\alpha)\boldsymbol{p}_{t+1} | \alpha \in [0,1]\}$, where $\boldsymbol{p}_{T+1} = \boldsymbol{p}_1$. Hence we have

$$g_0(\boldsymbol{x}) = \min_{t \in [T]} \min_{\alpha : 0 \leq \alpha \leq 1} \|\boldsymbol{x} - \alpha \boldsymbol{p}_t - (1-\alpha)\boldsymbol{p}_{t+1}\|.$$

The minimizer of the inner optimization is $\alpha_t = \min\{1, \max\{0, \frac{\langle \boldsymbol{p}_t - \boldsymbol{p}_{t+1}, \boldsymbol{x} - \boldsymbol{p}_{t+1}\rangle}{\|\boldsymbol{p}_t - \boldsymbol{p}_{t+1}\|^2}\}\}$. Hence, we obtain $g_0(\boldsymbol{x}) = \min_{t\in[T]} \|\boldsymbol{x} - \alpha_t \boldsymbol{p}_t - (1-\alpha_t)\boldsymbol{p}_{t+1}\|$, and $\nabla_{\boldsymbol{x}} g_0(\boldsymbol{x}) = \mathbf{n}/\|\mathbf{n}\|$, where $\mathbf{n} = \boldsymbol{x} - \alpha_{t^*}\boldsymbol{p}_{t^*} - (1-\alpha_{t^*})\boldsymbol{p}_{t^*+1}$ and where $t^* \in [T]$ is the minimizer of $g_0(\boldsymbol{x})$.

In these cases, the computation of g_0 can be done in polynomial time with respect to d (if T in convex polytope is a polynomial function of d).

8. Numerical and Real-world Data Analysis

To demonstrate the efficacy of TruncSM empirically, we conduct a wide range of experiments. This section uses the Euclidean metric for g_0 and $\bar{g}_L$. The code and data sets to reproduce our experiments are available at `https://github.com/anewgithubname/Truncated-Score-Matching`.

8.1 Illustrative Example and Computation Time

In the first experiment, we show an illustrative example comparing TruncSM and Rejection Sampling MLE (RJ-MLE) and their computational times. RJ-MLE uses rejection sampling to approximate the intractable normalizing term $Z_V(\boldsymbol{\theta})$ and perform maximum likelihood estimation. It can be seen as an example of Monte Carlo MLE (Geyer, 1994). Samples are generated from a Gaussian mixture on $\mathbb{R}^2$, with centers at

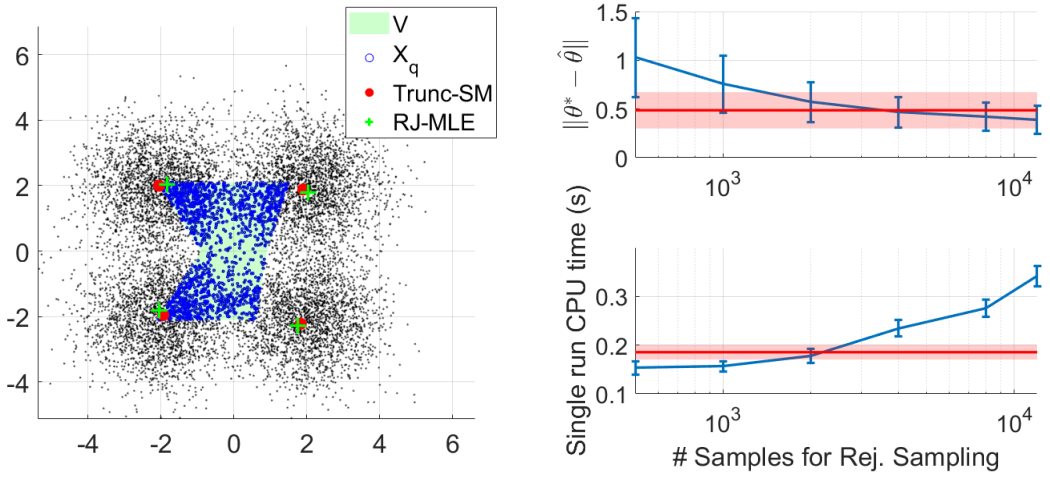
$$\boldsymbol{\mu}_1 = [2,2], \quad \boldsymbol{\mu}_2 = [-2,2], \quad \boldsymbol{\mu}_3 = [-2,-2], \quad \boldsymbol{\mu}_4 = [2,-2],$$

and standard deviations all set to 1. The pre-truncated data set can be seen in Figure 5(a) as black dots. To create a truncated data set, we limit our observation window to be a green polygon region in the middle, thus only samples inside the green polygon (blue points) can be observed. The task is to find all four centers of the mixture model using *only blue points*. We generate 10,000 samples, only 1417 of which can be used for parameter estimation within the truncation boundary. Our unnormalized density model is a Gaussian mixture model with four components (parametrized by $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_4$) and the unit variance-covariance matrix: $\bar{p}_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_4}(\boldsymbol{x}) = \sum_{i=1}^4 \mathcal{N}_{\boldsymbol{x}}(\boldsymbol{\theta}_i, \boldsymbol{I})$.

As the polygon boundary ∂V of the non-convex domain V in $\mathbb{R}^2$ consists of line segments, the analytical algorithm in Section 7 is available to compute g_0 and $\nabla_{\boldsymbol{x}} g_0$ efficiently. We compare with RJ-MLE which uses rejection sampling to approximate the normalizing constant $Z_V(\boldsymbol{\theta}) := \int_V \bar{p}_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_4}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$. In this experiment, 500,000 particles are used to approximate $Z_V(\boldsymbol{\theta})$ and they are drawn from a bivariate uniform distribution with density $\mathcal{U}_{x_1}(-2,2) \cdot \mathcal{U}_{x_2}(-2,2)$. The estimated mixture component centers are plotted as green crosses (RJ-MLE) and red dots (TruncSM) in Figure 5(a). It can be seen that both methods give estimates close to the true mixture centers. However, the computation time for TruncSM and RJ-MLE are 0.35 seconds and 3.35 seconds respectively[2] [3].

It is worth noting that our particle distribution is carefully chosen so that it tightly covers the truncation domain. It ensures the best rejection sampling performance: Over-coverage would increase the computational cost, and insufficient coverage would lower the

2. For TruncSM, we include the calculation time for both g_0 and $\partial_k g_k$. Same below.
3. Our experiments are run on a workstation with an AMD Ryzen 1700 CPU with 32GB memory.



(a) Estimation of truncated Gaussian mixture centers (b) Computational cost and estimation accuracy comparison for TruncSM (red) and RJ-MLE (blue)

Figure 5: Gaussian mixture centers truncated by a polygon

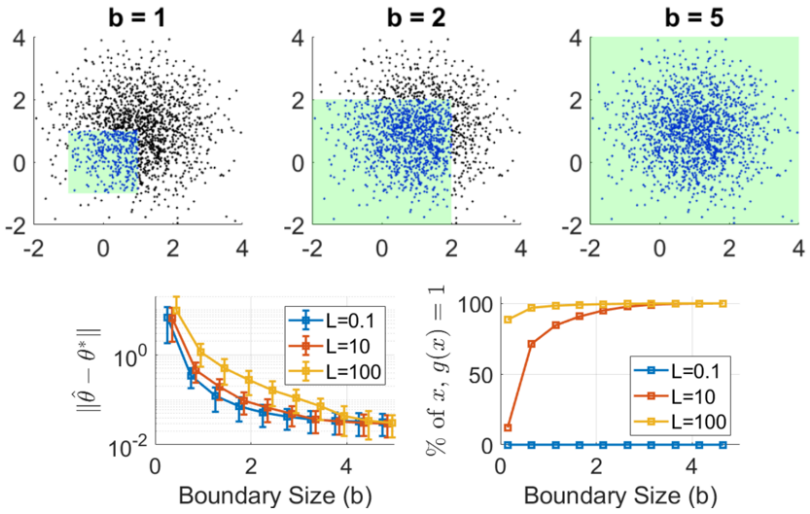
approximation accuracy of $Z_V(\boldsymbol{\theta})$. While in a lower-dimensional space, the visualization helps, it is unclear how to come up with a good rejection sampling distribution in a higher-dimensional space.

To further investigate the computation time between TruncSM and RJ-MLE, we study the estimation accuracy and computation time (with standard deviation) against the number of particles used for rejection sampling by RJ-MLE. We optimize both objective functions using MATLAB's `fminunc` function with default settings. From Figure 5(b), we can see that when using a large number of particles to approximate $Z_V(\boldsymbol{\theta})$, RJ-MLE can achieve a slightly better performance than TruncSM. However, the slight improvement of estimation accuracy comes with a significant penalty in computation cost even with an optimal choice of the rejection sampling distribution.

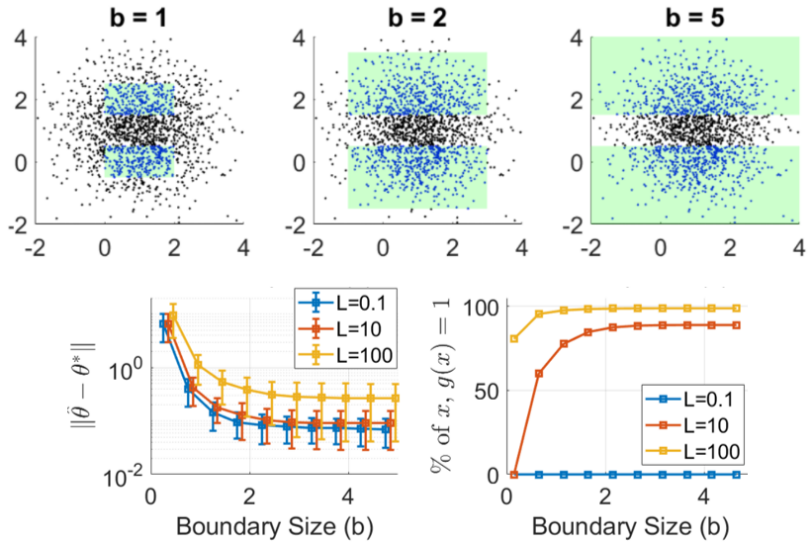
8.2 Capped Weight Function

Now we investigate the performance of a capped weight function: $\bar{g}_L := \min(1, L \cdot g_0(\boldsymbol{x}))$. It can be seen that when L is small, $\bar{g}_L$ will never be capped. Thus it is equivalent to g_0 after multiplying a constant, which does not affect the estimator. In this experiment, two different truncation domains are used: a single rectangle and two disjoint rectangles. 1600 samples are drawn from a normal distribution $\mathcal{N}([1,1], \boldsymbol{I})$. We monitor the performance of TruncSM using $\bar{g}_L$ as the weight function as V grows in size. We measure the growth of V using a scaling factor b (see Appendix I on how polygons are re-scaled). As shown in Remark 1, the true parameter is identifiable by TruncSM even when the domain consists of two disjoint rectangles.

Figure 6 illustrates the truncated data sets as V grows, the estimation performance and the percentage of data points whose $\bar{g}_L$ are capped.



(a)



(b)

Figure 6: The truncated data sets, the estimation performance, and the percentage of data points whose $\bar{g}_L$ are capped as V enlarges. b is the scaling multiplier of polygon vertices. It can be seen that as the area of V grows, $\bar{g}_L(\boldsymbol{x})$ becomes capped for more and more samples in our data set (the bottom right plots in both (a) and (b)). This phenomenon is expected as the boundary stretches, fewer and fewer points are adjacent to the boundary.

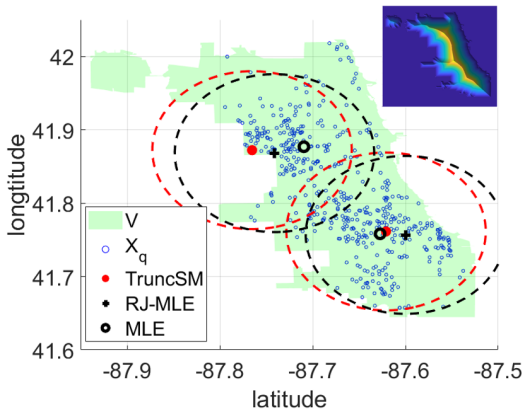


Figure 7: Chicago Crime data set, whose truncation boundary is a polygon. Blue circles are homicide locations. g_0 is visualized at the upper right corner. The estimated component centers of TruncSM, RJ-MLE and MLE are plotted.

In Figure 6(a), we can observe that the performance gaps between $L = 0.1$, $L = 10$, and $L = 100$ widen and shrink: If the truncation domain is small, not many samples are included in the truncation domain. Thus algorithms with all choices of L would suffer. However, as the truncation boundary grows, the difference starts to show: TruncSM with a smaller L performs better, as we analyzed in Section 6.3. As V grows beyond a certain point, the data set essentially becomes non-truncated, and TruncSM using large L reduces to a classic SM since $\bar{g}_L$ are always capped at 1. At this time, TruncSM with all choices of L converges to the same level of performance. Figure 6(b) shows a similar story, but estimation errors with different L converge to different levels as V enlarges: V never covers the center of our data set; thus, our data sets are always truncated. Again TruncSM with a smaller L gives a better performance, as we analyzed in Section 6.3.

8.3 2008 Chicago Crime Data Set

We also test the performance of TruncSM on a real-world truncated density estimation problem: Analyzing the crime occurrences in Chicago. The data set contains locations of homicides that happened in Chicago during 2008. We fit a Gaussian mixture model with two components on this data set. The standard deviations of the two components are fixed to the same value, roughly half of the "width" of the city.

In this experiment, we compare TruncSM with vanilla MLE using the non-truncated density model (MLE for short) and RJ-MLE using the truncated density model. In this data set, the Chicago city boundary is expressed via a polygon in $\mathbb{R}^2$, so the distance weight g_0 is calculated using the analytical solution given in Section 7.

The estimated means of two components are plotted in Figure 7. The estimated 95% confidence region is plotted for TruncSM and RJ-MLE as red and black dotted circles, respectively. We can see that TruncSM, MLE, and RJ-MLE all picked centers on the north

and south side of the city. However, MLE picked a northern location inside the city, while TruncSM and RJ-MLE picked a spot right next to the western border of Chicago.

In this case, TruncSM and RJ-MLE tend to put observed crimes on the decaying slope of a Gaussian density which would better explain the declining crime rate from the west to the east. Unaware of the truncation, MLE puts the Gaussian center in the middle of the city, while the crimes rarely happen in the east.

Although all estimators tested in this section solve non-convex optimization problems, we observe only minor changes in estimated Gaussian centers between different runs in the vast majority of runs with different initializations. See Appendix J for more discussion on this.

8.4 Outlier Over-trimming Compensation

Removing outliers is a vital data preprocessing step. If we know the percentage of outliers in our data set, we can adopt methods such as One-class Support Vector Machines (OSVM) (Schölkopf et al., 1999) to remove them. However, it is often impossible to determine the percentage of outliers a priori and aggressively setting the outlier percentage may result in inliers being removed from the data set.

Consider outlier trimming using OSVM. The method outputs a "decision function" $u(\boldsymbol{x}) := \sum_j \hat{\alpha}_j \phi_j(\boldsymbol{x})$ which defines a domain $V := \{\boldsymbol{x} \in \mathbb{R}^d | u(\boldsymbol{x}) > 0\}$, where $\hat{\alpha}_j$ is the parameter obtained from OSVM procedure and ϕ_j is a kernel function. In this experiment, we use Gaussian kernel, defined as $\phi_j(\boldsymbol{x}) := \exp(-\|\boldsymbol{x}_j - \boldsymbol{x}\|^2/2\sigma^2)$, where j is the j-th datapoint in the contaminated data set. OSVM chooses a V such that a certain proportion (e.g., 80%) of our data set is included in V. We discard any data point not in V as "outliers". However, if the specified outlier percentage is larger, we will trim inliers from our data set too, and our data set becomes a truncated data set. Estimation without considering the truncation boundary would lead to biased estimates.

This phenomenon is demonstrated using the following simulated experiment. We sample

$$\boldsymbol{x}_{\text{inlier}} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1.5^2 & 0 \\ 0 & 1 \end{bmatrix}\right), \qquad \boldsymbol{x}_{\text{outlier}} \sim \mathcal{N}\left(\begin{bmatrix} 3.5 \\ 3 \end{bmatrix}, \begin{bmatrix} .8^2 & 0 \\ 0 & .8^2 \end{bmatrix}\right).$$

In total, 500 inlier samples and 50 outlier samples are drawn. The outlier percentage (ν) in OSVM is set to 20%. In this experiment, we allow MATLAB to automatically determine the kernel bandwidth σ according to its predefined protocol. The data set (black dots), the selected inliers (blue dots), and the truncation domain V given by OSVM are visualized in the top left plot in Figure 8. On one hand, we can see that OSVM does separate the inliers from outliers using a boundary function $u(\boldsymbol{x})$. On the other hand, some inlier samples are also removed due to the aggressive setting of the outlier proportion.

We use the trimmed data set to estimate a truncated multivariate normal distribution $p_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) \propto \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is restricted to be a diagonal matrix. In this experiment, TruncSM uses the distance weight g_0. g_0 and $\nabla_{\boldsymbol{x}} g_0$ are computed numerically using the constrained optimization described in Section 7. We compare TruncSM with vanilla MLE, which does not use truncation information. Ellipses visualize the true and estimated 95% confidence regions in Figure 8. The vanilla MLE underestimates the variance of the inlier distribution, while TruncSM accurately recovers the 95% confidence region using the

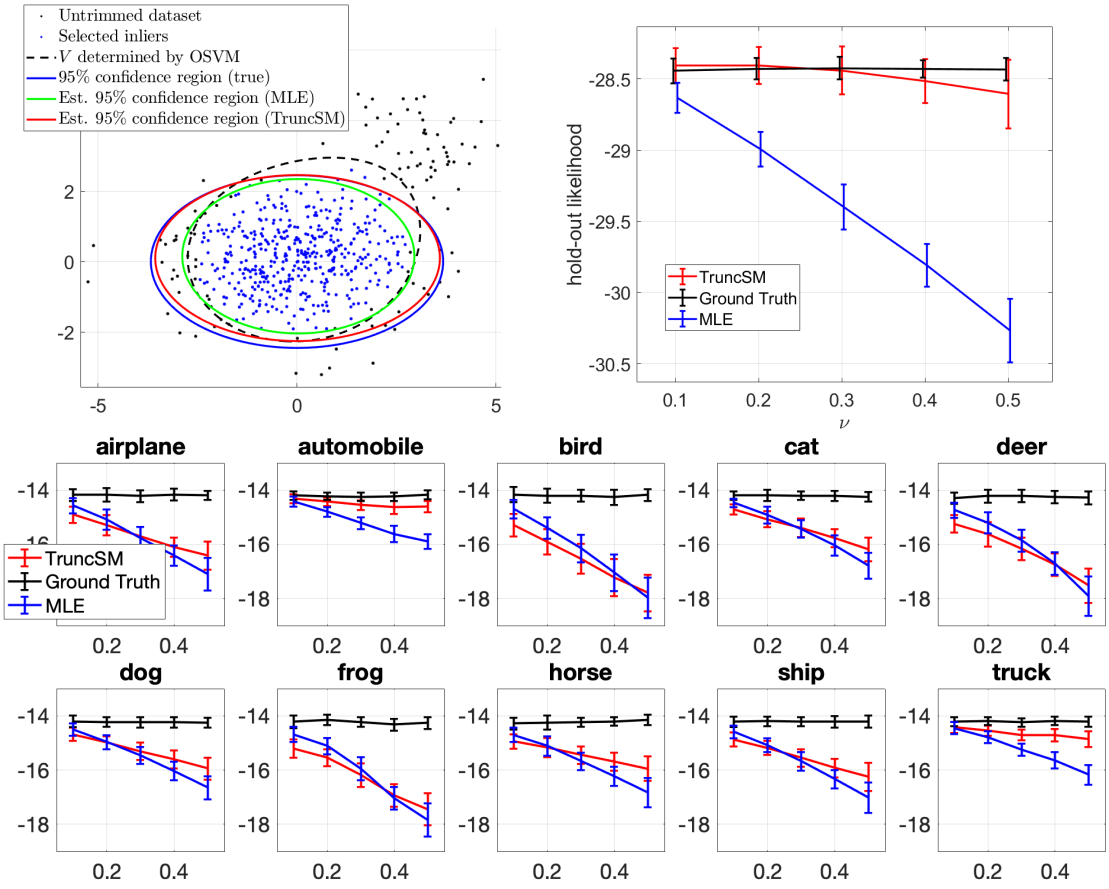


Figure 8: Top-left: MLE underestimates the 95% confidence region due to the truncation. The confidence region estimated by TruncSM is much closer to the true confidence region. Top-right, MLE is getting less and less accurate as the percentage of truncated samples (controlled by ν) increases while TruncSM maintains good accuracy. Bottom two rows: TruncSM achieves a comparable or better performance than vanilla MLE as the percentage of truncation increases on the CIFAR-10 data set.

estimated inlier density function. We then perform the same experiment in a much higher dimensional space: The inlier distribution is a 20-dimensional standard normal distribution, while the outlier distribution is a 20-dimensional normal distribution with mean $\mathbf{1}$ and unit variance. The truncation boundary (induced by a kernel function) is highly irregular in this experiment.

In the top-right plot of Figure 8, we plot the hold-out likelihood versus different settings of outlier proportion (ν) in OSVM. The likelihood is computed using both vanilla MLE and TruncSM solutions. As we can see, the more aggressive our outlier trimming is, the worse the vanilla MLE performs. On the other hand, TruncSM only drops very slightly in performance as more and more inlier data points are truncated.

We now experiment on a real-world data set, CIFAR-10, which contains ten different classes of 32 by 32 images. To speed up the computation, we reduced the dimension of the data set to 10 using PCA. We artificially add 10% outliers drawn from a normal distribution for each class with mean $\mathbf{1}$ and unit variance. We use TruncSM and vanilla MLE to estimate a multivariate Normal distribution for each class in the reduced 10-dimensional space and plot the hold-out likelihood of each method. The hold-out likelihood is plotted at the bottom two rows of Figure 8. For a large ν, we find that TruncSM performs better than MLE. The hold-out likelihood shows that TruncSM can correct the estimation bias induced by the trimmed data. Although the advantage of TruncSM is more minor on this data set, we argue that this is a very challenging problem: The data set is hardly Gaussian, so the model assumption used by TruncSM is wrong. In some cases (automobile and truck), TruncSM significantly outperforms MLE.

9. Conclusions and Future Works

We propose an estimator for truncated statistical models with complex truncation boundaries based on generalized SM. The proposed method uses the shortest distance (or capped distance) from a data point to the truncation boundary as the weight function. Such a choice of weight function naturally arises from minimizing Stein Discrepancy and lowering the estimation error upper bound. The proposed weight function is also computationally favorable for high dimensional truncation domains. Experiments on synthetic data and the Chicago crime data set show promising results. The proposed estimator was later applied to outlier trimming bias correction.

Although the proposed method achieves promising results in truncated density estimation, there are interesting open questions:

- How to choose an optimal weight function $\boldsymbol{g}$ when $p_{\boldsymbol{\theta}}, q$ and V are given? As we have seen from (14) in Theorem 4, the statistical estimation error depends on the ratio $\frac{\Gamma_{\boldsymbol{g}}}{C_{\boldsymbol{g}}}$ and both $\Gamma_{\boldsymbol{g}}$ and $C_{\boldsymbol{g}}$ depends on the weight function $\boldsymbol{g}$, the density model $p_{\boldsymbol{\theta}}$, the data density q and the truncation boundary V. A natural idea is to choose a $\boldsymbol{g}$ that minimizes $\frac{\Gamma_{\boldsymbol{g}}}{C_{\boldsymbol{g}}}$. Can we find an efficient numerical procedure for such a minimization?

- The related question is, how to choose an appropriate distance metric in g_0? For example, ℓ^1 and ℓ^2 distances are both good choices for g_0. Again, Theorem 4 suggests that, in terms of upper bounding the statistical estimation error, the answer depends on q, $p_{\boldsymbol{\theta}}$ as well as V. However, how would the geometry of V, q, and $p_{\boldsymbol{\theta}}$ affect the choice of the distance metric? See Appendix H for an empirical comparison between ℓ^1 and ℓ^2 distances.

- The computation of g_0 for a generic V is not trivial. In particular, when the dimensionality of data is large, computing g_0 can become computationally infeasible. Finding efficient ways of evaluating or approximately evaluating g_0 may help us extend the usage of the TruncSM estimator to higher-dimensional data sets.

- In many applications, our data set is automatically filtered by some algorithm. Outlier trimming by OSVM is just one example. Suppose we have mixed images of cats and dogs and a binary classifier. We can quickly identify the high confidence region of each

class. However, the classification step would also create artificial truncation boundaries around the distributions of individual classes. Can we use TruncSM to reconstruct the true underlying distribution using these pre-classified images?

Acknowledgments

We thank two anonymous reviewers for their insightful feedbacks. This work was supported by Japan Society for the Promotion of Science under KAKENHI Grant Number 17H00764, 19H04071, and 20H00576. Daniel J. Williams was supported by a PhD studentship from the EPSRC Centre for Doctoral Training in Computational Statistics and Data Science (COMPASS).

A. Generalized Score Matching

Assume that $p(\boldsymbol{x}), q(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x})$ take strictly positive values on the domain $V \subset \mathbb{R}^d$. Suppose that generalized SM objective function with weight $\boldsymbol{g}$ is equal to zero, i.e.,

$$\mathbb{E}_q[\|\boldsymbol{g}^{1/2}(\boldsymbol{x}) \circ \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) - \boldsymbol{g}^{1/2}(\boldsymbol{x}) \circ \nabla_{\boldsymbol{x}} \log q(\boldsymbol{x})\|^2] = 0. \tag{A.18}$$

Then, we have $\nabla_{\boldsymbol{x}}(\log p(\boldsymbol{x}) - \log q(\boldsymbol{x})) = \boldsymbol{0}$, meaning that $p(\boldsymbol{x}) = Cq(\boldsymbol{x})$ on (the connected domain) V with a positive constant C. Since both p and q are the probability densities on V, we have $C = 1$.

B. Proof of Lemma 3

Proof First, we show that g_0 is Lipschitz continuous with respect to the metric $\mathrm{dist}(\cdot, \cdot)$.

$$\begin{aligned}\forall \boldsymbol{x}_a, \boldsymbol{x}_b \in V, g_0(\boldsymbol{x}_a) - g_0(\boldsymbol{x}_b) &= \left(\min_{\boldsymbol{x}' \in \partial V} \max_{\boldsymbol{x}'' \in \partial V} \mathrm{dist}(\boldsymbol{x}_a, \boldsymbol{x}') - \mathrm{dist}(\boldsymbol{x}_b, \boldsymbol{x}'')\right) \\ &\leq \left(\max_{\boldsymbol{x}'' \in \partial V} \mathrm{dist}(\boldsymbol{x}_a, \boldsymbol{x}'') - \mathrm{dist}(\boldsymbol{x}_b, \boldsymbol{x}'')\right) \leq \mathrm{dist}(\boldsymbol{x}_a, \boldsymbol{x}_b).\end{aligned}$$

The last inequality is due to the triangle inequality. Likewise, $g_0(\boldsymbol{x}_b) - g_0(\boldsymbol{x}_a)$ is also bounded above by $\mathrm{dist}(\boldsymbol{x}_a, \boldsymbol{x}_b)$. Therefore g_0 is a Lipschitz function with Lipschitz constant 1. Rademacher's theorem (Evans and Gariepy, 1992) asserts that a Lipschitz continuous function is differentiable at every point in a Lipschitz domain outside a set of *measure zero*. Therefore, we can construct

$$D_k g_0(\boldsymbol{x}) := \begin{cases} \partial_k g_0(\boldsymbol{x}), & \text{if } g_0 \text{ is differentiable,} \\ \text{arbitary constant}, & \text{otherwise} \end{cases} \tag{B.19}$$

We can check that $D_k g_0(\boldsymbol{x})$ is a valid weak derivative (Definition 7.1.3. in Atkinson and Han 2005):

$$\int_V g_0(\boldsymbol{x}) \partial_k \phi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = -\int_V D_k g_0(\boldsymbol{x}) \phi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

where ϕ is any m-times differentiable function on a compact support on V. The equality holds due to the classic integration by parts formula for continuous differentiable functions. g_0 is only non-differentiable over a zero set, so the arbitary constant we set in (B.19) does not affect the outcome of the integration. Since V is a bounded domain and g_0 is 1-Lipschitz, g_0 and $D_k g_0$ are both bounded in terms of the $\|\cdot\|_{L^2(V)}$ norm. Therefore, g_0 is weakly differentiable and in a Sobolev-Hilbert space. ■

C. Proof of Theorem 2

Proof In this proof, we apply Theorem 4 to derive a tractable expression for $M(\boldsymbol{\theta})$. Let us consider the second term of $M(\boldsymbol{\theta})$. As $q(\boldsymbol{x})$ and $g_0(\boldsymbol{x})\partial_k \log p_\theta(\boldsymbol{x})$ are functions in the

Sobolev-Hilbert space of the first order, the direct application of Theorem 4 leads to

$$\begin{aligned}
&\sum_{k=1}^{d} \int_V g_k(\boldsymbol{x})[\partial_k \log p_\theta(\boldsymbol{x})][\partial_k \log q(\boldsymbol{x})] q(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= \sum_{k=1}^{d} \int_V g_k(\boldsymbol{x})[\partial_k \log p_\theta(\boldsymbol{x})][\partial_k q(\boldsymbol{x})] \mathrm{d}\boldsymbol{x} \\
&= \sum_{k=1}^{d} \left\{ \int_{\partial V} g_k(\boldsymbol{x})[\partial_k \log p_\theta(\boldsymbol{x})] q(\boldsymbol{x}) \nu_k(\boldsymbol{x}) \mathrm{d}s - \int_V \partial_k [g_k(\boldsymbol{x}) \partial_k \log p_\theta(\boldsymbol{x})] q(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right\} \\
&= -\sum_{k=1}^{d} \int_V \partial_k [g_k(\boldsymbol{x}) \partial_k \log p_\theta(\boldsymbol{x})] q(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.
\end{aligned}$$

The second equality is ensured by Theorem 4 and the third equality holds from the boundary condition imposed in the theorem. ■

D. Proofs of Theorem 3 and Corollary 1 in Section 5

D.1 Proof of Lemma 5

$$\max_{g_k \in \mathrm{Lip}_0^L(V)} \mathbb{E}_q \left[\sum_k (\partial_k \ell_{\boldsymbol{\theta}} - \partial_k \log q)^2 g_k \right] = L \cdot \mathbb{E}_q \left[\sum_k (\partial_k \ell_{\boldsymbol{\theta}} - \partial_k \log q)^2 g_0 \right] = L \cdot \mathrm{FH}_{g_0}(q, p_{\boldsymbol{\theta}}),$$

The first equality holds since $\forall k, (\partial_k \ell_{\boldsymbol{\theta}} - \partial_k \log q)^2 \geq 0$. Since for all $g_k \in \mathrm{Lip}_0^L(V)$, $g_k(\boldsymbol{z}) = 0, \forall \boldsymbol{z} \in \partial V$,

$$g_k(\boldsymbol{x}) = g_k(\boldsymbol{x}) - g_k(\boldsymbol{x}') \leq L d(\boldsymbol{x}, \boldsymbol{x}'), \forall \boldsymbol{x}' \in \partial V \implies g_k(\boldsymbol{x}) \leq L \cdot \min_{\boldsymbol{z} \in \partial V} d(\boldsymbol{x}, \boldsymbol{z}) = L g_0(\boldsymbol{x}), \tag{D.20}$$

Hence the second equality.

D.2 Proofs of Theorem 3

Proof As we stated in the proof of v Lemma 3, due to Rademacher's theorem (Evans and Gariepy, 1992), any Lipschitz function defined on a Lipschitz domain is in $H^1(V)$, hence $g_k \in H^1(V)$ and $\forall k, f_k \in H^1(V)$. Using the fact that $f_k, q, \ell_{\boldsymbol{\theta}} \in H^1(V), f_k(\boldsymbol{x}) = 0, \forall \boldsymbol{x} \in \partial V$, we can verify that $\boldsymbol{f}$ is in a Stein class of q by applying the same integration by parts techniques used in Section C. Therefore, we can write the Stein discrepancy between $p_{\boldsymbol{\theta}}$ and q as:

$$\max_{\boldsymbol{f} \in \mathcal{F}} \mathbb{E}_q \left[T_{p_{\boldsymbol{\theta}}} \boldsymbol{f} \right] = \max_{\boldsymbol{f} \in \mathcal{F}} \mathbb{E}_q \left[T_{p_{\boldsymbol{\theta}}} \boldsymbol{f} - T_q \boldsymbol{f} \right]. \tag{D.21}$$

Now we optimize (D.21) analytically using Theorem 2 in Barp et al. (2019)

$$\begin{aligned}\max_{\boldsymbol{f}\in\mathcal{F}} \mathbb{E}_q\left[T_{p_{\boldsymbol{\theta}}}\boldsymbol{f} - T_q\boldsymbol{f}\right] &= \max_{g_k\in \mathrm{Lip}_0^L(V)} \max_{\boldsymbol{h}} \mathbb{E}_q\left[\sum_k (\partial_k \ell_{\boldsymbol{\theta}} - \partial_k \log q)\cdot g_k^{\frac{1}{2}} h_k\right] \\ &= \max_{g_k\in \mathrm{Lip}_0^L(V)} \sqrt{\mathbb{E}_q\left[\sum_k (\partial_k \ell_{\boldsymbol{\theta}} - \partial_k \log q)^2 g_k\right]}.\end{aligned}$$

Applying Lemma 5, we obtain the desired result. ■

D.3 Proof of Corollary 1

Proof The proof is mostly the same as the proof of Theorem 3. However we need to prove a different version of Lemma 5 for the capped $\mathrm{Lip}_0^L(V)$ family.

Lemma 5 states $g_k(\boldsymbol{x}) \le L\cdot \min_{\boldsymbol{z}\in\partial V} \mathrm{dist}(\boldsymbol{x},\boldsymbol{z}) = g_0(\boldsymbol{x})$ for all $g_k \in \mathrm{Lip}_0^L(V)$. Further, we know that $g_k \le 1$ by definition. Therefore $g_k(\boldsymbol{x}) \le \min(1, Lg_0(\boldsymbol{x})) = \bar{g}_L$. Notice that $\bar{g}_L \in \overline{\mathrm{Lip}}_0^L(V)$. Using the same argument in Section D.1, we can see

$$\max_{g_k\in\overline{\mathrm{Lip}}_0^L(V)} \mathbb{E}_q\left[\sum_k (\partial_k \ell_{\boldsymbol{\theta}} - \partial_k \log q)^2 g_k\right] = \mathrm{FH}_{\bar{g}_L}(q, p_{\boldsymbol{\theta}}).$$

Applying this result to the last step in Section D.2, gives the desired result. ■

E. Proof of Theorem 4

We assume that

$$\mathbb{E}\Big[\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\le\delta} |\mathbb{G}_n(m_{\boldsymbol{\theta}} - m_{\boldsymbol{\theta}^*})|\Big] \le C'_{\boldsymbol{g}}\delta^{\beta}, \tag{E.22}$$

where $\mathbb{G}_n f = \frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - \mathbb{E}_q[f(X)])$ for the independent random variables $X_1,\ldots,X_n$ from q, and $C'_{\boldsymbol{g}}$ is a positive constant depending on the function $\boldsymbol{g}$ satisfying the same property as $C_{\boldsymbol{g}}$. Then, the estimation accuracy is given by the following theorem.

Theorem E.5 (Theorem 5.52 in van der Vaart (2000)) *Suppose Assumption 1, 2 and* (E.22) *hold for $\alpha > \beta > 0$ and $\alpha > 1$. Then, for any positive integer K we have*

$$P\left\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| > \frac{2^K}{n^{1/(2(\alpha-\beta))}}\right\} \le 2^{K(\beta-\alpha)}\frac{2^{2\alpha}}{2^{\alpha-1}-1}\frac{C'_{\boldsymbol{g}}}{C_{\boldsymbol{g}}}.$$

Hence, we have $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = O_p(n^{-1/(2(\alpha-\beta))})$. Our concern is the relation between the coefficient of the convergence rate and the weight function $\boldsymbol{g}$.

The upper bound of the expectation (E.22) is closely related to the covering number of the parameter space $\Theta \subset \mathbb{R}^r$. Let us define $N_{[]}(\varepsilon, \mathcal{F}, L^2(Q))$ be the bracketing number of $\mathcal{F}$

with the radius ε under the norm of $L^2(Q)$ and $J_{[]}(\delta, \mathcal{F}, L^2(Q))$ be

$$J_{[]}(\delta, \mathcal{F}, L^2(Q)) = \int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L^2(Q))}\mathrm{d}\varepsilon$$

Then, the following theorem holds.

Theorem E.6 (Corollary 19.35 of van der Vaart (2000)) *Let F be an envelope for the function class $\mathcal{F} \subset L^2(Q)$, i.e., $\sup_{f\in\mathcal{F}} \|f(\boldsymbol{x})\|_\infty \leq F(\boldsymbol{x})$ for every $\boldsymbol{x} \in V$ and suppose $\mathbb{E}_Q[|F(X)|^2] < \infty$. Then, we have*

$$\mathbb{E}\big[\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)|\big] \leq C\, J_{[]}(\|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q)),$$

where C is a universal constant.

The bracketing number of the parametric model is given by the following proposition.

Proposition 2 (Example 19.6 in van der Vaart (2000)) *The bracketing number of the parametrized loss function $m_{\boldsymbol{\theta}}(\boldsymbol{x})$ is given as follows. Let $\Theta \subset \mathrm{R}^r$ be contained in a ball of radius R. Let $\mathcal{F} = \{m_{\boldsymbol{\theta}}(\boldsymbol{x}) \,|\, \boldsymbol{\theta} \in \Theta\}$ be a function class indexed by Θ. Suppose there exists a function $\dot{m}(\boldsymbol{x})$ with $\|\dot{m}\|_{L^2(Q)} < \infty$ such that*

$$|m_{\boldsymbol{\theta}_1}(\boldsymbol{x}) - m_{\boldsymbol{\theta}_2}(\boldsymbol{x})| \leq \dot{m}(\boldsymbol{x})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$$

for all $\boldsymbol{x} \in V$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$. Then, for every $\varepsilon > 0$,

$$N_{[]}(\varepsilon\|\dot{m}\|_{L^2(Q)}, \mathcal{F}, L^2(Q)) \leq \left(1 + \frac{4R}{\varepsilon}\right)^r.$$

In our case, we need to evaluate $\mathbb{E}[\sup_{f\in\mathcal{F}_\delta} |\mathbb{G}_n(f)|]$ for

$$\mathcal{F}_\delta := \{m_{\boldsymbol{\theta}}(\boldsymbol{x}) - m_{\boldsymbol{\theta}^*}(\boldsymbol{x}) \,|\, \boldsymbol{\theta} \in \Theta,\, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq \delta\},$$

where $\boldsymbol{\theta}^*$ is the minimizer of the expected loss $M(\boldsymbol{\theta})$. The function $\dot{m}(\boldsymbol{x})$ in Proposition 2 should satisfy

$$|(m_{\boldsymbol{\theta}_1}(\boldsymbol{x}) - m_{\boldsymbol{\theta}^*}(\boldsymbol{x})) - (m_{\boldsymbol{\theta}_2}(\boldsymbol{x}) - m_{\boldsymbol{\theta}^*}(\boldsymbol{x}))| = |m_{\boldsymbol{\theta}_1}(\boldsymbol{x}) - m_{\boldsymbol{\theta}_2}(\boldsymbol{x})| \leq \dot{m}(\boldsymbol{x})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Then, Proposition 2 leads to

$$N_{[]}(\varepsilon\|\dot{m}\|_{L^2(Q)}, \mathcal{F}_\delta, L^2(Q)) \leq \left(1 + \frac{4\delta}{\varepsilon}\right)^r.$$

The envelope function $F_\delta(\boldsymbol{x})$ of $\mathcal{F}_\delta$ is given by $F_\delta(\boldsymbol{x}) = \dot{m}(\boldsymbol{x})\delta$, because

$$|m_{\boldsymbol{\theta}}(\boldsymbol{x}) - m_{\boldsymbol{\theta}^*}(\boldsymbol{x})| \leq \dot{m}(\boldsymbol{x})\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq \dot{m}(\boldsymbol{x})\delta.$$

Hence, we obtain

$$N_{[]}(\varepsilon\|F_\delta\|_{L^2(Q)}, \mathcal{F}_\delta, L^2(Q)) = N_{[]}(\varepsilon\delta\|\dot{m}\|_{L^2(Q)}, \mathcal{F}_\delta, L^2(Q)) \leq \left(1 + \frac{4}{\varepsilon}\right)^r,$$

and thus,

$$\mathbb{E}[\sup_{f\in\mathcal{F}_\delta}|\mathbb{G}_n(f)|] \leq C\delta\|\dot{m}\|_{L^2(Q)}\sqrt{r}\int_0^1\sqrt{\log\left(1+\frac{4}{\varepsilon}\right)}\mathrm{d}\varepsilon \leq C'\sqrt{r}\delta\|\dot{m}\|_{L^2(Q)}$$

holds, where C and C' are universal constants. We find that β in (E.22) is given by $\beta = 1$.

Let us evaluate the norm $\|\dot{m}\|_{L^2(Q)}$. For the function A_k and B_k in (8), we have the following inequalities

$$\begin{aligned}|A_k(\boldsymbol{x},\boldsymbol{\theta}_1) - A_k(\boldsymbol{x},\boldsymbol{\theta}_2)| &\leq \dot{A}_k(\boldsymbol{x})\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|,\\ |B_k(\boldsymbol{x},\boldsymbol{\theta}_1) - B_k(\boldsymbol{x},\boldsymbol{\theta}_2)| &\leq \dot{B}_k(\boldsymbol{x})\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|.\end{aligned} \tag{E.23}$$

It is straightforward to see that the following $\dot{m}$ satisfies the required condition:

$$\dot{m}(\boldsymbol{x}) = \sum_k\{g_k(\boldsymbol{x})\dot{A}_k(\boldsymbol{x}) + |\partial_k g_k(\boldsymbol{x})|\dot{B}_k(\boldsymbol{x})\}.$$

Cauchy-Schwarz inequality leads to $\|g_k\dot{A}_k\|_{L^2(Q)} \leq (\mathbb{E}_q[\dot{A}_k^4])^{1/4}(\mathbb{E}_q[g_k^4])^{1/4}$. The similar inequality holds for $\|\dot{B}_k\partial_k g_k\|_{L^2(Q)}$. Hence, we have

$$\|\dot{m}\|_{L^2(Q)} \leq \Gamma(\boldsymbol{g};A,B) := \sum_{k=1}^d\left\{(\mathbb{E}_q[\dot{A}_k^4]\mathbb{E}_q[g_k^4])^{1/4} + (\mathbb{E}_q[\dot{B}_k^4]\mathbb{E}_q[|\partial_k g_k|^4])^{1/4}3\right\}. \tag{E.24}$$

In summary, we have the estimation error bound in Theorem 4.

F. Some More Analysis of Weight Functions

In this subsection, we assume that the statistical model is realizable, i.e., $q = p_{\boldsymbol{\theta}^*}$ holds. Let U be a subset in the domain V. Under the regularity condition later shown in Appendix F.1, one can find that the constant $C_{\boldsymbol{g}}$ in (9) is given by

$$C_{\boldsymbol{g}} = \min_{\boldsymbol{x}\in U}\min_{k\in[d]} g_k(\boldsymbol{x}) \tag{F.25}$$

up to a constant independent of $\boldsymbol{g}$. As the continuous and non-negative weight function $g_k(\boldsymbol{x})$ can take zero only on the boundary ∂V, $C_{\boldsymbol{g}} > 0$ holds if U is a closed subset which does not include boundary points of V.

Let us consider $\Gamma(\boldsymbol{g};A,B)/C_{\boldsymbol{g}}$ in the upper bound in Theorem 4. The following theorem ensures that the minimum function

$$h(\boldsymbol{x}) = \min_{k\in[d]} g_k(\boldsymbol{x}) \tag{F.26}$$

improves the upper bound of the estimation error under some conditions on g_k.

Theorem F.7 *Let $\boldsymbol{g} = (g_1,\ldots,g_d)$ and $\boldsymbol{h} = (h,\ldots,h)$ be the weight functions, where h is defined by* (F.26)*. Suppose that g_k is differentiable on V except a measure zero set and that the set $\{\boldsymbol{x}\in V \mid g_k(\boldsymbol{x}) = g_{k'}(\boldsymbol{x}),\ \exists k,k'\in[d], k\neq k'\}$ is measure zero. We assume that*

$$|\partial_k g_k(\boldsymbol{x})| \geq |\partial_k g_{k^*}(\boldsymbol{x})|,$$

holds for $\boldsymbol{x} \in V$ and $k \in [d]$, where $k^* \in [d]$ is the number such that $h(\boldsymbol{x}) = g_{k^*}(\boldsymbol{x})$. Then, we have

$$\frac{\Gamma(\boldsymbol{h}; A, B)}{C_{\boldsymbol{h}}} \leq \frac{\Gamma(\boldsymbol{g}; A, B)}{C_{\boldsymbol{g}}},$$

where $C_{\boldsymbol{g}}$ and $C_{\boldsymbol{h}}$ are defined by (F.25).

The proof is found in Appendix F.2.

Example 4 (Bounded Rectangular Domain) *For the d-dimensional rectangular $V = \prod_{k=1}^{d}[0, c_k]$, let us consider the statistical accuracy of the generalized score matching (SM) method (Yu et al., 2019) with $g_k(\boldsymbol{x}) = \min\{x_k, c_k - x_k\}$, and $h(\boldsymbol{x}) = \min_k g_k(\boldsymbol{x})$. Note that the weight h is nothing but g_0 in* (6). *According to Theorem F.7, we find that the estimator with the weight h is superior to the estimator with the above $g_1, \ldots, g_d$ in the sense of the estimation error bound.*

Example 5 (Unit ball under p-norm) *Let us define V as the d-dimensional unit ball under p-norm, $V = \{\boldsymbol{x} \in \mathbb{R}^d | \|\boldsymbol{x}\|_p \leq 1\}$. The weight $g_k(\boldsymbol{x})$ is defined by the distance from $\boldsymbol{x}$ to the boundary of V along the k-th axis. This is expressed by*

$$g_k(\boldsymbol{x}) = \max\{|\varepsilon| \,|\, \boldsymbol{x} + \varepsilon \boldsymbol{e}_k \in V\} = \min\{(1 - \|\boldsymbol{x}_{(k)}\|_p^p)^{1/p} - x_k,\, x_k + (1 - \|\boldsymbol{x}_{(k)}\|_p^p)^{1/p}\},$$

where $\boldsymbol{e}_k$ is the unit vector along with the k-th axis, $\boldsymbol{x}_{(k)}$ is the $d-1$ dimensional vector dropped the k-th element x_k from $\boldsymbol{x}$. Some calculation yields the inequality $|\partial_k g_k(\boldsymbol{x})| = 1 \geq |\partial_k g_\ell(\boldsymbol{x})|$ for $\ell = \operatorname{argmin}_k g_k(\boldsymbol{x})$. Hence, the minimum function $h(\boldsymbol{x}) = \min_k g_k(\boldsymbol{x})$ improves the error bound of the estimator with the weight g_k. One can confirm that $h(\boldsymbol{x}) = \min_{\boldsymbol{z} \in \partial V} \|\boldsymbol{x} - \boldsymbol{z}\|_1$ holds. Likewise, for the truncated domain $V = \{\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d | \|\boldsymbol{x}\|_p \leq 1, x_i \geq c_i\}$, $c_i \in \mathbb{R}$, one can prove the inequality $|\partial_k g_k(\boldsymbol{x})| = 1 \geq |\partial_k g_\ell(\boldsymbol{x})|$. The above assertions are explained in Appendix F.3.

Under the assumption of Theorem F.7, we find that the homogeneous weight $\boldsymbol{h}$ improves the upper bound of the estimation error for the estimator with $\boldsymbol{g}$. Furthermore, the weight $\boldsymbol{h}$ corresponds to the minimum ℓ^1 distance to the boundary as shown in Example 5. To compare the distance-based homogeneous weights, we need to evaluate the estimation error upper bound for each distance. In Appendix H, we show the numerical comparison between ℓ^1 and ℓ^2 distances and numerical evaluation of the theoretical upper bound.

F.1 Derivation of (F.25)

Suppose the true probability $q(\boldsymbol{x})$ is $p_{\boldsymbol{\theta}^*}(\boldsymbol{x})$. Let U be an open subset of V. Then, we have

$$\begin{aligned} M(\boldsymbol{\theta}) - M(\boldsymbol{\theta}^*) &= \int_V \sum_k g_k(\boldsymbol{x})(\partial_k \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \partial_k \log p_{\boldsymbol{\theta}^*}(\boldsymbol{x}))^2 p_{\boldsymbol{\theta}^*}(\boldsymbol{x}) d\boldsymbol{x} \\ &\geq \int_U \sum_k g_k(\boldsymbol{x})(\partial_k \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \partial_k \log p_{\boldsymbol{\theta}^*}(\boldsymbol{x}))^2 p_{\boldsymbol{\theta}^*}(\boldsymbol{x}) d\boldsymbol{x} \\ &\geq \min_{\boldsymbol{x} \in U} \min_k \{g_k(\boldsymbol{x})\} \int_U \sum_k (\partial_k \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \partial_k \log p_{\boldsymbol{\theta}^*}(\boldsymbol{x}))^2 p_{\boldsymbol{\theta}^*}(\boldsymbol{x}) d\boldsymbol{x}. \end{aligned}$$

Suppose that the Hessian matrix of

$$\int_U \sum_k (\partial_k \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \partial_k \log p_{\boldsymbol{\theta}^*}(\boldsymbol{x}))^2 p_{\boldsymbol{\theta}^*}(\boldsymbol{x}) d\boldsymbol{x} \tag{F.27}$$

as the function of $\boldsymbol{\theta}$ is non-degenerate, there exists a constant C_0 independent of $\boldsymbol{g}$ such that the above integral is bounded below by $C_0\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2$. Hence, $C_{\boldsymbol{g}} = \min_{\boldsymbol{x}\in U} \min_k \{g_k(\boldsymbol{x})\}$ satisfies the required condition.

F.2 Proof of Theorem F.7

From the definition of h, one can find

$$C_{\boldsymbol{g}} = \min_{\boldsymbol{x}\in U} \min_k g_k(\boldsymbol{x}) = \min_{\boldsymbol{x}\in U} h(\boldsymbol{x}) = C_{\boldsymbol{h}}.$$

We evaluate $\Gamma(\boldsymbol{g}; A, B)$ and $\Gamma(\boldsymbol{h}; A, B)$. As for the derivative function, the assumption guarantees that $\partial_k h(\boldsymbol{x}) = \partial_k g_{k^*}(\boldsymbol{x})$ and $|\partial_k g_k(\boldsymbol{x})| \geq |\partial_k g_{k^*}(\boldsymbol{x})|$. Hence, we have

$$\begin{aligned}
\Gamma(\boldsymbol{g}; A, B) &= \sum_{k=1}^d (\mathbb{E}_q[\dot{A}_k^4]\mathbb{E}_q[g_k^4])^{1/4} + \sum_{k=1}^d (\mathbb{E}_q[\dot{B}_k^4]\mathbb{E}_q[|\partial_k g_k|^4])^{1/4} \\
&\geq \sum_{k=1}^d (\mathbb{E}_q[\dot{A}_k^4]\mathbb{E}_q[(\min_k g_k)^4])^{1/4} + \sum_{k=1}^d (\mathbb{E}_q[\dot{B}_k^4]\mathbb{E}_q[|\partial_k g_{k^*}|^4])^{1/4} \\
&= \Gamma(\boldsymbol{h}; A, B).
\end{aligned}$$

As a result, we obtain $\frac{\Gamma(\boldsymbol{g};A,B)}{C_{\boldsymbol{g}}} \geq \frac{\Gamma(\boldsymbol{h};A,B)}{C_{\boldsymbol{h}}}$.

F.3 Some Equations in Example 5

Proof of $|\partial_k g_k(\boldsymbol{x})| \geq |\partial_k g_{k^*}(\boldsymbol{x})|$: Without loss of generality, we suppose $\boldsymbol{x} = (x_1, \ldots, x_d) \geq \boldsymbol{0}$. Then, $g_k(\boldsymbol{x}) = (1 - \|\boldsymbol{x}_{(k)}\|_p^p)^{1/p} - x_k$ holds. Firstly, we prove that $g_\ell(\boldsymbol{x}) < g_k(\boldsymbol{x})$ leads to $x_k < x_\ell$. For the sake of simplicity let us assume $k = 1$ and $\ell = 2$. Let us define $c^p = 1 - x_3^p - \cdots - x_d^p$ for $c \geq 0$. Then, we find that $g_2(\boldsymbol{x}) < g_1(\boldsymbol{x})$ leads to $(c^p - x_1^p)^{1/p} + x_1 < (c^p - x_2^p)^{1/p} + x_2$. The function $f(x) = (c^p - x^p)^{1/p} + x$ for $0 \leq x \leq c$ is concave, takes the maximum value at $x = c/2^{1/p} (< c)$ and satisfies $f(x) = f((c^p - x^p)^{1/p})$. When y lies on the interval between x and $(c^p - x^p)^{1/p}$, $f(x) = f((c^p - x^p)^{1/p}) \leq f(y)$ holds. Suppose $x_1^p + x_2^p \leq c^2$ and $f(x_1) < f(x_2) = f((c^p - x_2^p)^{1/p})$. If x_1 lies on the interval between x_2 and $(c^p - x_2^p)^{1/p}$, $f(x_2) \leq f(x_1)$ holds and it is the contradiction. Hence we have $x_1 < x_2$. Next, we evaluate the absolute value of the derivatives. When $g_\ell(\boldsymbol{x}) < g_k(\boldsymbol{x})$, $x_k < x_\ell$ holds. For $k \neq \ell$, we obtain $|\partial_k g_k(\boldsymbol{x})| = 1$ and $|\partial_k g_\ell(\boldsymbol{x})| = |x_k|^{p-1}(1 - \|\boldsymbol{x}_{(\ell)}\|_p^p)^{1/p-1} \leq |x_\ell|^{p-1}(|x_\ell|^p)^{1/p-1} = 1$ for $p \geq 1$.

Rough sketch of the proof of $h(\boldsymbol{x}) = \min_{\boldsymbol{z}\in\partial V} \|\boldsymbol{x} - \boldsymbol{z}\|_1$: Suppose that all the vertexes of the polytope $B_1(\boldsymbol{x}, c) := \{\boldsymbol{z} \in \mathbb{R}^d | \|\boldsymbol{x} - \boldsymbol{z}\|_1 \leq c\}$ are included in V. Then, $B_1(\boldsymbol{x}, c) \subset V$ holds as V is convex. Clearly, $h(\boldsymbol{x})$ is expressed by $\sup_{c\geq 0}\{c | B_1(\boldsymbol{x}, c) \subset V\}$. This is nothing but the ℓ^1-distance from $\boldsymbol{x}$ to the boundary of V.

G. Derivation of (15)

Let us define U by $U = \{\boldsymbol{x} \in V | g_0(\boldsymbol{x}) \geq c\}$. We see that $\min_{\boldsymbol{x}\in U} \bar{g}_L(\boldsymbol{x}) = 1$ holds from Section F.1. Hence we have $C_{\bar{g}_L} = 1$. Let us evaluate the fourth order moments of $\bar{g}_L$:

$$\begin{aligned}\mathbb{E}[\bar{g}_L^4] &= L^4 \int_0^{1/L} z^4 p_{\mathrm{dist}}(z)dz + \int_{z\geq 1/L} p_{\mathrm{dist}}(z)dz \\ &= L^4 \int_0^{1/L} z^4 p_{\mathrm{dist}}(z)dz + 1 - \int_0^{1/L} p_{\mathrm{dist}}(z)dz = 1 - \frac{C}{L^{1+\beta}},\end{aligned}$$

where C is a positive constant such as $c_{b,\beta} := \frac{4b}{\beta^2+6\beta+5} \leq C \leq C_{b',\beta} := \frac{4b'}{\beta^2+6\beta+5}$. Due to the non-negativity of $\mathbb{E}[\bar{g}_L^4]$, $L^{1+\beta} \geq C$ should hold. This inequality is guaranteed from $1 \geq \int_0^{1/L} b' z^\beta dz \geq \int_0^{1/L} p_{\mathrm{dist}}(z)dz$. Let us evaluate the second term of $\Gamma(\boldsymbol{g}; A, B)$. The derivative $\partial_k \bar{g}_L$ is given by

$$\partial_k \bar{g}_L(\boldsymbol{x}) = \begin{cases} L\dfrac{x_k - \widetilde{x}_k}{\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|}, & g_0(\boldsymbol{x}) < 1/L, \\ 0, & g_0(\boldsymbol{x}) > 1/L, \end{cases}$$

where $\widetilde{\boldsymbol{x}}$ is the minimum solution of $\min_{\boldsymbol{z}\in\partial V} \|\boldsymbol{x} - \boldsymbol{z}\|$. Hence, we have

$$\mathbb{E}[|\partial_k \bar{g}_L|^4] = L^4 \int_{g_0(\boldsymbol{x})\leq 1/L} \frac{|x_k - \widetilde{x}_k|^4}{g_0(\boldsymbol{x})^4} q(\boldsymbol{x})d\boldsymbol{x} \leq L^4 \int_0^{1/L} p_{\mathrm{dist}}(z)dz \leq \frac{b'}{\beta+1} L^{-\beta+3}.$$

Next, we evaluate the lower bound of the fourth moment of the derivative $\partial_k g_k$. Jensen's inequality yields that

$$\begin{aligned}\mathbb{E}[|\partial_k \bar{g}_L|^4] &= L^4 \sum_{k=1}^d \int_{g_0(\boldsymbol{x})\leq 1/L} \frac{|x_k - \widetilde{x}_k|^4}{g_0(\boldsymbol{x})^4} q(\boldsymbol{x})d\boldsymbol{x} \\ &\geq dL^4 \left(\frac{1}{d}\sum_{k=1}^d \int_{g_0(\boldsymbol{x})\leq 1/L} \frac{|x_k - \widetilde{x}_k|^2}{g_0(\boldsymbol{x})^2} q(\boldsymbol{x})d\boldsymbol{x}\right)^2 \\ &= \frac{L^4}{d}\left(\int_{g_0(\boldsymbol{x})\leq 1/L} q(\boldsymbol{x})d\boldsymbol{x}\right)^2 \geq \frac{b^2}{d(\beta+1)^2} L^{2(1-\beta)}\end{aligned}$$

Let us define $C_A = \sum_{k=1}^d (\mathbb{E}_q[\dot{A}_k^4])^{1/4}, C_B = \sum_{k=1}^d (\mathbb{E}_q[\dot{B}_k^4])^{1/4}, c_0 = (b^2/(d(\beta+1)^2))^{1/4}$, and $c_1 = (b'/(\beta+1))^{1/4}$. Then, we have

$$C_A\left(1 - \frac{C_{b',\beta}}{L^{\beta+1}}\right)^{1/4} + C_B c_0 L^{(1-\beta)/2} \leq \frac{\Gamma(\bar{g}_L; A, B)}{C_{\bar{g}_L}} \leq C_A\left(1 - \frac{c_{b,\beta}}{L^{\beta+1}}\right)^{1/4} + C_B c_1 L^{(3-\beta)/4}.$$

H. Choice of Weight Function: ℓ^1 vs. ℓ^2 distance

In this section, we test different metrics for g_0: ℓ^1 distance $\mathrm{dist}(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\|_1$ and ℓ^2 distance $\mathrm{dist}(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\|_2$. The data set is generated from $\mathcal{N}(\mathbf{1}_d \cdot 0.5, \boldsymbol{I}_d)$, where $\mathbf{1}_d$ and $\boldsymbol{I}_d$ are the d dimensional all one vector and the $d \times d$ identity matrix respectively. The

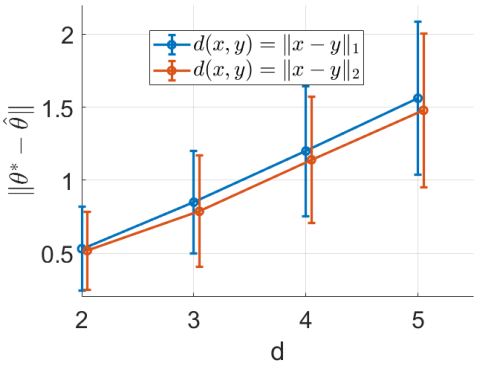


Figure 9: Estimation accuracy of TruncSM using ℓ^1 vs. ℓ^2 distance.

Table 1: Comparison of the weight h_p, $p = 1, 2$. The table shows $\{Rd\mathbb{E}_q[|h_p|^4]^{1/4} + (1-R)\sum_{k=1}^d \mathbb{E}_q[|\partial_k h_p|^4]^{1/4}\}/C_{\boldsymbol{h}_p}$ for $R = 0.2, 0.8$.

	$R = 0.2$					$R = 0.8$				
weight \ dim	2	3	4	5	6	2	3	4	5	6
h_1	3.160	4.558	5.939	7.309	8.685	1.569	2.088	2.537	2.975	3.371
h_2	3.073	4.279	5.412	6.496	7.552	1.545	2.012	2.402	2.764	3.081

truncation domain is $V = \{\boldsymbol{x} | \|\boldsymbol{x}\|_1 < 1 \text{ and } x_d > 0\}$. 150 samples are generated. We plot the estimation error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|$ and its standard deviation for both choices of distances: ℓ^1 and ℓ^2 versus the dimensionality d in Figure 9. It can be seen that the Euclidean distance, i.e., ℓ^2 distance consistently achieves slightly lower error than the ℓ^1 distance.

Let us numerically evaluate the upper bound $\Gamma(\boldsymbol{h}_p; A, B)/C_{\boldsymbol{h}_p}$ for the weight $\boldsymbol{h}_p = (h_p, \ldots, h_p)$, where $h_p(\boldsymbol{x}) = \min_{\boldsymbol{z} \in \partial V} \|\boldsymbol{x} - \boldsymbol{z}\|_p$, $p = 1, 2$. Let "$\sup_U$" be the supremum over all open subsets of V. Then, we have $\sup_U \min_{\boldsymbol{x} \in U} h_1(\boldsymbol{x}) = 1/2$ and $\sup_U \min_{\boldsymbol{x} \in U} h_2(\boldsymbol{x}) = 1/(\sqrt{d}+1)$. In order to obtain the tight bound, we use $C_{\boldsymbol{h}_1} = 1/2$ and $C_{\boldsymbol{h}_2} = 1/(\sqrt{d}+1)$ for our evaluation. For $\bar{A} = \max_k \mathbb{E}_q[|\dot{A}_k|^4]^{1/4}$, $\bar{B} = \max_k \mathbb{E}_q[|\dot{B}_k|^4]^{1/4}$, and $R = \bar{A}/(\bar{A} + \bar{B})$, clearly we have

$$\Gamma(\boldsymbol{h}_p; A, B) \leq (\bar{A} + \bar{B})\{Rd\mathbb{E}_q[|h_p|^4]^{1/4} + (1-R)\sum_{k=1}^d \mathbb{E}_q[|\partial_k h_p|^4]^{1/4}\}.$$

For $\boldsymbol{x} \sim N_d(\mathbf{1}_d \cdot 0.5, \boldsymbol{I}_d)$, we numerically evaluate $\mathbb{E}_q[|h_p|^4]$ and $\mathbb{E}_q[|\partial_k h_p|^4]$. Table 1 shows $\{Rd\mathbb{E}_q[|h_p|^4]^{1/4} + (1-R)\sum_{k=1}^d \mathbb{E}_q[|\partial_k h_p|^4]^{1/4}\}/C_{\boldsymbol{h}_p}$ of each dimension and weight for $R = 0.2, 0.8$.

In this problem setup, one can confirm that the upper bound for $p = 2$ is slightly smaller than that of $p = 1$ for all $R \in [0, 1]$, as the bound is the linear function of R. Our theoretical investigation verifies the numerical results.

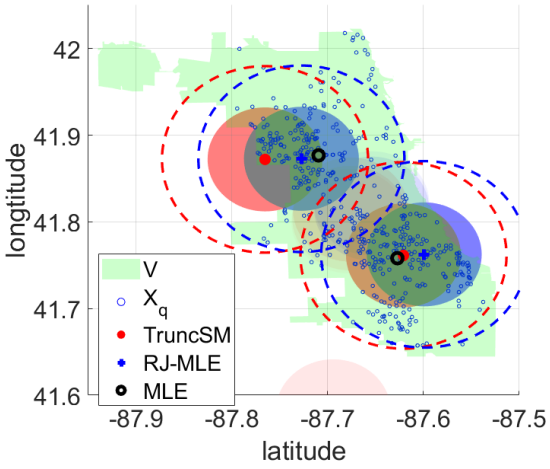


Figure 10: Chicago data set: Gaussian mixture fitting with random initialization.

I. Increasing Boundary Size in Section 8.2

When experimenting with different values of L used in capping the distance function $\bar{g}_L$, an increasing boundary size (shown by values of b in Figure 6) is used in the experiments. The two regions, the square boundary and the disjoint boundary, are created by supplying set(s) of vertices Ω, detailing the locations of the polygonal truncation domain. The variable b controls the boundary size by offsetting the supplied vertices in Ω. For the square, this was

$$\Omega_{\text{sq}} := \{(-b,-b),(-b,b),(b,b),(b,-b)\}.$$

Increasing b in this scenario leads to the corners of the square boundary being shifted by an equal amount. For the disjoint boundary, we use two sets of vertices for two disjoint domains:

$$\begin{aligned}\Omega_{\text{dis1}} :=& \{(1-b,0.5-b),(1-b,0.5),(1+b,0.5),(1+b,0.5-b)\},\\ \Omega_{\text{dis2}} :=& \{(1-b,1.5),(1-b,1.5+b),(1+b,1.5+b),(1+b,1.5)\}.\end{aligned}$$

Increasing b, in this case, will enlarge the truncation domain while the "disjoint" section in the middle remains unchanged.

J. Chicago Crime data set with Different Random Initialization

As we mentioned in Section 8.3, both TruncSM and RJ-MLE solve non-convex optimization problems, so they can return local optima. Therefore, to show the stability of both methods' solutions, in this experiment, we randomly initialize both methods with the initial point $\boldsymbol{\mu}_q + \epsilon$, where $\boldsymbol{\mu}_q$ is the mean of the data set X_q and $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d \cdot 0.06^2)$. We ran both methods 500 times and plotted two lightly shaded balls centered at the estimated mixture centers with their radius equal to the standard deviation. From the blue/red shades in Figure 10, both algorithms consistently place centers in northern and southern Chicago, and both balls are roughly centered at locations reported in Section 8.3. TruncSM also placed a

Gaussian center outside the boundary of Chicago in one of the simulations. This placement is a rare event and can be easily ruled out.

References

K. Atkinson and W. Han. *Theoretical numerical analysis*, volume 39. Springer, 2005.

A. Barp, F-X. Briol, A. Duncan, M. Girolami, and L. Mackey. Minimum stein discrepancy estimators. In *Advances in Neural Information Processing Systems 32 (NIPS 2019)*, pages 12964–12976, 2019.

W. Briec. Minimum distance to the complement of a convex set: Duality result. *Journal of Optimization Theory and Applications*, 93(2):301–319, 1997.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 2606–2615, 2016.

C. Daskalakis, T. Gouleakis, C. Tzamos, and M. Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 639–649. IEEE, 2018.

C. Daskalakis, T. Gouleakis, C. Tzamos, and M. Zampetakis. Computationally and statistically efficient truncated regression. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pages 955–960, 2019.

L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, Boca Raton, Florida, 1992.

C. J. Geyer. On the convergence of monte carlo maximum likelihood calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):261–274, 1994.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

A. Hyvärinen. Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512, 2007.

R. Kannan, L. Lovász, and M. Simonovits. Random walks and an $O*(n^5)$ volume algorithm for convex bodies. *Random Structures & Algorithms*, 11(1):1–50, 1997.

Q Liu, J. D. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, pages 276–284, 2016.

S. Lyu. Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 359–366. AUAI Press, 2009.

K. Mardia and P. E. Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

K. Mardia, J. Kent, and A. Laha. Score matching estimators for directional distributions. *arXiv:1604.08470*, 2016.

P. Milgrom and I. Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2): 583–601, 2002.

C. Moreira and J. de Uña-Álvarez. Kernel density estimation with doubly truncated data. *Electronic Journal of Statistics*, 6:501–521, 2012.

C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems 12 (NIPS 1999)*, volume 12, pages 582–588, 1999.

B. W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3): 290–295, 1976.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

S. Yu, M. Drton, and A. Shojaie. Generalized score matching for non-negative data. *Journal of Machine Learning Research*, 20(76):1–70, 2019.

S. Yu, M. Drton, and A. Shojaie. Generalized score matching for general domains. *Information and Inference: A Journal of the IMA*, 01 2021. URL `https://doi.org/10.1093/imaiai/iaaa041`.