Nonparametric Score Estimators

Yuhao Zhou [1] **Jiaxin Shi** [1] **Jun Zhu** [1]

Abstract

Estimating the score, i.e., the gradient of log density function, from a set of samples generated by an unknown distribution is a fundamental task in inference and learning of probabilistic models that involve flexible yet intractable densities. Kernel estimators based on Stein's methods or score matching have shown promise, however their theoretical properties and relationships have not been fully-understood. We provide a unifying view of these estimators under the framework of regularized nonparametric regression. It allows us to analyse existing estimators and construct new ones with desirable properties by choosing different hypothesis spaces and regularizers. A unified convergence analysis is provided for such estimators. Finally, we propose score estimators based on iterative regularization that enjoy computational benefits from curl-free kernels and fast convergence.

1. Introduction

Intractability of density functions has long been a central challenge in probabilistic learning. This may arise from various situations such as training implicit models like GANs (Goodfellow et al., 2014), or marginalizing over a non-conjugate hierarchical model, e.g., evaluating the output density of stochastic neural networks (Sun et al., 2019). In these situations, inference and learning often require evaluating such intractable densities or optimizing an objective that involves them.

Among various solutions, one important family of methods are based on *score estimation*, which rely on a key step of estimating the *score*, i.e., the derivative of the log density $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ from a set of samples drawn from some unknown probability density p. These methods include parametric score matching (Hyvärinen, 2005; Sasaki et al., 2014; Song et al., 2019), its denoising variants as autoencoders (Vincent, 2011), nonparametric score matching (Sriperumbudur et al., 2017; Sutherland et al., 2018), and kernel score estimators based on Stein's methods (Li & Turner, 2018; Shi et al., 2018). They have been successfully applied to applications such as estimating gradients of mutual information for representation learning (Wen et al., 2020), score-based generative modeling (Song & Ermon, 2019; Saremi & Hyvarinen, 2019), gradient-free adaptive MCMC (Strathmann et al., 2015), learning implicit models (Warde-Farley & Bengio, 2016), and solving intractability in approximate inference algorithms (Sun et al., 2019).

Recently, nonparametric score estimators are growing in popularity, mainly because they are flexible, have well-studied statistical properties, and perform well when samples are very limited. Despite a common goal, they have different motivations and expositions. For example, the work Sriperumbudur et al. (2017) is motivated from the density estimation perspective and the richness of kernel exponential families (Canu & Smola, 2006; Fukumizu, 2009), where the estimator is obtained by score matching. Li & Turner (2018) and Shi et al. (2018) are mainly motivated by Stein's methods. The solution of Li & Turner (2018) gives the score prediction at sample points by minimizing the kernelized Stein discrepancy (Chwialkowski et al., 2016; Liu et al., 2016) and at an out-of-sample point by adding it to the training data, while the estimator of Shi et al. (2018) is obtained by a spectral analysis in function space.

As these estimators are studied in different contexts, their relationships and theoretical properties are not fully-understood. In this paper, we provide a unifying view of them under the regularized nonparametric regression framework. This framework allows us to construct new estimators with desirable properties, and to justify the consistency and improve the convergence rate of existing estimators. It also allows us to clarify the relationships between these estimators. We show that they differ only in hypothesis spaces and regularization schemes.

Our contributions are both theoretical and algorithmic:

- We provide a unifying perspective of nonparametric score estimators. We show that the major distinction of the KEF estimator (Sriperumbudur et al., 2017) from

[1] Dept. of Comp. Sci. & Tech., BNRist Center, Institute for AI, Tsinghua-Bosch ML Center, Tsinghua University. Correspondence to: J. Zhu <dcszj@tsinghua.edu.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

the other two estimators lies in the use of curl-free kernels, while Li & Turner (2018) and Shi et al. (2018) differ mostly in regularization schemes, with the former additionally ignores a one-dimensional subspace in the hypothesis space. We provide a unified convergence analysis under the framework.

- We justify the consistency of the Stein gradient estimator (Li & Turner, 2018), although the originally proposed out-of-sample extension is heuristic and expensive. We provide a natural and principled out-of-sample extension derived from our framework. For both approaches we provide explicit convergence rates.
- From the convergence analysis we also obtain the explicit rate for Shi et al. (2018), which can be shown to improve the error bound of Shi et al. (2018).
- Our results suggest favoring curl-free estimators in high dimensions. To address the scalability challenge, we propose iterative score estimators by adopting the ν-method (Engl et al., 1996) as the regularizer. We show that the structure of curl-free kernels can further accelerate such algorithms. Inspired by a similar idea, we propose a conjugate gradient solver of KEF that is significantly faster than previous approximations.

Notation We always assume ρ is a probability measure with probability density function $p(\mathbf{x})$ supported on $\mathcal{X} \subset \mathbb{R}^d$, and $\mathcal{L}^2(\mathcal{X}, \rho; \mathbb{R}^d)$ is the Hilbert space of all square integrable functions $f : \mathcal{X} \to \mathbb{R}^d$ with inner product $\langle f, g\rangle_{\mathcal{L}^2(\mathcal{X},\rho;\mathbb{R}^d)} = \mathbb{E}_{\mathbf{x}\sim\rho}[\langle f(\mathbf{x}), g(\mathbf{x})\rangle_{\mathbb{R}^d}]$. We denote by $\langle\cdot,\cdot\rangle_\rho$ and $\|\cdot\|_\rho$ the inner product and the norm in $\mathcal{L}^2(\mathcal{X}, \rho; \mathbb{R}^d)$, respectively. We denote k as a scalar-valued kernel, and $\mathcal{K}$ as a matrix-valued kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$ satisfying the following conditions: (1) $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \mathcal{K}(\mathbf{x}', \mathbf{x})^\intercal$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$; (2) $\sum_{i,j=1}^m \mathbf{c}_i^\intercal \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)\mathbf{c}_j \geq 0$ for any $\{\mathbf{x}_i\} \subset \mathcal{X}$ and $\{\mathbf{c}_i\} \subset \mathbb{R}^d$. We denote a vector-valued *reproducing kernel Hilbert space* (RKHS) associated to $\mathcal{K}$ by $\mathcal{H}_\mathcal{K}$, which is the closure of $\left\{\sum_{i=1}^m \mathcal{K}(\mathbf{x}_i, \cdot)\mathbf{c}_i : \mathbf{x}_i \in \mathcal{X}, \mathbf{c}_i \in \mathbb{R}^d, m \in \mathbb{N}\right\}$ under the norm induced by the inner product $\langle\mathcal{K}(\mathbf{x}_i, \cdot)\mathbf{c}_i, \mathcal{K}(\mathbf{x}_j, \cdot)\mathbf{s}_j\rangle := \mathbf{c}_i^\intercal \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)\mathbf{s}_j$. We define $\mathcal{K}_\mathbf{x} := \mathcal{K}(\mathbf{x}, \cdot)$ and $[M] := \{1, \cdots, M\}$ for $M \in \mathbb{Z}_+$. For $\mathbf{A}_1, \cdots, \mathbf{A}_n \in \mathbb{R}^{s\times t}$, we use $(\mathbf{A}_1, \cdots, \mathbf{A}_n)$ to represent a block matrix $\mathbf{A} \in \mathbb{R}^{ns\times t}$ with $A_{(i-1)s+j,k}$ being the (j, k)-th component of $\mathbf{A}_i$, and we similarly define $[\mathbf{A}_1, \cdots, \mathbf{A}_n] := (\mathbf{A}_1^\intercal, \cdots, \mathbf{A}_n^\intercal)^\intercal$.

2. Background

In this section, we briefly introduce the nonparametric regression method of learning vector-valued functions (Baldassarre et al., 2012). We also review existing kernel-based approaches to score estimation.

2.1. Vector-Valued Learning

Supervised vector-valued learning amounts to learning a vector-valued function $f_\mathbf{z} : \mathcal{X} \to \mathcal{Y}$ from a training set $\mathbf{z} = \{(\mathbf{x}^m, \mathbf{y}^m)\}_{m\in[M]}$, where $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} \subseteq \mathbb{R}^q$. Here we assume the training data is sampled from an unknown distribution $\rho(\mathbf{x}, \mathbf{y})$, which can be decomposed into $\rho(\mathbf{y}|\mathbf{x})\rho_\mathcal{X}(\mathbf{x})$. A criterion for evaluating such an estimator is the mean squared error (MSE) $\mathcal{E}(f) := \mathbb{E}_{\rho(\mathbf{x},\mathbf{y})} \|f(\mathbf{x}) - \mathbf{y}\|_2^2$. It is well-known that the conditional expectation $f_\rho(\mathbf{x}) := \mathbb{E}_{\rho(\mathbf{y}|\mathbf{x})}[\mathbf{y}]$ minimizes $\mathcal{E}$. In practice, we minimize the empirical error $\mathcal{E}_\mathbf{z}(f) := \frac{1}{M}\sum_{m=1}^M \|f(\mathbf{x}^m) - \mathbf{y}^m\|_2^2$ in a certain hypothesis space $\mathcal{F}$. However, the minimization problem is typically ill-posed for large $\mathcal{F}$. Hence, it is convenient to consider the regularized problem:

$$f_{\mathbf{z},\lambda} := \arg\min_{f\in\mathcal{F}} \mathcal{E}_\mathbf{z}(f) + \lambda \|f\|_\mathcal{F}^2, \tag{1}$$

where $\|\cdot\|_\mathcal{F}$ is the norm in $\mathcal{F}$. In the vector-valued case, it is typical to consider a vector-valued RKHS $\mathcal{H}_\mathcal{K}$ associated with a matrix-valued kernel $\mathcal{K}$ as the hypothesis space. Then the estimator is $f_{\mathbf{z},\lambda} = \sum_{m=1}^M \mathcal{K}_{\mathbf{x}^m}\mathbf{c}^m$, where $\mathcal{K}_{\mathbf{x}^m}$ denotes the function $\mathcal{K}(\mathbf{x}^m, \cdot)$. $\mathbf{c}^m$ solves the linear system $(\frac{1}{M}\mathbf{K} + \lambda I)\mathbf{c} = \frac{1}{M}\mathbf{y}$ with $\mathbf{K}_{ij} = \mathcal{K}(\mathbf{x}^i, \mathbf{x}^j)$, $\mathbf{c} = (\mathbf{c}^1, \cdots, \mathbf{c}^M)$, $\mathbf{y} = (\mathbf{y}^1, \cdots, \mathbf{y}^M)$.

For convenience, we define the *sampling operator* $S_\mathbf{x} : \mathcal{H}_\mathcal{K} \to \mathbb{R}^{Mq}$ as $S_\mathbf{x}(f) := (f(\mathbf{x}^1), \cdots, f(\mathbf{x}^M))$. Its adjoint $S_\mathbf{x}^* : \mathbb{R}^{Mq} \to \mathcal{H}_\mathcal{K}$ that satisfies $\langle S_\mathbf{x}(f), \mathbf{c}\rangle_{\mathbb{R}^{Mq}} = \langle f, S_\mathbf{x}^*(\mathbf{c})\rangle_{\mathcal{H}_\mathcal{K}}, \forall f \in \mathcal{H}_\mathcal{K}, \mathbf{c} \in \mathbb{R}^{Mq}$ is $S_\mathbf{x}^*(\mathbf{c}^1, \cdots, \mathbf{c}^M) = \sum_{m=1}^M \mathcal{K}_{\mathbf{x}^m}\mathbf{c}^m$. Since $(\frac{1}{M}S_\mathbf{x}^* S_\mathbf{x} + \lambda I)f_{\mathbf{z},\lambda} = \frac{1}{M}S_\mathbf{x}^*\mathbf{K}\mathbf{c} + \lambda S_\mathbf{x}^*\mathbf{c} = \frac{1}{M}S_\mathbf{x}^*\mathbf{y}$, the estimator now can be written as $f_{\mathbf{z},\lambda} = \left(\frac{1}{M}S_\mathbf{x}^* S_\mathbf{x} + \lambda I\right)^{-1} \frac{1}{M}S_\mathbf{x}^*\mathbf{y}$. In fact, if we consider the data-free limit of (1): $\arg\min_{f\in\mathcal{H}_\mathcal{K}} \mathcal{E}(f) + \lambda\|f\|_{\mathcal{H}_\mathcal{K}}^2$, the minimizer is unique when $\lambda > 0$ and is given by $f_\lambda := (L_\mathcal{K} + \lambda I)^{-1}L_\mathcal{K}f_\rho$, where $L_\mathcal{K} : \mathcal{H}_\mathcal{K} \to \mathcal{H}_\mathcal{K}$ is the *integral operator* defined as $L_\mathcal{K}f := \int_\mathcal{X} \mathcal{K}_\mathbf{x} f(\mathbf{x})d\rho_\mathcal{X}$ (Smale & Zhou, 2007). It turns out that $\frac{1}{M}S_\mathbf{x}^* S_\mathbf{x}$ is an empirical estimate of $L_\mathcal{K}$: $\hat{L}_\mathcal{K}f := \frac{1}{M}\sum_{m=1}^M \mathcal{K}_{\mathbf{x}^m} f(\mathbf{x}^m) = \frac{1}{M}S_\mathbf{x}^* S_\mathbf{x} f$. It can also be shown that $\hat{L}_\mathcal{K}f_\rho = \frac{1}{M}S_\mathbf{x}^*\mathbf{y}$. Hence, we can write $f_{\mathbf{z},\lambda} = (\hat{L}_\mathcal{K} + \lambda I)^{-1}\hat{L}_\mathcal{K}f_\rho$.

As we have mentioned, the role of regularization is to deal with the ill-posedness. Specifically, $\hat{L}_\mathcal{K}$ is not always invertible as it has finite rank and $\mathcal{H}_\mathcal{K}$ is usually of infinite dimension. Many regularization methods are studied in the context of solving inverse problems (Engl et al., 1996) and statistical learning theory (Bauer et al., 2007). The regularization method we presented in (1) is the famous *Tikhonov regularization*, which belongs to a class of regularization techniques called spectral regularization (Bauer et al., 2007). Specifically, spectral regularization corresponds to a family of estimators defined as

$$f_{\mathbf{z},\lambda}^g := g_\lambda(\hat{L}_\mathcal{K})\hat{L}_\mathcal{K}f_\rho,$$

where $g_\lambda : \mathbb{R}^+ \to \mathbb{R}$ is a regularizer such that $g_\lambda(\hat{L}_\mathcal{K})$ approximates the inverse of $\hat{L}_\mathcal{K}$. Note that $\hat{L}_\mathcal{K}$ can be decomposed into $\sum \sigma_i \langle e_i, \cdot \rangle e_i$, where (σ_i, e_i) is a pair of eigenvalue and eigenfunction, we can define $g_\lambda(\hat{L}_\mathcal{K}) := \sum g_\lambda(\sigma_i) \langle e_i, \cdot \rangle e_i$. The Tikhonov regularization corresponds to $g_\lambda(\sigma) = (\lambda + \sigma)^{-1}$. There are several different regularizers. For example, the *spectral cut-off regularizer* is defined by $g_\lambda(\sigma) = \sigma^{-1}$ for $\sigma \geq \lambda$ and $g_\lambda(\sigma) = 0$ otherwise. We refer the readers to Smale & Zhou (2007); Bauer et al. (2007); Baldassarre et al. (2012) for more details.

2.2. Related Work

We assume $\log p(\mathbf{x})$ is differentiable, and define the *score* as $s_p := \nabla \log p$. By score estimation we aim to estimate s_p from a set of i.i.d. samples $\{\mathbf{x}^m\}_{m \in [M]}$ drawn from ρ. There have been many kernel-based score estimators studied in different contexts (Sriperumbudur et al., 2017; Sutherland et al., 2018; Li & Turner, 2018; Shi et al., 2018). Below we give a brief review of them.

Kernel Exponential Family Estimator The kernel exponential family (KEF) (Canu & Smola, 2006; Fukumizu, 2009) was originally proposed as an infinite-dimensional generalization of exponential families. It was shown to be useful in density estimation as it can approximate a broad class of densities arbitrarily well (Sriperumbudur et al., 2017). The KEF is defined as:

$$\mathcal{P}_k := \{p_f(\mathbf{x}) = e^{f(\mathbf{x}) - A(f)} : f \in \mathcal{H}_k, e^{A(f)} < \infty\},$$

where $\mathcal{H}_k$ is a scalar-valued RKHS, and $A(f) := \log \int_\mathcal{X} e^{f(\mathbf{x})} dx$ is the normalizing constant. Since $A(f)$ is typically intractable, Sriperumbudur et al. (2017) proposed to estimate f by matching the model score $\nabla \log p_f$ and the data score s_p, thus the KEF can naturally be used for score estimation (Strathmann et al., 2015). This approach works by minimizing the regularized score matching loss:

$$\min_{f \in \mathcal{H}_k} J(p \| p_f) + \lambda \|f\|^2_{\mathcal{H}_k}, \tag{2}$$

where $J(p\|q) := \mathbb{E}_p \|\nabla \log p - \nabla \log q\|_2^2$ is the Fisher divergence between p and q. Integration by parts was used to eliminated $\nabla \log p$ from $J(p\|q)$ (Hyvärinen, 2005) and the exact solution of (2) was given as follows (Sriperumbudur et al., 2017, Theorem 5):

$$\hat{f}_{p,\lambda} = \sum_{m=1}^{M} \sum_{j=1}^{d} c_{(m-1)d+j} \partial_j k(\mathbf{x}^m, \cdot) - \frac{\hat{\xi}}{\lambda}, \tag{3}$$

where $\hat{\xi}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M \sum_{j=1}^d \partial_j^2 k(\mathbf{x}^m, \cdot)$, $\mathbf{c} \in \mathbb{R}^{Md}$ is obtained by solving $(\mathbf{G} + M\lambda \mathbf{I})\mathbf{c} = \mathbf{b}/\lambda$ with $\mathbf{G}_{(m-1)d+i,(\ell-1)d+j} = \partial_i \partial_{j+d} k(\mathbf{x}^m, \mathbf{x}^\ell)$ and $\mathbf{b}_{(m-1)d+i} = \frac{1}{M} \sum_{\ell=1}^M \sum_{j=1}^d \partial_i \partial_{j+d}^2 k(\mathbf{x}^m, \mathbf{x}^\ell)$, where ∂_{j+d} denotes taking derivative w.r.t. the j-th component of the second parameter $\mathbf{x}^\ell$. This solution suffers from computational drawbacks due to the large linear system of size $Md \times Md$. Sutherland et al. (2018) proposed to use the Nyström method to accelerate KEF. Instead of minimizing the loss in the whole RKHS, they minimized it in a low dimensional subspace.

Stein Gradient Estimator The Stein gradient estimator proposed by Li & Turner (2018) is based on inverting the following generalized Stein's identity (Stein, 1981; Gorham & Mackey, 2015)

$$\mathbb{E}_p[h(\mathbf{x}) \nabla \log p(\mathbf{x})^\top + \nabla h(\mathbf{x})] = 0, \tag{4}$$

where $h : \mathcal{X} \to \mathbb{R}^d$ is a test function satisfying some regularity conditions. An empirical approximation of the identity is $-\frac{1}{M}\mathbf{HS} \approx \overline{\nabla_\mathbf{x} h}$, where $\mathbf{H} = (h(\mathbf{x}^1), \cdots, h(\mathbf{x}^M)) \in \mathbb{R}^{d \times M}$, $\mathbf{S} = (\nabla \log p(\mathbf{x}^1), \cdots, \nabla \log p(\mathbf{x}^M)) \in \mathbb{R}^{M \times d}$ and $\overline{\nabla_\mathbf{x} h} = \frac{1}{M} \sum_{m=1}^M \nabla_{\mathbf{x}^m} h(\mathbf{x}^m)$. Li & Turner (2018) proposed to minimize $\left\|\overline{\nabla_\mathbf{x} h} + \frac{1}{M}\mathbf{HS}\right\|_F^2 + \frac{\eta}{M^2} \|\mathbf{S}\|_F^2$ to estimate $\mathbf{S}$, where $\|\cdot\|_F$ denotes the Frobenius norm. The kernel trick $k(\mathbf{x}^i, \mathbf{x}^j) := h(\mathbf{x}^i)^\top h(\mathbf{x}^j)$ was then exploited to obtain the estimator. From the above we only have score estimates at the sample points. Li & Turner (2018) proposed a heuristic out-of-sample extension at $\mathbf{x}$ by adding it to $\{\mathbf{x}^m\}$ and recompute the minimizer. Such an approach is unjustified. It is still unclear whether the estimator is consistent.

Spectral Stein Gradient Estimator The Spectral Stein Gradient Estimator (SSGE) (Shi et al., 2018) was derived by a spectral analysis of the score function. Unlike Li & Turner (2018) it was shown to have convergence guarantees and principled out-of-sample extension. The idea is to expand each component of the score in a scalar-valued function space $\mathcal{L}^2(\mathcal{X}, \rho)$: $g_i(\mathbf{x}) = \sum_{j=1}^\infty \beta_{ij} \psi_j(\mathbf{x})$, where g_i is the i-th component of the score and $\{\psi_j\}$ are the eigenfunctions of the integral operator $L_k f := \int_\mathcal{X} k(\mathbf{x}, \cdot) f(\mathbf{x}) d\rho(\mathbf{x})$ associated with a scalar-valued kernel k. By using the Stein's identity in (4) Shi et al. (2018) showed that $\beta_{ij} = -\mathbb{E}_\rho[\partial_i \psi_j(\mathbf{x})]$. The Nyström method (Baker, 1977; Williams & Seeger, 2001) was then used to estimate ψ_j:

$$\hat{\psi}_j(\mathbf{x}) = \frac{\sqrt{M}}{\lambda_j} \sum_{m=1}^M k(\mathbf{x}, \mathbf{x}^m) w_{jm}, \tag{5}$$

where $\{\mathbf{x}^m\}_{m \in [M]}$ are i.i.d. samples drawn from ρ, w_{jm} is the m-th component of the eigenvector that corresponds to its j-th largest eigenvalue of the kernel matrix constructed from $\{\mathbf{x}^m\}_{m \in [M]}$. The final estimator was obtained by truncating g_i to $\sum_{j=1}^J \beta_{ij} \psi_j(\mathbf{x})$ and plugging in $\hat{\psi}_j$. Shi et al. (2018, Theorem 2) provided an error bound of SSGE depending on J and M. However, the convergence rate is still unknown.

3. Nonparametric Score Estimators

The kernel score estimators discussed in Sec. 2.2 were proposed in different contexts. The KEF estimator is motivated from the density estimation perspective, while Stein and SSGE have no explicit density models. SSGE relies on spectral analysis in the function space, while the other two are derived by minimizing a loss function. Despite sharing a common goal, it is still unclear how these estimators relate to each other. In this section, we present a unifying framework of score estimation using regularized vector-valued regression. We show that several existing kernel score estimators are special cases under the framework, which allows us to thoroughly investigate their strengths and weaknesses.

3.1. A Unifying Framework

As introduced in Sec. 2.2, the goal is to estimate the score s_p from a set of i.i.d. samples $\{\mathbf{x}^m\}_{m\in[M]}$ drawn from ρ. We first consider the ideal case where we have the ground truth values of s_p at the sample locations. Then we can estimate s_p with vector-valued regression as described in Sec. 2.1:

$$\hat{s}_{p,\lambda} = \operatorname*{arg\,min}_{s\in\mathcal{H}_\mathcal{K}} \frac{1}{M}\sum_{m=1}^{M} \|s(\mathbf{x}^m) - s_p(\mathbf{x}^m)\|_2^2 + \frac{\lambda}{2}\|s\|_{\mathcal{H}_\mathcal{K}}^2. \tag{6}$$

The solution is given by $\hat{s}_{p,\lambda} = (\hat{L}_\mathcal{K} + \lambda I)^{-1}\hat{L}_\mathcal{K} s_p$. We could replace the Tikhonov regularizer with other spectral regularization, for which the general solution is

$$\hat{s}^g_{p,\lambda} := g_\lambda(\hat{L}_\mathcal{K})\hat{L}_\mathcal{K} s_p. \tag{7}$$

In reality, the values of s_p at $\mathbf{x}^{1:M}$ are unknown and we cannot compute $\hat{L}_\mathcal{K} s_p$ as $\frac{1}{M}\sum_{m=1}^M \mathcal{K}_{\mathbf{x}^m} s_p(\mathbf{x}^m)$. Fortunately, we could exploit integration by parts to avoid this problem. Under some mild regularity conditions (Assumptions B.1-B.3), we have

$$L_\mathcal{K} s_p = \mathbb{E}_\rho[\mathcal{K}_\mathbf{x}\nabla\log p(\mathbf{x})] = -\mathbb{E}_\rho[\operatorname{div}_\mathbf{x}\mathcal{K}_\mathbf{x}^\intercal],$$

where the divergence of $\mathcal{K}_\mathbf{x}^\intercal$ is defined as a vector-valued function, whose i-th component is the divergence of the i-th column of $\mathcal{K}_\mathbf{x}^\intercal$. The empirical estimate $\hat{L}_\mathcal{K} s_p$ is then available as $-\frac{1}{M}\sum_{m=1}^M \operatorname{div}_{\mathbf{x}^m}\mathcal{K}_{\mathbf{x}^m}^\intercal$, which leads to the following general formula of nonparametric score estimators:

$$\hat{s}^g_{p,\lambda} = -g_\lambda(\hat{L}_\mathcal{K})\hat{\zeta}, \tag{8}$$

where $\hat{\zeta} := \frac{1}{M}\sum_{m=1}^M \operatorname{div}_{\mathbf{x}^m}\mathcal{K}_{\mathbf{x}^m}^\intercal$.

3.2. Regularization Schemes

We now derive the final form of the estimator under three regularization schemes (Bauer et al., 2007). The choice of regularization will impact the convergence rate of the estimator, which will be studied in Sec. 4.

Theorem 3.1 (Tikhonov Regularization)**.** *Let $\hat{s}^g_{p,\lambda}$ be defined as in (8), and $g_\lambda(\sigma) = (\sigma+\lambda)^{-1}$. Then*

$$\hat{s}^g_{p,\lambda}(\boldsymbol{x}) = \boldsymbol{K}_{\boldsymbol{x}\boldsymbol{X}}\boldsymbol{c} - \hat{\zeta}(\boldsymbol{x})/\lambda, \tag{9}$$

where $\boldsymbol{c}$ is obtained by solving

$$(\boldsymbol{K} + M\lambda I)\boldsymbol{c} = \boldsymbol{h}/\lambda. \tag{10}$$

Here $\boldsymbol{c}\in\mathbb{R}^{Md}$, $\boldsymbol{h} = (\hat{\zeta}(\boldsymbol{x}^1),\cdots,\hat{\zeta}(\boldsymbol{x}^M))\in\mathbb{R}^{Md}$, $\boldsymbol{K}_{\boldsymbol{x}\boldsymbol{X}} = [\mathcal{K}(\boldsymbol{x},\boldsymbol{x}^1),\cdots,\mathcal{K}(\boldsymbol{x},\boldsymbol{x}^M)]\in\mathbb{R}^{d\times Md}$, and $\boldsymbol{K}\in\mathbb{R}^{Md\times Md}$ is given by $\boldsymbol{K}_{(m-1)d+i,(\ell-1)d+j} = \mathcal{K}(\boldsymbol{x}^m,\boldsymbol{x}^\ell)_{ij}$.

The proof is given in appendix C.4, where the general representer theorem (Sriperumbudur et al., 2017, Theorem A.2) is used to show that the solution lies in the subspace generated by

$$\{\mathcal{K}_{\mathbf{x}^m}\mathbf{c}_m : m\in[M], \mathbf{c}_m\in\mathbb{R}^d\}\cup\{\hat{\zeta}\}. \tag{11}$$

Unlike the Tikhonov regularizer that shifts all eigenvalues simultaneously, the *spectral cut-off regularization* sets $g_\lambda(\sigma) = \sigma^{-1}$ for $\sigma\ge\lambda$, and $g_\lambda(\sigma) = 0$ otherwise. To obtain such estimator, we need the following lemma that relates the spectral properties of $\mathbf{K}$ and $\hat{L}_\mathcal{K}$.

Lemma 3.2. *Let σ be a non-zero eigenvalue of $\frac{1}{M}\boldsymbol{K}$ such that $\frac{1}{M}\boldsymbol{K}\boldsymbol{u} = \sigma\boldsymbol{u}$, where $\boldsymbol{u}\in\mathbb{R}^{Md}$ is the unit eigenvector. Then σ is an eigenvalue of $\hat{L}_\mathcal{K}$ and the corresponding unit eigenfunction is*

$$v = \frac{1}{\sqrt{M\sigma}}\sum_{m=1}^M \mathcal{K}_{\boldsymbol{x}^m}\boldsymbol{u}^{(m)},$$

where $\boldsymbol{u}$ is splitted into $(\boldsymbol{u}^{(1)},\cdots,\boldsymbol{u}^{(M)})$ and $\boldsymbol{u}^{(i)}\in\mathbb{R}^d$.

The lemma is a direct generalization of Rosasco et al. (2010, Proposition 9) to vector-valued operators.

Theorem 3.3 (Spectral Cut-Off Regularization)**.** *Let $\hat{s}^g_{p,\lambda}$ be defined as in (8), and*

$$g_\lambda(\sigma) = \begin{cases}\sigma^{-1} & \sigma > \lambda,\\ 0 & \sigma\le\lambda.\end{cases}$$

Let $(\sigma_j,\boldsymbol{u}_j)_{j\ge1}$ be the eigenvalue and eigenvector pairs that satisfy $\frac{1}{M}\boldsymbol{K}\boldsymbol{u}_j = \sigma_j\boldsymbol{u}_j$. Then we have

$$\hat{s}^g_{p,\lambda}(\boldsymbol{x}) = -\boldsymbol{K}_{\boldsymbol{x}\boldsymbol{X}}\left(\sum_{\sigma_j\ge\lambda}\frac{\boldsymbol{u}_j\boldsymbol{u}_j^\intercal}{M\sigma_j^2}\right)\boldsymbol{h}, \tag{12}$$

where $\boldsymbol{K}_{\boldsymbol{x}\boldsymbol{X}}$ and $\boldsymbol{h}$ are defined as in Theorem 3.1.

Apart from the above methods with closed-form solutions, early stopping of iterative solvers like gradient descent can also play the role of regularization (Engl et al., 1996). Iterative methods replace the expensive inversion or eigendecomposition of the $Md\times Md$ size kernel matrix with fast matrix-vector multiplication. In Sec. 3.5 we show that such methods can be further accelerated by utilizing the structure of our kernel matrix.

We consider two iterative methods: the Landweber iteration and the ν-method (Engl et al., 1996). The Landweber iteration solves $\hat{L}_{\mathcal{K}} s_p = -\hat{\zeta}$ with the fixed-point iteration:

$$\hat{s}_p^{(t+1)} := \hat{s}_p^{(t)} - \eta\left(\hat{\zeta} + \hat{L}_{\mathcal{K}}\hat{s}_p^{(t)}\right), \tag{13}$$

where η is a step-size parameter. It can be regarded as using the following regularization:

Theorem 3.4 (Landweber Iteration)**.** *Let $\hat{s}^g_{p,\lambda}$ and $\hat{s}^{(k)}_p$ be defined as in (8) and (13), respectively. Let $\hat{s}^{(0)} = 0$ and $g_\lambda(\sigma) = \eta\sum_{i=0}^{t-1}(1-\eta\sigma)^i$, where $t := \lfloor \lambda^{-1} \rfloor$. Then,*

$$\hat{s}^g_{p,\lambda} = \hat{s}^{(t)}_p = -t\eta\hat{\zeta} + \boldsymbol{K}_{\boldsymbol{xX}}\boldsymbol{c}_t,$$

where $\boldsymbol{c}_0 = 0$, $\boldsymbol{c}_{t+1} = (\mathbf{I}_d - \eta\boldsymbol{K}/M)\boldsymbol{c}_t - t\eta^2\boldsymbol{h}/M$, and $\boldsymbol{K}, \boldsymbol{K}_{\boldsymbol{xX}}, \boldsymbol{h}$ are defined as in Theorem 3.1.

The Landweber iteration often requires a large number of iterations. An accelerated version of it is the ν-method, where ν is a parameter controlling the maximal convergence rate (see Sec. 4). The regularizer of the ν-method can be represented by a family of polynomials $g_\lambda(\sigma) = \mathrm{poly}(\sigma)$. These polynomials approximate $1/\sigma$ better than those in the Landweber iteration. As a result, the ν-method only requires a polynomial of degree $\lfloor \lambda^{-1/2} \rfloor$ to define g_λ, which significantly reduces the number of iterations (Engl et al., 1996; Bauer et al., 2007). The next iterate of the ν-method can be generated by the current and the previous ones:

$$\hat{s}_p^{(t+1)} = \hat{s}_p^{(t)} + u_t(\hat{s}_p^{(t)} - \hat{s}_p^{(t-1)}) - \omega_t(\hat{\zeta} + \hat{L}_{\mathcal{K}}\hat{s}_p^{(t)}),$$

where u_t, ω_t are carefully chosen constants (Engl et al., 1996, Algorithm 6.13). We describe the full algorithm in Example C.4 (appendix C.4.3).

3.3. Hypothesis Spaces

In this framework, the hypothesis space is characterized by the matrix-valued kernel that induces the RKHS (Alvarez et al., 2012). Below we discuss two choices of the kernel: the diagonal ones are computationally more efficient, while curl-free kernels capture the conservative property of score vector fields.

Diagonal Kernels The simplest way to define a diagonal matrix-valued kernel is $\mathcal{K}(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{y})\mathbf{I}_d$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a scalar-valued kernel. This induces a product RKHS $\mathcal{H}_k^d := \otimes_{i=1}^d \mathcal{H}_k$ where all output dimensions of a function are independent. In this case the kernel matrix for $\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^M)$ is $\mathbf{K} = k(\mathbf{X}, \mathbf{X}) \otimes \mathbf{I}_d$, where $k(\mathbf{X}, \mathbf{X})$ denotes the Gram matrix of the scalar-valued kernel k. Therefore, the computational cost of matrix inversion and eigendecomposition is the same as in the scalar-valued case. On the other hand, the independence assumption may not hold for score functions, whose output dimensions are correlated as they form the gradient of the log density. As we shall see in Sec. 5, such misspecification of the hypothesis space degrades the performance in high dimensions.

Curl-Free Kernels Noticing that score vector fields are gradient fields, we can use curl-free kernels (Fuselier Jr, 2007; Macêdo & Castro, 2010) to capture this property. A curl-free kernel can be constructed from the negative Hessian of a translation-invariant kernel $k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x} - \mathbf{y})$: $\mathcal{K}_{\mathrm{cf}}(\mathbf{x}, \mathbf{y}) := -\nabla^2\phi(\mathbf{x} - \mathbf{y})$, where $\phi : \mathcal{X} \to \mathbb{R} \in C^2$. It is easy to see that $\mathcal{K}_{\mathrm{cf}}$ is positive definite. A nice property of $\mathcal{H}_{\mathcal{K}_{\mathrm{cf}}}$ is that any element in it is a gradient of some function. To see this, notice that the j-th column of $\mathcal{K}_{\mathrm{cf}}$ is $-\nabla(\partial_j\phi)$ and each element in $\mathcal{H}_{\mathcal{K}_{\mathrm{cf}}}$ is a linear combination of columns of $\mathcal{K}_{\mathrm{cf}}$. We also note that the unnormalized log-density function can be recovered from the estimated score when using curl-free kernels (see appendix C.3). The cost of inversion and eigendecomposition of the kernel matrix is $O(M^3d^3)$, compared to $O(M^3)$ for diagonal kernels.

3.4. Examples

In the following we provide examples of nonparametric score estimators derived from the framework. We show that existing estimators can be recovered with certain types of kernels and regularization schemes (Table 1).

Example 3.5 (KEF)**.** Consider using curl-free kernels for the Tikhonov regularized estimator in (9). By substituting $\mathcal{K}_{\mathrm{cf}}(\mathbf{x}, \mathbf{y}) = -\nabla^2\phi(\mathbf{x} - \mathbf{y})$ for $\mathcal{K}$, we get

$$\hat{s}^g_{p,\lambda}(\mathbf{x}) = -\sum_{m=1}^{M}\sum_{j=1}^{d} c_{(m-1)d+j}\nabla\partial_j\phi(\mathbf{x} - \mathbf{x}^m) - \frac{\hat{\zeta}_{\mathrm{cf}}(\mathbf{x})}{\lambda},$$

where $\zeta_{\mathrm{cf}}(\mathbf{x})_i := -\frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{d}\partial_i\partial_j^2\phi(\mathbf{x} - \mathbf{x}^m)$. Noticing that $\mathcal{K}_{\mathrm{cf}}(\mathbf{x}, \mathbf{y})_{ij} = -\partial_i\partial_j\phi(\mathbf{x} - \mathbf{y}) = \partial_i\partial_{j+d}k(\mathbf{x}, \mathbf{y})$, we could check that the definition of $\mathbf{c}$ here, which follows from (9), is the same as in (3). Thus by comparing with (3), we have

$$\hat{s}^g_{p,\lambda}(\mathbf{x}) = \nabla\hat{f}_{p,\lambda}(\mathbf{x}) = \nabla\log p_{\hat{f}_{p,\lambda}}(\mathbf{x}). \tag{14}$$

Therefore, *the KEF estimator is equivalent to choosing curl-free kernels and the Tikhonov regularization in* (8).

We note that, although the solutions are equivalent, the space $\{\nabla\log p_f, f \in \mathcal{H}_k\}$ looks different from the curl-free RKHS constructed from the negative Hessian of k. Such equivalence of regularized minimization problems may be of independent interest.

Example 3.6 (SSGE)**.** For the estimator (12) obtained from the spectral cut-off regularization. Consider letting

$\mathcal{K}(\mathbf{x},\mathbf{y}) = k(\mathbf{x},\mathbf{y})\mathbf{I}_d$. Then $\mathbf{K} = k(\mathbf{X},\mathbf{X}) \otimes \mathbf{I}_d$, and it can be decomposed as $\sum_{m=1}^{M}\sum_{i=1}^{d} \lambda_m(\mathbf{w}_m\mathbf{w}_m^\top \otimes \mathbf{e}_i\mathbf{e}_i^\top)$, where $\{(\lambda_m, \mathbf{w}_m)\}$ is the eigenpairs of $k(\mathbf{X},\mathbf{X})$ with $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_M$ and $\{\mathbf{e}_i\}$ is the standard basis of $\mathbb{R}^d$. The estimator reduces to

$$\hat{s}^{g}_{p,\lambda}(\mathbf{x})_i = -k(\mathbf{x},\mathbf{X})\left(\sum_{\lambda_j \ge \lambda} \frac{\mathbf{w}_j\mathbf{w}_j^\top}{\lambda_j^2}\right)\mathbf{r}_i, \qquad (15)$$

where $\frac{1}{M}\mathbf{r}_i := (h_i, h_{d+i}, \cdots, h_{(M-1)d+i})$. When we choose $\lambda = \lambda_J$, simple calculations (see appendix C.2) show that (15) equals the SSGE estimator $\hat{s}^{g}_{p,\lambda}(\mathbf{x})_i = -\frac{1}{M}\sum_{j=1}^{J}\sum_{m=1}^{M} \partial_i\hat{\psi}_j(\mathbf{x}^m)\hat{\psi}_j(\mathbf{x})$, where $\hat{\psi}_j$ is defined as in (5). Therefore, *SSGE is equivalent to choosing the diagonal kernel $\mathcal{K}(\boldsymbol{x},\boldsymbol{y}) = k(\boldsymbol{x},\boldsymbol{y})\boldsymbol{I}_d$ and the spectral cut-off regularization in* (8).

Example 3.7 (Stein)**.** We consider modifying the Tikhonov regularizer to $g_\lambda(\sigma) = (\lambda+\sigma)^{-1}\mathbf{1}_{\{\sigma>0\}}$. In this case, we obtain an estimator $\hat{s}^{g}_{p,\lambda}(\mathbf{x}) = -\mathbf{K}_{\mathbf{x}\mathbf{X}}\mathbf{K}^{-1}(\frac{1}{M}\mathbf{K}+\lambda I)^{-1}\mathbf{h}$ by Lemma C.2. At sample points, the estimated score is $-(\frac{1}{M}\mathbf{K}+\lambda I)^{-1}\mathbf{h}$, which coincides with the Stein gradient estimator. This suggests a principled out-of-sample extension of the Stein gradient estimator.

To gain more insights, we consider to minimize (6) in the subspace generated by $\{\mathcal{K}_{\mathbf{x}^m}\mathbf{c}_m : m \in [M], \mathbf{c}_m \in \mathbb{R}^d\}$. Compared with (11), the one-dimensional subspace $\mathbb{R}\hat{\zeta}$ is ignored. We could check that (in appendix C.2) this is equivalent to exploiting the previous mentioned regularizer. Therefore, *the Stein estimator is equivalent to using the diagonal kernel $\mathcal{K}(\boldsymbol{x},\boldsymbol{y}) = k(\boldsymbol{x},\boldsymbol{y})\boldsymbol{I}_d$ and the Tikhonov regularization with a one-dimensional subspace ignored.*

All the above examples can be extended to use a subset of the samples with Nyström methods (Williams & Seeger, 2001). Specifically, we can modify the general formula in (8) as $\hat{s}^{g,\mathbf{Z}}_{p,\lambda} = -g_\lambda(P_{\mathbf{Z}}\hat{L}_{\mathcal{K}}P_{\mathbf{Z}})P_{\mathbf{Z}}\hat{\zeta}$, where $P_{\mathbf{Z}} : \mathcal{H}_{\mathcal{K}} \to \mathcal{H}_{\mathcal{K}}$ is the projection onto a low-dimensional subspace generated by the subset $\mathbf{Z}$. When the curl-free kernel and the same truncated Tikhonov regularizer as in Example 3.7 are used, this estimator is equivalent to the Nyström KEF (NKEF) (Sutherland et al., 2018). More details can be found in appendix C.1.

3.5. Scalability

When using curl-free kernels, we need to deal with an $Md \times Md$ matrix. In such cases, the Tikhonov and the spectral cut-off regularization cost $O(M^3d^3)$ and have difficulties scaling with the sample size and the input dimensions. Fortunately, as the unifying perspective suggests, we could modify the regularization schemes with iterative methods that only require matrix-vector multiplications, e.g., the Landweber iteration and the ν-method (see Sec. 3.2).

Table 1. Existing nonparametric score estimators, their kernel types, and regularization schemes. ϕ is from $k(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x}-\mathbf{y})$.

ALGORITHM	KERNEL	REGULARIZER
SSGE	$k(\mathbf{x},\mathbf{y})\mathbf{I}_d$	$\mathbf{1}_{\{\sigma\ge\lambda\}}\sigma^{-1}$
Stein	$k(\mathbf{x},\mathbf{y})\mathbf{I}_d$	$\mathbf{1}_{\{\sigma>0\}}(\lambda+\sigma)^{-1}$
KEF	$-\nabla^2\phi(\mathbf{x}-\mathbf{y})$	$(\lambda+\sigma)^{-1}$
NKEF	$-\nabla^2\phi(\mathbf{x}-\mathbf{y})$	$\mathbf{1}_{\{\sigma>0\}}(\lambda+\sigma)^{-1}$

Interestingly, we get further acceleration by utilizing the structure of curl-free kernels.

Example 3.8 (Iterative curl-free estimators)**.** We observe that when using a curl-free kernel $\mathcal{K}_{\mathrm{cf}}$ constructed from a radial scalar-valued kernel $k(\mathbf{x},\mathbf{y}) = \phi(\|\mathbf{x}-\mathbf{y}\|)$,

$$\mathcal{K}_{\mathrm{cf}}(\mathbf{x},\mathbf{y}) = \left(\frac{\phi'}{r^3} - \frac{\phi''}{r^2}\right)\mathbf{r}\mathbf{r}^\top - \frac{\phi'}{r}\mathbf{I},$$

where $\mathbf{r} = \mathbf{x}-\mathbf{y}$, $r = \|\mathbf{r}\|_2$. Consider in matrix-vector multiplications, for a vector $\mathbf{a} \in \mathbb{R}^d$, $\mathcal{K}_{\mathrm{cf}}(\mathbf{x},\mathbf{y})\mathbf{a}$ can be computed as $\left(\frac{\phi'}{r^3} - \frac{\phi''}{r^2}\right)(\mathbf{r}^\top\mathbf{a})\mathbf{r} - \frac{\phi'}{r}\mathbf{a}$, where only a vector-vector multiplication is required with time complexity $O(d)$, compared to general $O(d^2)$. Thus, we only need $O(M^2d)$ time to compute $\mathbf{Kb}$ for any $\mathbf{b} \in \mathbb{R}^{Md}$. In practice, we only need to store samples for computing $\mathbf{Kb}$ instead of constructing the whole kernel matrix. This reduces the memory usage from $O(M^2d^2)$ to $O(M^2d)$.

We note that the same idea in Example 3.8 can be used to accelerate the KEF estimator if we adopt the conjugate gradient methods (Van Loan & Golub, 1983) to solve (10), because we have shown that the KEF estimator is equivalent to our Tikhonov regularized estimators with curl-free kernels. As we shall see in experiments, this method is extremely fast in high dimensions.

4. Theoretical Properties

In this section, we give a general theorem on the convergence rate of score estimators in our framework, which provides a tighter error bound of SSGE (Shi et al., 2018). We also investigate the case where samples are corrupted by a small set of points, and provide the convergence rate of the heuristic out-of-sample extension proposed in Li & Turner (2018). Proofs and assumptions are deferred to appendix B.

First, we follow Bauer et al. (2007); Baldassarre et al. (2012) to characterize the regularizer.

Definition 4.1 (Bauer et al. (2007))**.** We say a family of functions $g_\lambda : [0,\kappa^2] \to \mathbb{R}$, $0 < \lambda \le \kappa^2$ is a *regularizer* if there are constants B, D, γ such that $\sup_{0<\sigma\le\kappa^2}|\sigma g_\lambda(\sigma)| \le D$, $\sup_{0<\sigma\le\kappa^2}|g_\lambda(\sigma)| \le B/\lambda$ and $\sup_{0<\sigma\le\kappa^2}|1-\sigma g_\lambda(\sigma)| \le \gamma$. The *qualification* of g_λ is the maximal r such that $\sup_{0<\sigma\le\kappa^2}|1-\sigma g_\lambda(\sigma)|\sigma^r \le \gamma_r\lambda^r$, where γ_r does not depend on λ.

Now, we can use the idea of Bauer et al. (2007, Theorem 10) to obtain an error bound of our estimator.

Theorem 4.2. *Assume Assumptions B.1-B.5 hold. Let $\bar{r}$ be the qualification of the regularizer g_λ, and $\hat{s}^g_{p,\lambda}$ be defined as in (8). Suppose there exists $f_0 \in \mathcal{H}_\mathcal{K}$ such that $s_p = L^r_\mathcal{K} f_0$ for some $r \in [0, \bar{r}]$. Then we have for $\lambda = M^{-\frac{1}{2r+2}}$,*

$$\|\hat{s}^g_{p,\lambda} - s_p\|_{\mathcal{H}_\mathcal{K}} = O_p\left(M^{-\frac{r}{2r+2}}\right),$$

and for $r \in [0, \bar{r} - 1/2]$, we have

$$\|\hat{s}^g_{p,\lambda} - s_p\|_\rho = O_p\left(M^{-\frac{r+1/2}{2r+2}}\right),$$

where O_p is the Big-O notation in probability.

Note the qualification impacts the maximal convergence rate. As the qualification of Tikhonov regularization is 1, from the error bound, we observe the well-known saturation phenomenon of Tikhonov regularization (Engl et al., 1996), i.e., the convergence rate does not improve even if $s_p = L^r_\mathcal{K} f_0$ for $r > 1$. To alleviate this, we can choose the regularizer with a larger qualification. For example, the spectral cut-off regularization and the Landweber iteration have qualification ∞, and the ν-method has qualification ν. This suggests that the ν-method is appealing as it has a smaller iteration number than the Landweber iteration and a better maximal convergence rate than the Tikhonov regularization.

Remark 4.3 (Stein)**.** The consistency and convergence rate of the Stein estimator and its out-of-sample extension suggested in Example 3.7 follow from Theorem 4.2. The rate in $\|\cdot\|_\rho$ is $O_p\left(M^{-\theta_1}\right)$, where $\theta_1 \in [1/4, 1/3]$. The convergence rate of the original out-of-sample extension in Li & Turner (2018) will be given in Corollary 4.7.

Remark 4.4 (SSGE)**.** From Theorem 4.2, the convergence rate in $\|\cdot\|_\rho$ of SSGE is $O_p(M^{-\theta_2})$, where $\theta_2 \in [1/4, 1/2)$, which improves Shi et al. (2018, Theorem 2). To see this, we assume the eigenvalues of $L_\mathcal{K}$ are $\mu_1 > \mu_2 > \cdots$ and they decay as $\mu_J = O(J^{-\beta})$. The error bound provided by Shi et al. (2018) is

$$\|\hat{s}_{p,\lambda} - s_p\|^2_\rho = O_p\left(\frac{J^2}{\mu_J(\mu_J - \mu_{J+1})^2 M} + \mu_J\right).$$

We can choose $J = M^{\frac{1}{4(\beta+1)}}$ to obtain $\|\hat{s}_{p,\lambda} - s_p\|_\rho = O_p(M^{-\frac{\beta}{8(\beta+1)}})$. The convergence rate is slower than $O_p(M^{-1/4})$, the worst case of Theorem 4.2.

Remark 4.5 (KEF)**.** Compared with Theorem 7(ii) in Sriperumbudur et al. (2017), where they bound the Fisher divergence, which is the square of the L^2-norm in our Theorem 4.2, we see that the two results are exactly the same. The rate in this norm is $O_p\left(M^{-\theta_3}\right)$, where $\theta_3 \in [1/4, 1/3]$.

Next, we consider the case where estimators are not obtained from i.i.d. samples. Specifically, we consider how the convergence rate is affected when our data is the mixture of a set of i.i.d. samples and a set of fixed points.

Theorem 4.6. *Under the same assumption of Theorem 4.2, we define $g_\lambda(\sigma) := (\lambda + \sigma)^{-1}$, and choose $\mathbf{Z} := \{\boldsymbol{z}^n\}_{n\in[N]} \subseteq \mathcal{X}$. Let $\mathbf{Y} := \{\boldsymbol{y}^m\}_{m\in[M]}$ be a set of i.i.d. samples drawn from ρ, and $\hat{s}_{p,\lambda,\mathbf{Z}}$ be defined as in (8) with $\mathbf{X} = \mathbf{Z} \cup \mathbf{Y}$. Suppose $N = O(M^\alpha)$, then we have for $\lambda = M^{-\frac{1}{2r+2}}$,*

$$\sup_{\mathbf{Z}} \|\hat{s}_{p,\lambda,\mathbf{Z}} - s_p\|_{\mathcal{H}_\mathcal{K}} = O_p\left(M^{-\frac{r}{2r+2}}\right) + O(M^{\alpha - \frac{r}{r+1}}),$$

where the $\sup_{\mathbf{Z}}$ is taken over all $\{\boldsymbol{z}^n\}_{n\in[N]} \subseteq \mathcal{X}$.

Proof Outline. Define $T_\mathbf{Z} := \frac{1}{N} S^*_\mathbf{Z} S_\mathbf{Z}$, where $S_\mathbf{Z}$ is the sampling operator. Let $\hat{s}_{p,\lambda}$ be defined as in (8) with $\mathbf{X} = \mathbf{Y}$. We can write the estimator as $\hat{s}_{p,\lambda,\mathbf{Z}} := g_\lambda(\hat{L}_\mathcal{K} + R_\mathbf{Z})(\hat{L}_\mathcal{K} + R_\mathbf{Z})s_p$, where $R_\mathbf{Z} := \frac{N}{M+N}(T_\mathbf{Z} - \hat{L}_\mathcal{K})$, and bound $\|\hat{s}_{p,\lambda,\mathbf{Z}} - \hat{s}_{p,\lambda}\|$ by $\|(g_\lambda(\hat{L}_\mathcal{K} + R_\mathbf{Z}) - g_\lambda(\hat{L}_\mathcal{K}))\hat{L}_\mathcal{K} s_p\| + \|g_\lambda(\hat{L}_\mathcal{K} + R_\mathbf{Z})R_\mathbf{Z} s_p\|$. It can be shown that the first term is $O\left(NM^{-1}\lambda^{-2}\right)$, and the second term is $O\left(NM^{-1}\lambda^{-1}\right)$. Combining these with Theorem 4.2, we finish the proof. □

From Theorem 4.6, we see that the convergence rate is not affected when data is corrupted by at most $O(M^{\frac{r}{2r+2}})$ points. Under the same notation of this theorem, the out-of-sample extension of the Stein estimator proposed in Li & Turner (2018) can be written as $\hat{s}_{p,\lambda,\mathbf{x}}(\mathbf{x})$, which corrupts the i.i.d. data by a single test point. Then we can obtain the following bound for this estimator.

Corollary 4.7. *With the same assumptions and notations of Theorem 4.6, we have*

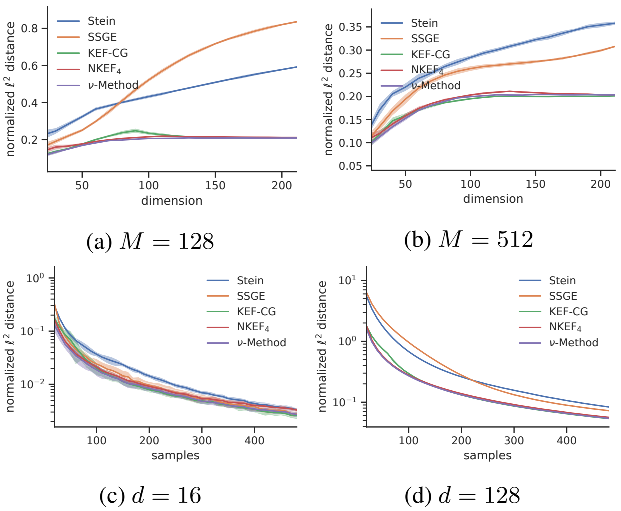
$$\sup_{\boldsymbol{x}\in\mathcal{X}} \|\hat{s}_{p,\lambda,\boldsymbol{x}}(\boldsymbol{x}) - s_p(\boldsymbol{x})\|_2 = O_p(M^{-\frac{r}{2r+2}}).$$

5. Experiments

We evaluate our estimators on both synthetic and real data. In Sec. 5.1, we consider a challenging grid distribution as described in the experiment of Sutherland et al. (2018) to test the accuracy of nonparametric score estimators in high dimensions and out-of-sample points, In Sec. 5.2 we train Wasserstein autoencoders (WAE) with score estimation and compare the accuracy and the efficiency of different estimators. We mainly compare the following score estimators[1]:

Existing nonparametric estimators: Stein (Li & Turner, 2018), SSGE (Shi et al., 2018), KEF (Sriperumbudur et al., 2017), and its low rank approximation NKEF$_\alpha$ (Sutherland et al., 2018), where α represents to use $\alpha M/10$ samples.

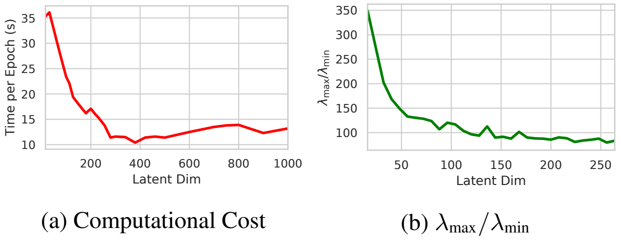
[1] Code is available at `https://github.com/miskcoo/kscore`.



(a) $M = 128$ (b) $M = 512$

(c) $d = 16$ (d) $d = 128$

Figure 1. Normalized distance $\mathbb{E}[\|s_p - \hat{s}_{p,\lambda}\|_2^2]/d$ on grid data. In the first row, M is fixed and d varies. In the second row, d is fixed and M varies. Shaded areas are three times the standard deviation.



(a) Computational Cost (b) $\lambda_{\max}/\lambda_{\min}$

Figure 2. (a) Computational costs of KEF-CG for $\lambda = 10^{-5}$ on MNIST; (b) The ratio of the maximum and the minimum eigenvalues of kernel matrices.

Parametric estimators: In the WAE experiment, we also consider the sliced score matching (SSM) estimator (Song et al., 2019), which is a parametric method and requires amortized training.

Proposed: The iterative curl-free estimator with the ν-method, and the conjugate gradient version of the KEF estimator (KEF-CG), both described in Sec. 3.5.

5.1. Synthetic Distributions

We follow Sutherland et al. (2018, Sec. 5.1) to construct a d-dimensional grid distribution. It is the mixture of d standard Gaussian distributions centered at d fixed vertices in the unit hypercube. We change d and M respectively to test the accuracy and the convergence of score estimators, and use 1024 samples from the grid distribution to evaluate the ℓ^2. We report the result of 32 runs in Fig. 1.

We can see that the effect of hypothesis space is significant. The diagonal kernels used in SSGE and Stein degrade the accuracy in high dimensions, while curl-free kernels provide better performance. In low dimensions, all estimators are comparable, and the computational cost of diagonal kernels is lower than that of curl-free kernels. This suggests favoring the diagonal kernels in low dimensions. Possibly because this dataset does not make the convergence rate saturate, we find different regularization schemes produce similar results. The iterative score estimator based on the ν-method is among the best and KEF-CG closely tracked them even with large d and M.

Table 2. Negative log-likelihoods on MNIST datasets and per epoch time on 128 latent dimension. All models are timed on GeForce GTX TITAN X GPU.

LATENT DIM	8	32	64	128	TIME
STEIN	97.15	92.10	101.60	114.41	4.2s
SSGE	97.24	92.24	101.92	114.57	9.2s
KEF	97.07	90.93	91.58	92.40	201.1s
NKEF$_2$	97.71	92.29	92.82	94.14	36.4s
NKEF$_4$	97.59	91.19	91.80	92.94	97.5s
NKEF$_8$	97.23	90.86	92.39	92.49	301.2s
KEF-CG	97.39	90.77	92.66	**92.05**	13.7s
ν-METHOD	97.28	90.94	**91.48**	92.10	78.1s
SSM	**96.98**	**89.06**	93.06	96.92	6.0s

5.2. Wasserstein Autoencoders

Wasserstein autoencoder (WAE) (Tolstikhin et al., 2017) is a latent variable model $p(\mathbf{z}, \mathbf{x})$ with observed and latent variables $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$, respectively. $p(\mathbf{z}, \mathbf{x})$ is defined by a prior $p(\mathbf{z})$ and a distribution of $\mathbf{z}$ conditioned on $\mathbf{x}$, and can be written as $p(\mathbf{z}, \mathbf{x}) = p(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$. WAEs aim at minimizing Wasserstein distance $\mathcal{W}_c(p_X, p_G)$ between the data distribution $p_X(\mathbf{x})$ and $p_G(\mathbf{x}) := \int p(\mathbf{z}, \mathbf{x})d\mathbf{z}$, where c is a metric on $\mathcal{X}$. Tolstikhin et al. (2017) showed that when $p_\theta(\mathbf{x}|\mathbf{z})$ maps $\mathbf{z}$ to $\mathbf{x}$ deterministically by a function $G : \mathcal{Z} \to \mathcal{X}$, it suffices to minimize $\mathbb{E}_{p_X(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\|\mathbf{x} - G(\mathbf{z})\|_2^2] + \lambda\mathcal{D}(q_\phi(\mathbf{z}), p(\mathbf{z}))$, where $\mathcal{D}$ is a divergence of two distributions and $q_\phi(\mathbf{z}|\mathbf{x})$ is a parametric approximation of the posterior. When we choose $\mathcal{D}$ to be the KL divergence, the entropy term of $q_\phi(\mathbf{z}) := \int q_\phi(\mathbf{z}|\mathbf{x})p_X(\mathbf{x})d\mathbf{x}$ in the loss function is intractable (Song et al., 2019). If $\mathbf{z}$ can be parameterized by $f_\phi(\mathbf{x})$ with $\mathbf{x} \sim p_X$, the gradient of the entropy can be estimated using score estimators as $\mathbb{E}_{p_X(\mathbf{x})}[\nabla_\mathbf{z} \log q_\phi(\mathbf{z})\nabla_\phi f_\phi(\mathbf{x})]$ (Shi et al., 2018; Song et al., 2019).

We train WAEs on MNIST and CelebA and repeat each configuration 3 times. The average negative log-likelihoods for MNIST estimated by AIS (Neal, 2001) are reported in Table 2. The results for CelebA are reported in appendix A. We can see that the performance of these estimators is close in low latent dimensions, and the parametric method is slightly better than nonparametric ones as we have continuously generated samples. However, in high dimensions, estimators based on curl-free kernels significantly outperform those based on diagonal kernels and parametric methods. This is probably due to guarantee that the estimates at all locations form a gradient field.

As discussed in Sec. 3.5, curl-free kernels are computationally expensive. This is shown in Table 2 by the running time of the original KEF algorithm. By comparing the time and performance of NKEF$_\alpha$ with $\alpha = 2, 4, 8$, we see that in order to get meaningful speed-up in high dimensions, low-rank approximation methods have to sacrifice the performance, which are outperformed by the iterative curl-free estimators based on the ν-method. KEF-CG is the fastest curl-free method in high dimensions while the performance is comparable with the original KEF. Fig. 2a shows the training time of KEF-CG in different latent dimensions. Surprisingly, the speed rapidly increases with increasing latent dimension and then flattens out. The convergence rate of conjugate gradient is determined by the condition number, which means the kernel matrix **K** becomes well-conditioned in high dimensions (see Fig. 2b).

We found with large d, SSGE required at least 97% eigenvalues to attain the reported likelihood. We also ran SSGE with curl-free kernels and found only 13% eigenvalues are required to attain a comparable result when $d = 8$. From these observations, a possible reason why diagonal kernels degrade the performance in high dimensions is that the distribution is complicated while the hypothesis set is simple, so the small number of eigenfunctions are insufficient to approximate the target. This can also be observed from Fig. 1, where the performance of diagonal kernels and curl-free kernels are closer as M increases since more eigenfunctions are provided.

6. Conclusion

Our contributions are two folds. Theoretically, we present a unifying view of nonparametric score estimators, and clarify the relationships of existing estimators. Under this perspective, we provide a unified convergence analysis of existing estimators, which improves existing error bounds. Practically, we propose an iterative curl-free estimator with nice theoretical properties and computational benefits, and develop a fast conjugate gradient solver for the KEF estimator.

Acknowledgements

We thank Ziyu Wang for reading an earlier version of this manuscript and providing valuable feedback. This work was supported by the National Key Research and Development Program of China (No. 2017YFA0700904), NSFC Projects (Nos. 61620106010, U19B2034, U1811461), Beijing NSF Project (No. L172037), Beijing Academy of Artificial Intelligence (BAAI), Tsinghua-Huawei Joint Research Program, Tsinghua Institute for Guo Qiang, and the NVIDIA NVAIL Program with GPU/DGX Acceleration. JS was also supported by a Microsoft Research Asia Fellowship.

References

Alvarez, M. A., Rosasco, L., Lawrence, N. D., et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Baker, C. T. The numerical treatment of integral equations. 1977.

Baldassarre, L., Rosasco, L., Barla, A., and Verri, A. Multi-output learning via spectral filtering. *Machine learning*, 87(3):259–301, 2012.

Bauer, F., Pereverzev, S., and Rosasco, L. On regularization algorithms in learning theory. *Journal of complexity*, 23 (1):52–72, 2007.

Canu, S. and Smola, A. Kernel methods and the exponential family. *Neurocomputing*, 69(7-9):714–720, 2006.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *International Conference on Machine Learning*, pp. 2606–2615, 2016.

De Vito, E., Rosasco, L., and Toigo, A. Learning sets with separating kernels. *Applied and Computational Harmonic Analysis*, 37(2):185–217, 2014.

Engl, H. W., Hanke, M., and Neubauer, A. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Fukumizu, K. Exponential manifold by reproducing kernel hilbert spaces. *Algebraic and Geometric mothods in statistics*, pp. 291–306, 2009.

Fuselier Jr, E. J. *Refined error estimates for matrix-valued radial basis functions*. PhD thesis, Texas A&M University, 2007.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Gorham, J. and Mackey, L. Measuring sample quality with stein's method. In *Advances in Neural Information Processing Systems*, pp. 226–234, 2015.

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

Li, Y. and Turner, R. E. Gradient estimators for implicit models. In *International Conference on Learning Representations*, 2018.

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

Macêdo, I. and Castro, R. *Learning divergence-free and curl-free vector fields with matrix-valued kernels*. IMPA, 2010.

Neal, R. M. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Rosasco, L., Belkin, M., and Vito, E. D. On learning with integral operators. *Journal of Machine Learning Research*, 11(Feb):905–934, 2010.

Saremi, S. and Hyvarinen, A. Neural empirical bayes. *Journal of Machine Learning Research*, 20:1–23, 2019.

Sasaki, H., Hyvärinen, A., and Sugiyama, M. Clustering via mode seeking by direct estimation of the gradient of a log-density. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 19–34. Springer, 2014.

Shi, J., Sun, S., and Zhu, J. A spectral approach to gradient estimation for implicit distributions. In *International Conference on Machine Learning*, pp. 4651–4660, 2018.

Smale, S. and Zhou, D.-X. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pp. 11895–11907, 2019.

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced score matching: A scalable approach to density and score estimation. *arXiv preprint arXiv:1905.07088*, 2019.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. Density estimation in infinite dimensional exponential families. *The Journal of Machine Learning Research*, 18(1):1830–1888, 2017.

Stein, C. M. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pp. 1135–1151, 1981.

Strathmann, H., Sejdinovic, D., Livingstone, S., Szabo, Z., and Gretton, A. Gradient-free hamiltonian monte carlo with efficient kernel exponential families. In *Advances in Neural Information Processing Systems*, pp. 955–963, 2015.

Sun, S., Zhang, G., Shi, J., and Grosse, R. Functional variational Bayesian Neural Networks. In *International Conference on Learning Representations*, 2019.

Sutherland, D., Strathmann, H., Arbel, M., and Gretton, A. Efficient and principled score estimation with nyström kernel exponential families. In *International Conference on Artificial Intelligence and Statistics*, pp. 652–660, 2018.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Van Loan, C. F. and Golub, G. H. *Matrix computations*. Johns Hopkins University Press, 1983.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Vito, E. D., Rosasco, L., Caponnetto, A., Giovannini, U. D., and Odone, F. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6(May): 883–904, 2005.

Warde-Farley, D. and Bengio, Y. Improving generative adversarial networks with denoising feature matching. *International Conference on Learning Representations*, 2016.

Wen, L., Zhou, Y., He, L., Zhou, M., and Xu, Z. Mutual information gradient estimation for representation learning. In *International Conference on Learning Representations*, 2020.

Williams, C. K. and Seeger, M. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pp. 682–688, 2001.