Kernel Exponential Family Estimation via Doubly Dual Embedding

*Bo Dai[1] *Hanjun Dai[2] Arthur Gretton[3] Le Song[2] Dale Schuurmans[1] Niao He[4]
[1]Google Brain [2]Georgia Tech [3]UCL [4]UIUC

Abstract

We investigate penalized maximum log-likelihood estimation for exponential family distributions whose natural parameter resides in a reproducing kernel Hilbert space. Key to our approach is a novel technique, *doubly dual embedding*, that avoids computation of the partition function. This technique also allows the development of a flexible sampling strategy that amortizes the cost of Monte-Carlo sampling in the inference stage. The resulting estimator can be easily generalized to kernel conditional exponential families. We establish a connection between kernel exponential family estimation and MMD-GANs, revealing a new perspective for understanding GANs. Compared to the score matching based estimators, the proposed method improves both memory and time efficiency while enjoying stronger statistical properties, such as fully capturing smoothness in its statistical convergence rate while the score matching estimator appears to saturate. Finally, we show that the proposed estimator empirically outperforms state-of-the-art methods in both kernel exponential family estimation and its conditional extension.

1 Introduction

The exponential family is one of the most important classes of distributions in statistics and machine learning. The exponential family possesses a number of useful properties (Brown, 1986)and includes many commonly used distributions with finite-dimensional natural parameters, such as the Gaussian, Poisson and multinomial distributions, to name a few. It is natural to consider generalizing the richness and flexibility of the exponential family to an *infinite*-dimensional parameterization via reproducing kernel Hilbert spaces (RKHS) (Canu and Smola, 2006).

Maximum log-likelihood estimation (MLE) has already been well-studied in the case of finite-dimensional exponential families, where desirable statistical properties such as asymptotic unbiasedness, consistency and asymptotic normality have been established. However, it is difficult to extend MLE to the infinite-dimensional case. Beyond the intractability of evaluating the partition function for a general exponential family, the necessary conditions for maximizing log-likelihood might not be feasible; that is, there may exist no solution to the KKT conditions in the infinite-dimensional case (Pistone and Rogantin, 1999; Fukumizu, 2009). To address this issue, Barron and Sheu (1991); Gu and Qiu (1993); Fukumizu (2009) considered several ways to regularize the function space by constructing (a series of) finite-dimensional spaces that approximate the original RKHS, yielding a tractable estimator in the restricted finite-dimensional space. However, as Sriperumbudur et al. (2017) note, even with the finite-dimension approximation, these algorithms are still expensive as every update requires Monte-Carlo sampling from a current model to compute the partition function.

An alternative score matching based estimator has recently been introduced by Sriperumbudur et al. (2017). This approach replaces the Kullback-Leibler (KL) with the Fisher divergence, defined by the expected squared distance between the score of the model (*i.e.*, the derivative of the log-density) and the target distribution score (Hyvärinen, 2005). By minimizing the Tikhonov-regularized Fisher divergence, Sriperumbudur et al. (2017) develop a computable estimator for the infinite-dimensional exponential family that also obtains a consistency guarantee. Recently, this method has been generalized to *conditional* infinite-dimensional exponential family estimation (Arbel and Gretton, 2017). Although score matching avoids computing the generally intractable integral, it requires computing and saving the first- and second-order derivatives of the reproducing kernel for *each*

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s). * indicates equal contribution.

dimension on *each* sample. For n samples with d features, this results in $\mathcal{O}\left(n^2d^2\right)$ memory and $\mathcal{O}\left(n^3d^3\right)$ time cost respectively, which becomes prohibitive for datasets with moderate sizes of n and d. To alleviate this cost, Sutherland et al. (2017) utilize the m-rank Nystöm approximation to the kernel matrix, reducing memory and time complexity to $\mathcal{O}\left(nmd^2\right)$ and $\mathcal{O}\left(m^3d^3\right)$ respectively. Although this reduces the cost dependence on sample size, the dependence on d is unaffected, hence the estimator remains unsuitable for high-dimensional data. Estimating a general exponential family model by either MLE or score matching also generally requires some form of Monte-Carlo sampling, such as MCMC or HMC, to perform inference, which significantly adds to the computation cost.

In summary, both the MLE and score matching based estimators for the infinite-dimensional exponential family incur significant computational overhead, particularly in high-dimensional applications.

In this paper, we revisit penalized MLE for the kernel exponential family and propose a new estimation strategy. Instead of solving the log-likelihood equation directly, as in existing MLE methods, we exploit a *doubly dual embedding* technique that leads to a novel saddle-point reformulation for the MLE (along with its conditional distribution generalization) in Section 3. We then propose a stochastic algorithm for the new view of penalized MLE in Section 4. Since the proposed estimator is based on penalized MLE, it does not require the first- and second-order derivatives of the kernel, as in score matching, greatly reducing the memory and time cost. Moreover, since the saddle point reformulation of penalized MLE avoids the intractable integral, the need for a Monte-Carlo step is also bypassed, thus accelerating the learning procedure. This approach also learns a flexibly parameterized sampler simultaneously, therefore, further reducing inference cost. We present the consistency and rate of the new estimation strategy in the well-specified case, *i.e.*, the true density belongs to the kernel exponential family, and an algorithm convergence guarantee in Section 5. We demonstrate the empirical advantages of the proposed algorithm in Section 6, comparing to state-of-the-art estimators for both the kernel exponential family and its conditional extension.

2 Preliminaries

We first provide a preliminary introduction to the exponential family and Fenchel duality, which will play vital roles in the derivation of the new estimator.

2.1 Exponential family

The natural form of the exponential family over Ω with the sufficient statistics $f \in \mathcal{F}$ is defined as

$$p_f(x) = p_0(x) \exp\left(\lambda f(x) - A\left(\lambda f\right)\right), \qquad (1)$$

where $A\left(\lambda f\right) := \log \int_\Omega \exp\left(\lambda f\left(x\right)\right) p_0\left(x\right) dx$, $x \in \Omega \subset \mathbb{R}^d$, $\lambda \in \mathbb{R}$, and $\mathcal{F} := \left\{f \in \mathcal{H} : \exp\left(A\left(\lambda f\right)\right) < \infty\right\}$. In this paper, we mainly focus on the case where $\mathcal{H}$ is a reproducing Hilbert kernel space (RKHS) with kernel $k\left(x, x'\right)$, such that $f\left(x\right) = \left\langle f, k\left(x, \cdot\right)\right\rangle$. However, we emphasize that the proposed algorithm in Section 4 can be easily applied to arbitrary differentiable function parametrizations, such as deep neural networks.

Given samples $\mathcal{D} = \left[x_i\right]_{i=1}^N$, a model with a finite-dimensional parameterization can be learned via maximum log-likelihood estimation (MLE),

$$\max_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N \log p_f(x_i) = \widehat{\mathbb{E}}_{\mathcal{D}}\left[\lambda f(x) + \log p_0\left(x\right)\right] - A(\lambda f),$$

where $\widehat{\mathbb{E}}_{\mathcal{D}}\left[\cdot\right]$ denotes the empirical expectation over $\mathcal{D}$. The MLE (2) is well-studied and has nice properties. However, the MLE is known to be "ill-posed" in the infinite-dimensional case, since the optimal solution might not be exactly achievable in the representable space. Therefore, penalized MLE has been introduced for the finite (Dudík et al., 2007) and infinite dimensional Gu and Qiu (1993); Altun and Smola (2006) exponential families respectively. Such regularization essentially relaxes the moment matching constraints in terms of some norm, as shown later, guaranteeing the existence of a solution. In this paper, we will also focus on MLE with RKHS norm regularization.

One useful theoretical property of MLE for the exponential family is convexity w.r.t. f. With convex regularization, *e.g.*, $\|f\|_{\mathcal{H}}^2$, one can use stochastic gradient descent to recover a unique global optimum. Let $L\left(f\right)$ denote RKHS norm penalized log-likelihood, *i.e.*,

$$L\left(f\right) := \frac{1}{N} \sum_{i=1}^N \log p_f(x_i) - \frac{\eta}{2} \left\|f\right\|_{\mathcal{H}}^2. \qquad (2)$$

where $\eta > 0$ denotes the regularization parameter. The gradient of $L\left(f\right)$ w.r.t. f can be computed as

$$\nabla_f L = \widehat{\mathbb{E}}_{\mathcal{D}}\left[\lambda \nabla_f f(x)\right] - \nabla_f A(\lambda f) - \eta f. \qquad (3)$$

To calculate the $\nabla_f A(\lambda f)$ in (3), we denote $Z\left(\lambda f\right) = \int_\Omega \exp\left(\lambda f\left(x\right)\right) p_0\left(x\right) dx$ and expand the partition function by definition,

$$\begin{aligned} \nabla_f A(\lambda f) &= \frac{1}{Z(\lambda f)} \nabla_f \int_\Omega \exp\left(\lambda f(x)\right) p_0\left(x\right) dx \\ &= \int_\Omega \frac{p_0\left(x\right) \exp\left(\lambda f(x)\right)}{Z(\lambda f)} \nabla_f \lambda f(x) dx \\ &= \mathbb{E}_{p_f(x)}\left[\nabla_f \lambda f(x)\right]. \end{aligned} \qquad (4)$$

One can approximate the $\mathbb{E}_{p_f(x)}\left[\nabla_f \lambda f(x)\right]$ by AIS/MCMC samples (Vembu et al., 2009), which leads to the Contrastive Divergence (CD) algorithm (Hinton, 2002).

To avoid costly MCMC sampling in estimating the gradient, Sriperumbudur et al. (2017) construct an esti-

mator based on score matching instead of MLE, which minimizes the penalized Fisher divergence. Plugging the kernel exponential family into the empirical Fisher divergence, the optimization reduces to

$$J(f) := \frac{\lambda^2}{2}\langle f, \widehat{C} f\rangle_{\mathcal{H}} + \lambda\langle f, \hat{\delta}\rangle_{\mathcal{H}} + \frac{\eta}{2}\|f\|_{\mathcal{H}}^2, \quad (5)$$

where

$$\widehat{C} := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\partial_j k(x_i,\cdot)\otimes\partial_j k(x_i,\cdot),$$
$$\hat{\delta} := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\partial_j k(x_i,\cdot)\left(\partial_j \log p_0(x_i) + \partial_j^2 k(x_i,\cdot)\right).$$

As we can see, the score matching objective (5) is convex and does not involve the intractable integral in $A(\lambda f)$. However, such an estimator requires the computation of first- and second-order of derivatives of kernel for each dimension on each data, leading to memory and time cost of $\mathcal{O}\left(n^2d^2\right)$ and $\mathcal{O}\left(n^3d^3\right)$ respectively. This quickly becomes prohibitive for even modest n and d.

The same difficulty also appears in the score matching based estimator in (Arbel and Gretton, 2017) for the *conditional exponential family*, which is defined as

$$p(y|x) = p_0(y)\exp\left(\lambda f(x,y) - A_x(\lambda f)\right),\ f\in\mathcal{F} \quad (6)$$

where $y\in\Omega_y\subset\mathbb{R}^p$, $x\in\Omega_x\subset\mathbb{R}^d$, $\lambda\in\mathbb{R}$, and $A_x(\lambda f) := \log\int_{\Omega_y} p_0(y)\exp\left(\lambda f(x,y)\right)dy$, $\mathcal{F}$ is defined as $\{f\in\mathcal{H}_y : \exp\left(A_x(\lambda f)\right) < \infty\}$. We consider $f:\Omega_x\times\Omega_y\to\mathbb{R}$ such that $f(x,\cdot)$ is in RKHS $\mathcal{H}_y$ for $\forall x\in\Omega_x$. Denoting $\mathcal{T}\in\mathcal{H}_{\Omega_x}:\Omega_x\to\mathcal{H}_y$ such that $\mathcal{T}_x(y) = f(x,y)$, we can derive its kernel function following Micchelli and Pontil (2005); Arbel and Gretton (2017). By the Riesz representation theorem, $\forall x\in\Omega_x$ and $h\in\mathcal{H}_y$, there exists a linear operator $\Gamma_x:\mathcal{H}_y\to\mathcal{H}_{\Omega_x}$ such that

$$\langle h, \mathcal{T}_x\rangle_{\mathcal{H}_y} = \langle\mathcal{T}, \Gamma_x h\rangle_{\mathcal{H}_{\Omega_x}},\ \forall\mathcal{T}\in\mathcal{H}_{\Omega_x}.$$

Then, the kernel can be defined by composing Γ_x with its dual, *i.e.*, $k(x,x') = \Gamma_x^*\Gamma_{x'}$ and the function $f(x,y) = \mathcal{T}_x(y) = \langle\mathcal{T}, \Gamma_x k(y,\cdot)\rangle$. We follow such assumptions for kernel conditional exponential family.

2.2 Convex conjugate and Fenchel duality

Denote $h(\cdot)$ as a function $\mathbb{R}^d\to\mathbb{R}$, then its convex conjugate function is defined as

$$h^*(u) = \sup_{v\in\mathbb{R}^d}\{u^\top v - h(v)\}.$$

If $h(\cdot)$ is proper, convex and lower semicontinuous, the conjugate function, $h^*(\cdot)$, is also proper, convex and lower semicontinuous. Moreover, h and h^* are dual to each other, *i.e.*, $(h^*)^* = h$. Such a relationship is known as Fenchel duality (Rockafellar, 1970; Hiriart-Urruty and Lemaréchal, 2012). By the conjugate function, we can represent the h by as,

$$h(v) = \sup_{u\in\mathbb{R}^d}\{v^\top u - h^*(u)\}.$$

The supremum achieves if $v\in\partial h^*(u)$, or equivalently $u\in\partial h(v)$.

3 A Saddle-Point Formulation of Penalized MLE

As discussed in Section 2, the penalized MLE for the exponential family involves computing the log-partition functions, $A(\lambda f)$ and $A_x(\lambda f)$, which are intractable in general. In this section, we first introduce a saddle-point reformulation of the penalized MLE of the exponential family, using Fenchel duality to bypass the computation of the intractable log-partition function. This approach can also be generalized to the conditional exponential family. First, observe that we can rewrite the log-partition function $A(\lambda f)$ via Fenchel duality as follows.

Theorem 1 (Fenchel dual of log-partition)

$$A(\lambda f) = \max_{q\in\mathcal{P}}\ \lambda\langle q(x), f(x)\rangle_2 - KL(q||p_0), \quad (7)$$
$$p_f(x) = \operatorname*{argmax}_{q\in\mathcal{P}}\ \lambda\langle q(x), f(x)\rangle_2 - KL(q||p_0), \quad (8)$$

where $\langle f, g\rangle_2 := \int_\Omega f(x)g(x)dx$, $\mathcal{P}$ *denotes the space of distributions and* $KL(q||p_0) := \int_\Omega q(x)\log\frac{q(x)}{p_0(x)}dx$.

Proof Denote $l(q) := \lambda\langle q(x), f(x)\rangle - KL(q||p_0)$, which is strongly concave w.r.t. $q\in\mathcal{P}$, the optimal q^* can be obtained by setting the

$$\log q^*(x) \propto \lambda f(x) + \log p_0(x).$$

Since $q^*\in\mathcal{P}$, we have

$$q^*(x) = p_0(x)\exp\left(\lambda f(x) - A(\lambda f)\right) = p_f(x),$$

which leads to (8). Plugging q^* to $l(q)$, we obtain the maximum as $\log\int_\Omega\exp\left(\lambda f(x)\right)p_0(x)dx$, which is exactly $A(\lambda f)$, leading to (7). ■

Therefore, invoking the Fenchel dual of $A(\lambda f)$ into the penalized MLE, we achieve a saddle-point optimization, *i.e.*,

$$\max_{f\in\mathcal{F}}\ L(f) \propto \quad (9)$$
$$\min_{q\in\mathcal{P}}\underbrace{\widehat{\mathbb{E}}_{\mathcal{D}}[f(x)] - \mathbb{E}_{q(x)}[f(x)] - \frac{\eta}{2}\|f\|_{\mathcal{H}}^2 + \frac{1}{\lambda}KL(q||p_0)}_{\ell(f,q)}.$$

The saddle-point reformulation of the penalized MLE in (9) resembles the optimization in MMD GAN (Li et al., 2017; Bińkowski et al., 2018). In fact, the dual problem of the penalized MLE of the exponential family is a *KL-regularized* MMD GAN with a special design of the kernel family. Alternatively, if f is a Wasserstein-1 function, the optimization resembles the Wasserstein GAN (Arjovsky et al., 2017).

Next, we consider the duality properties.

Theorem 2 (weak and strong duality) *The weak duality holds in general,* i.e.*,*

$$\max_{f\in\mathcal{F}}\min_{q\in\mathcal{P}}\ \ell(f,q) \leqslant \min_{q\in\mathcal{P}}\max_{f\in\mathcal{F}}\ \ell(f,q).$$

The strong duality holds when $\mathcal{F}$ *is a closed RKHS and*

$\mathcal{P}$ is the distributions with bounded L_2 norm, i.e.,

$$\max_{f\in\mathcal{F}}\min_{q\in\mathcal{P}}\ \ell(f,q)=\min_{q\in\mathcal{P}}\max_{f\in\mathcal{F}}\ \ell(f,q). \tag{10}$$

Theorem 2 can be obtained by directly applying the minimax theorem (Ekeland and Temam, 1999)[Proposition 2.1]. We refer to the max-min problem in (10) as the primal problem, while the min-max form as the its dual form.

Remark (connections to MMD GAN): Consider the dual problem with the kernel learning, *i.e.*,

$$\min_{q\in\mathcal{P}}\max_{f\in\mathcal{F}_\phi,\phi}\widehat{\mathbb{E}}_{\mathcal{D}}\left[f(x)\right]-\mathbb{E}_{q(x)}\left[f(x)\right]-\frac{\eta}{2}\left\|f\right\|_{\mathcal{H}}^2+\frac{1}{\lambda}KL\left(q||p_0\right). \tag{11}$$

where we involve the parameters of the kernel ϕ to be learned in the optimization. By setting the gradient $\nabla_f\ell(f,q)=0$, for fixed q, we obtain the optimal witness function in the RKHS, $f_q=\frac{1}{\eta}\left(\widehat{\mathbb{E}}\left[k_\phi(x,\cdot)\right]-\mathbb{E}_q\left[k_\phi(x,\cdot)\right]\right)$, which leads to

$$\min_{q\in\mathcal{P}}\max_{\phi}\underbrace{\widehat{\mathbb{E}}\left[k_\phi(x,x')\right]-2\widehat{\mathbb{E}}\mathbb{E}_q\left[k_\phi(x,x')\right]+\mathbb{E}_q\left[k_\phi(x,x')\right]}_{MMD_\phi(\mathcal{D},q)}+\frac{2\eta}{\lambda}KL\left(q||p_0\right). \tag{12}$$

This can be regarded as the KL-divergence regularized MMD GAN. Thus, with KL-divergence regularization, the MMD GAN learns an infinite-dimension exponential family in an adaptive RKHS. Such a novel perspective bridges GAN and exponential family estimation, which appears to be of independent interest and potentially brings a new connection to the GAN literature for further theoretical development.

Remark (connections to Maximum Entropy Moment Matching): Altun and Smola (2006); Dudík et al. (2007) discuss the maximum entropy moment matching method for distribution estimation,

$$\begin{aligned}\min_{q\in\mathcal{P}}\quad & KL\left(q||p_0\right)\\ \text{s.t.}\quad & \left\|\mathbb{E}_q\left[k\left(x,\cdot\right)\right]-\widehat{\mathbb{E}}\left[k\left(x,\cdot\right)\right]\right\|_{\mathcal{H}_k}\leqslant\frac{\eta'}{2},\end{aligned} \tag{13}$$

whose dual problem will be reduced to the penalized MLE (2) with proper choice of η' (Altun and Smola, 2006)[Lemma 6]. Interestingly, the proposed saddle-point formulation (9) shares the solution to (13). From the maximum entropy view, the penalty $\|f\|_{\mathcal{H}}$ relaxes the moment matching constraints. However, the algorithms provided in Altun and Smola (2006); Dudík et al. (2007) simply ignore the difficulty in computing the expectation in $A\left(\lambda f\right)$, which is not practical, especially when f is infinite-dimensional.

Similar to Theorem 1, we can also represent $A_x\left(\lambda f\right)$ by its Fenchel dual,

$$A_x\left(\lambda f\right)=\max_{q(\cdot|x)\in\mathcal{P}}\lambda\left\langle q(y|x),f\left(x,y\right)\right\rangle-KL\left(q||p_0\right), \tag{14}$$

$$p_f\left(y|x\right)=\operatorname*{argmax}_{q(\cdot|x)\in\mathcal{P}}\lambda\left\langle q(x),f(x)\right\rangle-KL\left(q||p_0\right). \tag{15}$$

Then, we can recover the penalized MLE for the conditional exponential family as

$$\max_{f\in\mathcal{F}}\min_{q(\cdot|x)\in\mathcal{P}}\ \widehat{\mathbb{E}}_{\mathcal{D}}\left[f(x,y)\right]-\mathbb{E}_{q(y|x)}\left[f(x,y)\right]-\frac{\eta}{2}\left\|f\right\|_{\mathcal{H}}^2+\frac{1}{\lambda}KL\left(q||p_0\right). \tag{16}$$

The saddle-point reformulations of penalized MLE in (9) and (16) bypass the difficulty in the partition function. Therefore, it is very natural to consider learning the exponential family by solving the saddle-point problem (9) with an appropriate parametrized dual distribution $q\left(x\right)$. This approach is referred to as the "dual embedding" technique in Dai et al. (2016), which requires:

- **i)** the parametrization family should be flexible enough to reduce the extra approximation error;
- **ii)** the parametrized representation should be able to provide density value.

As we will see in Section 5, the flexibility of the parametrization of the dual distribution will have a significant effect on the consistency of the estimator.

One can of course use the kernel density estimator (KDE) as the dual distribution parametrization, which preserves convex-concavity. However, KDE will easily fail when approximating high-dimensional data. Applying the reparametrization trick with a suitable class of probability distributions (Kingma and Welling, 2013; Rezende et al., 2014) is another alternative to parametrize the dual distribution. However, the class of such parametrized distributions is typically restricted to simple known distributions, which might not be able to approximate the true solution, potentially leading to a huge approximation bias. At the other end of the spectrum, the distribution family generated by transport mapping is sufficiently flexible to model smooth distributions (Goodfellow et al., 2014; Arjovsky et al., 2017). However, the density value of such distributions, *i.e.*, $q\left(x\right)$, is not available for KL-divergence computation, and thus, is not applicable for parameterizing the dual distribution. Recently, flow-based parametrized density functions (Rezende and Mohamed, 2015; Kingma et al., 2016; Dinh et al., 2016) have been proposed for a trade-off between the flexibility and tractability. However, the expressive power of existing flow-based models remains restrictive even in our synthetic example and cannot be directly applied to conditional models.

3.1 Doubly Dual Embedding for MLE

Transport mapping is very flexible for generating smooth distributions. However, a major difficulty is that it lacks the ability to obtain the density value $q\left(x\right)$, making the computation of the KL-divergence impossible. To retain the flexibility of transport mapping and avoid the calculation of $q\left(x\right)$, we introduce

Algorithm 1 Doubly Dual Embedding-SGD for saddle-point reformulation of MLE (19)

1: **for** $l = 1, \ldots, L$ **do**
2: Compute (f, ν) by Algorithm 2.
3: Sample $\{\xi_0 \sim p(\xi)\}_{b=1}^{B}$.
4: Generate $x_b = g_{w_g}(\xi)$ for $b = 1, \ldots, B$.
5: Compute stochastic approximation $\widehat{\nabla}_{w_g} \widehat{L}(w_g)$ by (23).
6: Update $w_g^{l+1} = w_g^l - \rho_l \widehat{\nabla}_{w_g} L(w_g)$.
7: **end for**
8: Output w_g and f.

doubly dual embedding, which achieves a delicate balance between flexibility and tractability.

First, noting that KL-divergence is also a convex function, we consider the Fenchel dual of KL-divergence (Nguyen et al., 2008), *i.e.*,

$$KL(q||p_0) = \max_{\nu} \mathbb{E}_q[\nu(x)] - \mathbb{E}_{p_0}[\exp(\nu(x))] + 1, \quad (17)$$

$$\log \frac{q(x)}{p_0(x)} = \operatorname*{argmax}_{\nu} \mathbb{E}_q[\nu(x)] - \mathbb{E}_{p_0}[\exp(\nu(x))] + 1, \quad (18)$$

with $\nu(\cdot) : \Omega \to \mathbb{R}$. One can see in the dual representation of the KL-divergence that the introduction of the auxiliary optimization variable ν eliminates the explicit appearance of $q(x)$ in (17), which makes the transport mapping parametrization for $q(x)$ applicable. Since there is no extra restriction on ν, we can use arbitrary smooth function approximation for $\nu(\cdot)$, such as kernel functions or neural networks.

Applying the dual representation of KL-divergence into the dual problem of the saddle-point reformulation (9), we obtain the ultimate saddle-point optimization for the estimation of the kernel exponential family,

$$\min_{q \in \mathcal{P}} \max_{f,\nu \in \mathcal{F}} \tilde{\ell}(f, \nu, q) := \widehat{\mathbb{E}}_{\mathcal{D}}[f] - \mathbb{E}_q[f] - \frac{\eta}{2}\|f\|_{\mathcal{H}}^2 + \frac{1}{\lambda}\left(\mathbb{E}_q[\nu] - \mathbb{E}_{p_0}[\exp(\nu)]\right). \quad (19)$$

Note that several other work also applied Fenchel duality to KL-divergence (Nguyen et al., 2008; Nowozin et al., 2016), but the approach taken here is different as it employs dual of $KL(q||p_0)$, rather than $KL(q||\mathcal{D})$.

Remark (Extension for kernel conditional exponential family): Similarly, we can apply the doubly dual embedding technique to the penalized MLE of the kernel conditional exponential family, which leads to

$$\min_{q \in \mathcal{P}_x} \max_{f,\nu \in \mathcal{F}} \widehat{\mathbb{E}}_{\mathcal{D}}[f] - \mathbb{E}_{q(y|x),x\sim\mathcal{D}}[f] - \frac{\eta}{2}\|f\|_{\mathcal{H}}^2 + \frac{1}{\lambda}\left(\mathbb{E}_{q(y|x),x\sim\mathcal{D}}[\nu] - \mathbb{E}_{p_0(y)}[\exp(\nu)]\right).$$

In summary, with the *doubly dual embedding* technique, we derive a saddle-point reformulation of penalized MLE that bypasses the difficulty of handling the intractable partition function while allowing great flexibility in parameterizing the dual distribution.

Algorithm 2 Stochastic Functional Gradients for f and ν

1: **for** $k = 1, \ldots, K$ **do**
2: Sample $\xi \sim p(\xi)$.
3: Generate $x = g(\xi)$.
4: Sample $x' \sim p_0(x)$.
5: Compute stochastic function gradient w.r.t. f and ν with (20) and (21).
6: Update f_k and ν_k with (22) and (22).
7: **end for**
8: Output (f_K, ν_K) .

4 Practical Algorithm

In this section, we introduce the transport mapping parametrization for the dual distribution, $q(x)$, and apply the stochastic gradient descent for solving the optimization problems in (19) and (20). For simplicity of exposition, we only illustrate the algorithm for (19). The algorithm can be easily applied to (20).

Denote the parameters in the transport mapping for $q(x)$ as w_g, such that $\xi \sim p(\xi)$ and $x = g_{w_g}(\xi)$. We illustrate the algorithm with a kernel parametrize $\nu(\cdot)$. There are many alternative choices of parametrization of ν, *e.g.*, neural networks—the proposed algorithm is still applicable to the parametrization as long as it is differentiable. We abuse notation somewhat by using $\tilde{\ell}(f, \nu, w_g)$ as $\tilde{\ell}(f, \nu, q)$ in (19). with such parametrization, the (19) is an upper bound of the penalty MLE in general by Theorem 2.

With the kernel parametrized (f, ν), the inner maximization over f and ν is a standard concave optimization. We can solve it using existing algorithms to achieve the global optimal solution. Due to the existence of the expectation, we will use the stochastic functional gradient descent for scalability. Given $(f, \nu) \in \mathcal{H}$, following the definition of functional gradients (Kivinen et al., 2004; Dai et al., 2014), we have

$$\begin{aligned}\zeta_f(\cdot) &:= \nabla_f \tilde{\ell}(f, \nu, w_g) \quad (20)\\ &= \widehat{\mathbb{E}}_{\mathcal{D}}[k(x, \cdot)] - \mathbb{E}_{\xi}\left[k\left(g_{w_g}(\xi_0), \cdot\right)\right] - \eta f(\cdot).\end{aligned}$$

$$\begin{aligned}\zeta_\nu(\cdot) &:= \nabla_\nu \tilde{\ell}(f, \nu, w_g) \quad (21)\\ &= \frac{1}{\lambda}\left(\mathbb{E}_{\xi}\left[k\left(g_{w_g}(\xi), \cdot\right)\right] - \mathbb{E}_{p_0}[\exp(\nu(x))\, k(x, \cdot)]\right).\end{aligned}$$

In the k-th iteration, given sample $x \sim \mathcal{D}, \xi \sim p(\xi)$, and $x' \sim p_0(x)$, the update rule for f and ν will be

$$\begin{aligned}f_{k+1}(\cdot) &= (1 - \eta\tau_k) f_k(\cdot) + \tau_k\left(k(x, \cdot) - k\left(g_{w_g}(\xi), \cdot\right)\right),\\ \nu_{k+1}(\cdot) &= \nu_k(\cdot) \quad (22)\\ &\quad + \frac{\tau_k}{\lambda}\left(k\left(g_{w_g}(\xi), \cdot\right) - \exp(\nu_k(x'))\, k(x', \cdot)\right),\end{aligned}$$

where τ_k denotes the step-size.

Then, we consider the update rule for the parameters in dual transport mapping embedding.

Theorem 3 (Dual gradient) *Denoting*

$\left(f^*_{w_g}, \nu^*_{w_g}\right) = \operatorname{argmax}_{(f,\nu)\in\mathcal{H}} \tilde{\ell}\left(f, \nu, w_g\right)$ *and* $\widehat{L}\left(w_g\right) = \tilde{\ell}\left(f^*_{w_g}, \nu^*_{w_g}, w_g\right)$, *we have*

$$\nabla_{w_g}\widehat{L}\left(w_g\right) = -\mathbb{E}_\xi\left[\nabla_{w_g} f^*_{w_g}\left(g_{w_g}\left(\xi\right)\right)\right] + \frac{1}{\lambda}\mathbb{E}_\xi\left[\nabla_{w_g}\nu^*_{w_g}\left(g_{w_g}\left(\xi\right)\right)\right]. \quad (23)$$

Proof details are given in Appendix A.1. With this gradient estimator, we can apply stochastic gradient descent to update w_g iteratively. We summarize the updates for (f, ν) and w_g in Algorithm 1 and 2.

Compared to score matching based estimators (Sriperumbudur et al., 2017; Sutherland et al., 2017), although the convexity no longer holds for the saddle-point estimator, the doubly dual embedding estimator avoids representing f with derivatives of the kernel, thus avoiding the memory cost dependence on the square of dimension. In particular, the proposed estimator for f based on (19) reduces the memory cost from $\mathcal{O}\left(n^2d^2\right)$ to $\mathcal{O}\left(K^2\right)$ where K denotes the number of iterations in Algorithm 2. In terms of time cost, we exploit the stochastic update, which is naturally suitable for large-scale datasets and avoids the matrix inverse computation in score matching estimator whose cost is $\mathcal{O}\left(n^3d^3\right)$. We also learn the dual distribution simultaneously, which can be easily used to generate samples from the exponential family for inference, thus saving the cost of Monte-Carlo sampling in the inference stage. For detailed discussion about the computation cost, please refer to Appendix D.

Remark (random feature extension): Memory cost is a well-known bottleneck for applying kernel methods to large-scale problems. When we set $K = n$, the memory cost will be $\mathcal{O}\left(n^2\right)$, which is prohibitive for millions of data points. Random feature approximation (Rahimi and Recht, 2008; Dai et al., 2014; Bach, 2015) can be utilized for scaling up kernel methods. The proposed Algorithm 1 and Algorithm 2 are also compatible with random feature approximation, and hence, applicable to large-scale problems in the same way. With r random features, we can further reduce the memory cost of storing f to $\mathcal{O}\left(rd\right)$. However, even with a random feature approximation, the score matching based estimator will still require $\mathcal{O}\left(rd^2\right)$ memory. One can also learn random features by back-propagation, which leads to the neural networks extension. Due to space limitations, details of this variant of Algorithm 1 are given in Appendix B.

5 Theoretical Analysis

In this section, we will first provide the analysis of consistency and the sample complexity of the proposed estimator based on the saddle-point reformulation of the penalized MLE in the well-specified case, where the true density is assumed to be in the kernel (conditional) exponential family, following Sutherland et al. (2017); Arbel and Gretton (2017). Then, we consider the convergence property of the proposed algorithm. We mainly focus on the analysis for $p\left(x\right)$. The results can be easily extended to kernel conditional exponential family $p\left(y|x\right)$.

5.1 Statistical Consistency

We explicitly consider approximation error from the dual embedding in the consistency of the proposed estimator. We first establish some notation that will be used in the analysis. For simplicity, we set $\lambda = 1$ and $p_0\left(x\right) = 1$ improperly in the exponential family expression (1). We denote $p^*\left(x\right) = \exp\left(f^*\left(x\right) - A\left(f^*\right)\right)$ as the ground-true distribution with its true potential function f^* and $\left(\tilde{f}, \tilde{q}, \tilde{v}\right)$ as the optimal primal and dual solution to the saddle point reformulation of the penalized MLE (9). The parametrized dual space is denoted as $\mathcal{P}_w$. We denote $p_{\tilde{f}} := \exp\left(\lambda\tilde{f} - A\left(\lambda\tilde{f}\right)\right)$ as the exponential family generated by $\tilde{f}$. We have the consistency results as

Theorem 4 *Assume the spectrum of kernel $k\left(\cdot,\cdot\right)$ decays sufficiently homogeneously in rate l^{-r}. With some other mild assumptions listed in Appendix A.2, we have as $\eta \to 0$ and $n\eta^{\frac{1}{r}} \to \infty$,*

$$KL\left(p^*||p_{\tilde{f}}\right) + KL\left(p_{\tilde{f}}||p^*\right) = \mathcal{O}_{p^*}\left(n^{-1}\eta^{-\frac{1}{r}} + \eta + \epsilon^2_{approx}\right),$$

where $\epsilon_{approx} := \sup_{f\in\mathcal{F}}\inf_{q\in\mathcal{P}_w}\left\|p_f - q\right\|_{p^*}$. Therefore, when setting $\eta = \mathcal{O}\left(n^{-\frac{r}{1+r}}\right)$, $p_{\tilde{f}}$ converges to p^* in terms of Jensen-Shannon divergence at rate $\mathcal{O}_{p^*}\left(n^{-\frac{r}{1+r}} + \epsilon^2_{approx}\right)$.

For the details of the assumptions and the proof, please refer to Appendix A.2. Recall the connection between the proposed model and MMD GAN as discussed in Section 3. Theorem 4 also provides a learnability guarantee for a class of GAN models as a byproduct. The most significant difference of the bound provided above, compared to Gu and Qiu (1993); Altun and Smola (2006), is the explicit consideration of the bias from the dual parametrization. Moreover, instead of the Rademacher complexity used in the sample complexity results of Altun and Smola (2006), our result exploits the spectral decay of the kernel, which is more directly connected to properties of the RKHS.

From Theorem 4 we can clearly see the effect of the parametrization of the dual distribution: if the parametric family of the dual distribution is simple, the optimization for the saddle-point problem may become easy, however, ϵ_{approx} will dominate the error. The other extreme case is to also use the kernel exponential family to parametrize the dual distribution, then, ϵ_{approx} will reduce to 0, however, the optimization will be difficult to handle. The saddle-point reformulation

provides us the opportunity to balance the difficulty of the optimization with approximation error.

The statistical consistency rate of our estimator and the score matching estimator (Sriperumbudur et al., 2017) are derived under different assumptions, therefore, they are not directly comparable. However, since the smoothness is not fully captured by the score matching based estimator in Sriperumbudur et al. (2017), it only achieves $\mathcal{O}\left(n^{-\frac{2}{3}}\right)$ even if f is infinitely smooth. While under the case that dual distribution parametrization is relative flexible, *i.e.*, ϵ_{approx} is negligible, and the the spectrum of the kernel decay rate $r \to \infty$, the proposed estimator will converge in rate $\mathcal{O}\left(n^{-1}\right)$, which is significantly more efficient than the score matching method.

5.2 Algorithm Convergence

It is well-known that the stochastic gradient descent converges for saddle-point problem with convex-concave property (Nemirovski et al., 2009). However, for better the dual parametrization to reduce ϵ_{approx} in Theorem 4, we parameterize the dual distribution with the nonlinear transport mapping, which breaks the convexity. In fact, by Theorem 3, we obtain the unbiased gradient w.r.t. w_g. Therefore, the proposed Algorithm 1 can be understood as applying the stochastic gradient descent for the non-convex dual minimization problem, *i.e.*, $\min_{w_g} \widehat{L}(w_g) := \tilde{\ell}\left(f^*_{w_g}, \nu^*_{w_g}, w_g\right)$. From such a view, we can prove the sublinearly convergence rate to a stationary point when stepsize is diminishing following Ghadimi and Lan (2013); Dai et al. (2017). We list the result below for completeness.

Theorem 5 *Assume that the parametrized objective $\widehat{L}(w_g)$ is C-Lipschitz and variance of its stochastic gradient is bounded by σ^2. Let the algorithm run for L iterations with stepsize $\rho_l = \min\{\frac{1}{L}, \frac{D'}{\sigma\sqrt{L}}\}$ for some $D' > 0$ and output $w_g^1, \ldots, w_g^L$. Setting the candidate solution to be $\widehat{w}_g$ randomly chosen from $w_g^1, \ldots, w_g^L$ such that $P(w = w_g^j) = \frac{2\rho_j - C\rho_j^2}{\sum_{j=1}^L (2\rho_j - C\rho_j^2)}$, then it holds that $\mathbb{E}\left[\left\|\nabla \widehat{L}(\widehat{w}_g)\right\|^2\right] \leqslant \frac{CD^2}{L} + (D' + \frac{D}{D'})\frac{\sigma}{\sqrt{L}}$ where $D := \sqrt{2(\widehat{L}(w_g^1) - \min \widehat{L}(w_g))/L}$ represents the distance of the initial solution to the optimal solution.*

The above result implies that under the choice of the parametrization of f, ν and g, the proposed Algorithm 1 converges sublinearly to a stationary point, whose rate will depend on the smoothing parameter.

6 Experiments

In this section, we compare the proposed doubly dual embedding (DDE) with the current state-of-the-art score matching estimators for kernel exponential family (KEF) Sutherland et al. (2017)[1] and its conditional extension (KCEF) Arbel and Gretton (2017)[2], respectively, as well as several competitors. We test the proposed estimator empirically following their setting. We use Gaussian RBF kernel for both exponential family and its conditional extension, $k(x, x') = \exp\left(-\|x - x'\|_2^2/\sigma^2\right)$, with the bandwidth σ set by median-trick (Dai et al., 2014). For a fair comparison, we follow Sutherland et al. (2017) and Arbel and Gretton (2017) to set the $p_0(x)$ for kernel exponential family and its conditional extension, respectively. The dual variables are parametrized by MLP with 5 layers. More implementation details can be found in Appendix E and the code repository which is available at https://github.com/Hanjun-Dai/dde.

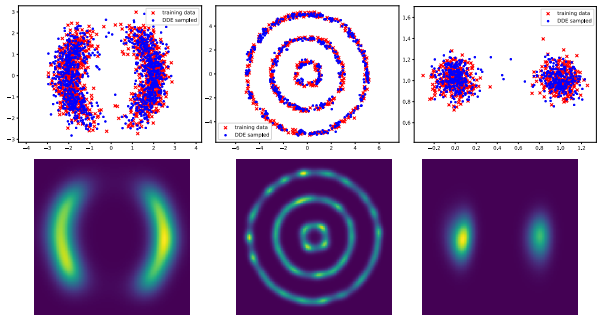
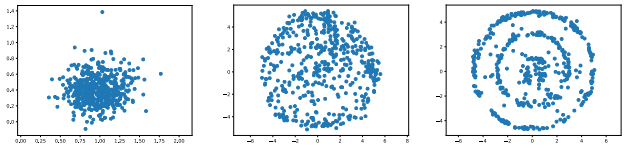


Figure 1: The blue points in the figures in first row show the generated samples by learned models, and the red points are the training samples. The learned f are illustrated in the second row.

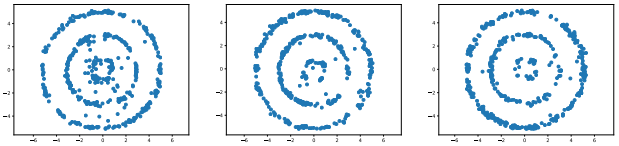
Density estimation We evaluate the DDE on the synthetic datasets, including `ring`, `grid` and `two moons`, where the first two are used in Sutherland et al. (2017), and the last one is from Rezende and Mohamed (2015). The `ring` dataset contains the points uniformly sampled along three circles with radii $(1, 3, 5) \in \mathbb{R}^2$ and $\mathcal{N}\left(0, 0.1^2\right)$ noise in the radial direction and extra dimensions. The d-dim `grid` dataset contains samples from mixture of d Gaussians. Each center lies on one dimension in the d-dimension hypercube. The `two moons` dataset is sampled from the exponential family with potential function as $\frac{1}{2}\left(\frac{\|x\|-2}{0.4}\right)^2 - \log\left(\exp\left(-\frac{1}{2}\left(\frac{x-2}{0.6}\right)^2\right) + \exp\left(-\frac{1}{2}\left(\frac{x+2}{0.6}\right)^2\right)\right)$. We use 500 samples for training, and for testing 1500 (`grid`) or 5000 (`ring, two moons`) samples, following Sutherland et al. (2017).

We visualize the samples generated by the learned sampler, and compare it with the training datasets in the first row in Figure 1. The learned f is also plotted in the second row in Figure 1. The DDE learned models generate samples that cover the training data,

[1] https://github.com/karlnapf/nystrom-kexpfam
[2] https://github.com/MichaelArbel/KCEF



(a) initialization (b) 500-th (c) 2500-th



(d) 5000-th (c) 10000-th (d) 20000-th

Figure 2: The DDE estimators on `rings` dataset in each iteration. The blue points are sampled from the learned model. With the algorithm proceeds, the learned distribution converges to the ground-truth.

Table 1: Quantitative evaluation on synthetic data. MMD in $\times 10^{-3}$ scale on test set is reported.

	MMD		time (s)	
Datasets	DDE	KEF	DDE-sample	KEF+HMC
two moons	**0.11**	1.32	**0.07**	13.04
ring	**1.53**	3.19	**0.07**	14.12
grid	**1.32**	40.39	**0.04**	4.43

showing the ability of the DDE for estimating kernel exponential family on complicated data.

Then, we demonstrate the convergence of the DDE in Figure 2 on `rings` dataset. We initialize with a random dual distribution. As the algorithm iterates, the distribution converges to the target true distribution, justifying the convergence guarantees. More results on 2-dimensional `grid` and `two moons` can be found in Figure 3 in Appendix C. The DDE algorithm behaves similarly on these two datasets.

Finally, we compare the MMD between the generated samples from the learned model to the training data with the current state-of-the-art method for KEF (Sutherland et al., 2017), which performs better than KDE and other alternatives (Strathmann et al., 2015). We use HMC to generate the samples from the KEF learned model, while in the DDE, we can bypass the HMC step by the learned sampler. The computation time for inference is listed in Table 1. It shows the DDE improves the inference efficiency in orders. The MMD comparison is listed in Table 1. The proposed DDE estimator performs significantly better than the best performances by KEF in terms of MMD.

Conditional model extension. In this part of the experiment, models are trained to estimate the conditional distribution $p(y|x)$ on the benchmark datasets for studying these methods (Arbel and Gretton, 2017; Sugiyama et al., 2010). We centered and normalized the data and randomly split the datasets into a training and a testing set with equal size, as in Arbel and Gretton (2017). We evaluate the performances by the negative *log*-likelihood. Besides the score matching based KCEF, we also compared with LS-CDE and ϵ-KDE, introduced in (Sugiyama et al., 2010). The empirical results are summarized in Table 2.

Table 2: The negative log-likelihood comparison on benchmarks. Mean and std are calculated on 20 runs of different train/test splits. KCEF gets numerically unstable on two datasets, where we mark as N/A.

	DDE	KCEF	ϵ-KDE	LSCDE
geyser	**0.55 ± 0.07**	1.21 ± 0.04	1.11 ± 0.02	0.7 ± 0.01
caution	**0.95 ± 0.19**	**0.99 ± 0.01**	1.25 ± 0.19	1.19 ± 0.02
ftcollinssnow	**1.49 ± 0.14**	**1.46 ± 0.0**	1.53 ± 0.05	1.56 ± 0.01
highway	**1.18 ± 0.30**	**1.17 ± 0.01**	2.24 ± 0.64	1.98 ± 0.04
snowgeese	**0.42 ± 0.31**	**0.72 ± 0.02**	1.35 ± 0.17	1.39 ± 0.05
GAGurine	**0.43 ± 0.15**	**0.46 ± 0.0**	0.92 ± 0.05	0.7 ± 0.01
topo	1.02 ± 0.31	**0.67 ± 0.01**	1.18 ± 0.09	0.83 ± 0.0
CobarOre	**1.45 ± 0.23**	3.42 ± 0.03	**1.65 ± 0.09**	**1.61 ± 0.02**
mcycle	**0.60 ± 0.15**	**0.56 ± 0.01**	1.25 ± 0.23	0.93 ± 0.01
BigMac2003	**0.47 ± 0.36**	**0.59 ± 0.01**	1.29 ± 0.14	1.63 ± 0.03
cpus	**-0.63 ± 0.77**	N/A	1.01 ± 0.10	1.04 ± 0.07
crabs	**-0.60 ± 0.26**	N/A	0.99 ± 0.09	-0.07 ± 0.11
birthwt	**1.22 ± 0.15**	**1.18 ± 0.13**	1.48 ± 0.01	1.43 ± 0.01
gilgais	**0.61 ± 0.10**	**0.65 ± 0.08**	1.35 ± 0.03	0.73 ± 0.05
UN3	**1.03 ± 0.09**	1.15 ± 0.21	1.78 ± 0.14	1.42 ± 0.12
ufc	**1.03 ± 0.10**	**0.96 ± 0.14**	1.40 ± 0.02	**1.03 ± 0.01**

Although these datasets are low-dimensional with few samples and the KCEF uses the anisotropic RBF kernel (*i.e.*, different bandwidth in each dimension, making the experiments preferable to the KCEF), the proposed DDE still outperforms the competitors on six datasets significantly, and achieves comparable performance on the rest, even though it uses a simple isotropic RBF kernel. This further demonstrates the statistical power of the proposed DDE, comparing to the score matching estimator.

7 Conclusion

In this paper, we exploit the *doubly dual embedding* to reformulate the penalized MLE to a novel saddle-point optimization, which bypasses the intractable integration and provides flexibility in parameterizing the dual distribution. The saddle point view reveals a unique understanding of GANs and leads to a practical algorithm, which achieves state-of-the-art performance. We also establish the statistical consistency and algorithm convergence guarantee for the proposed algorithm. Although the transport mapping parametrization is flexible enough, it requires extra optimization for the KL-divergence estimation. For the future work, we will exploit dynamic-based sampling methods to design new parametrization, which shares both flexibility and density tractability.

Acknowledgements
We thank Michael Arbel and the anonymous reviewers for their insightful comments and suggestions. NH is supported in part by NSF-CRII-1755829, NSF-CMMI-1761699, and NCSA Faculty Fellowship.

References

Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *COLT*, pages 139–153, 2006.

Michael Arbel and Arthur Gretton. Kernel conditional exponential family. In *AISTATS*, pages 1337–1346, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. In *ICML*, 2017.

Francis R. Bach. On the equivalence between quadrature rules and random features. *Journal of Machine Learning Research*, 18:714–751, 2017.

Andrew R Barron and Chyong-Hwa Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, pages 1347–1369, 1991.

Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018.

Lawrence D. Brown. *Fundamentals of Statistical Exponential Families*, volume 9 of *Lecture notes-monograph series*. Institute of Mathematical Statistics, Hayward, Calif, 1986.

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

Bo Dai, Bo Xie, Niao He, Yingyu Liang, Anant Raj, Maria-Florina Balcan, and Le Song. Scalable kernel methods via doubly stochastic gradients. In *NeurIPS*, 2014.

Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual embeddings. In *AISTATS*, pages 1458–1467, 2017.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *ICML*, pages 1133–1142, 2018.

A. Devinatz. Integral representation of pd functions. *Trans. AMS*, 74(1):56–77, 1953.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017.

Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, 8:1217–1260, 2007.

Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*, volume 28. SIAM 1999.

Kenji Fukumizu. *Exponential manifold by reproducing kernel Hilbert spaces*, page 291306. Cambridge University Press, 2009.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23 (4):2341–2368, 2013.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014.

Chong Gu and Chunfu Qiu. Smoothing spline density estimation: Theory. *The Annals of Statistics*, pages 217–234, 1993.

Matthias Hein and Olivier Bousquet. Kernels, associated structures, and generalizations. Technical Report 127, Max Planck Institute for Biological Cybernetics, 2004.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

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

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *NeurIPS*, pages 4743–4751, 2016.

Jyrki Kivinen, Alex Smola, and Robert C. Williamson. Online learning with kernels. *IEEE Transactions on Signal Processing*, 52(8), Aug 2004.

Hermann König. *Eigenvalue Distribution of Compact Operators*. Birkhäuser, Basel, 1986.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabas Poczos. MMD GAN: Towards deeper understanding of moment matching network. In *NeurIPS*, pages 2203–2213, 2017.

Charles Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization*, 19(4):1574–1609, January 2009. ISSN 1052-6234.

XuanLong Nguyen, Martin Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *NeurIPS*, pages 1089–1096, 2008.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *NeurIPS*, 2016.

Giovanni Pistone and Maria Piera Rogantin. The exponential statistical manifold: mean parameters, orthogonality and space transformations. *Bernoulli*, 5 (4):721–760, 1999.

Ali Rahimi and Ben Recht. Random features for large-scale kernel machines. In *NeurIPS*, 2008.

Ali Rahimi and Ben Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *NeurIPS*, 2009.

Danilo J Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286, 2014.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, 2015.

R. T. Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematics Series*. Princeton University Press, Princeton, NJ, 1970.

Maurice Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *The Journal of Machine Learning Research*, 18(1):1830–1888, 2017.

Heiko Strathmann, Dino Sejdinovic, Samuel Livingstone, Zoltan Szabo, and Arthur Gretton. Gradient-free hamiltonian monte carlo with efficient kernel exponential families. In *NeurIPS*, pages 955–963, 2015.

Masashi Sugiyama, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. Conditional density estimation via least-squares density ratio estimation. In *AISTATS*, pages 781–788, 2010.

Dougal J Sutherland, Heiko Strathmann, Michael Arbel, and Arthur Gretton. Efficient and principled score estimation with nyström kernel exponential families. In *AISTATS*, 2018.

Shankar Vembu, Thomas Gärtner, and Mario Boley. Probabilistic structured predictors. In *UAI*, pages 557–564, 2009.

arXiv:1811.02228v3 [cs.LG] 24 Apr 2019

Kernel Exponential Family Estimation via Doubly Dual Embedding

*Bo Dai[1], *Hanjun Dai[2], Arthur Gretton[3], Le Song[2], Dale Schuurmans[1], Niao He[4]

[1] Google Brain, [2]Georgia Institute of Technology
[3]University College London, [4] University of Illinois at Urbana Champaign

April 25, 2019

Abstract

We investigate penalized maximum log-likelihood estimation for exponential family distributions whose natural parameter resides in a reproducing kernel Hilbert space. Key to our approach is a novel technique, *doubly dual embedding*, that avoids computation of the partition function. This technique also allows the development of a flexible sampling strategy that amortizes the cost of Monte-Carlo sampling in the inference stage. The resulting estimator can be easily generalized to kernel conditional exponential families. We establish a connection between kernel exponential family estimation and MMD-GANs, revealing a new perspective for understanding GANs. Compared to the score matching based estimators, the proposed method improves both memory and time efficiency while enjoying stronger statistical properties, such as fully capturing smoothness in its statistical convergence rate while the score matching estimator appears to saturate. Finally, we show that the proposed estimator empirically outperforms state-of-the-art methods in both kernel exponential family estimation and its conditional extension.

1 Introduction

The exponential family is one of the most important classes of distributions in statistics and machine learning. The exponential family possesses a number of useful properties (Brown, 1986)and includes many commonly used distributions with finite-dimensional natural parameters, such as the Gaussian, Poisson and multinomial distributions, to name a few. It is natural to consider generalizing the richness and flexibility of the exponential family to an *infinite*-dimensional parameterization via reproducing kernel Hilbert spaces (RKHS) (Canu and Smola, 2006).

Maximum log-likelihood estimation (MLE) has already been well-studied in the case of finite-dimensional exponential families, where desirable statistical properties such as asymptotic unbiasedness, consistency and asymptotic normality have been established. However, it is difficult to extend MLE to the infinite-dimensional case. Beyond the intractability of evaluating the partition function for a general exponential family, the necessary conditions for maximizing log-likelihood might not be feasible; that is, there may exist no solution to the KKT conditions in the infinite-dimensional case (Pistone and Rogantin, 1999; Fukumizu, 2009). To address this issue, Barron and Sheu (1991); Gu and Qiu (1993); Fukumizu (2009) considered several ways to regularize the function space by constructing (a series of) finite-dimensional spaces that approximate the original RKHS, yielding a tractable estimator in the restricted finite-dimensional space. However, as Sriperumbudur et al. (2017) note, even with the finite-dimension approximation, these algorithms are still expensive as every update requires Monte-Carlo sampling from a current model to compute the partition function.

An alternative score matching based estimator has recently been introduced by Sriperumbudur et al. (2017). This approach replaces the KullbackLeibler (KL) with the Fisher divergence, defined by the expected squared distance between the score of the model (*i.e.*, the derivative of the log-density) and the target distribution

*indicates equal contribution. Email: bodai@google.com, hanjundai@gatech.edu

score (Hyvärinen, 2005). By minimizing the Tikhonov-regularized Fisher divergence, Sriperumbudur et al. (2017) develop a computable estimator for the infinite-dimensional exponential family that also obtains a consistency guarantee. Recently, this method has been generalized to *conditional* infinite-dimensional exponential family estimation (Arbel and Gretton, 2017). Although score matching avoids computing the generally intractable integral, it requires computing and saving the first- and second-order derivatives of the reproducing kernel for *each* dimension on *each* sample. For n samples with d features, this results in $\mathcal{O}\left(n^2d^2\right)$ memory and $\mathcal{O}\left(n^3d^3\right)$ time cost respectively, which becomes prohibitive for datasets with moderate sizes of n and d. To alleviate this cost, Sutherland et al. (2017) utilize the m-rank Nyström approximation to the kernel matrix, reducing memory and time complexity to $\mathcal{O}\left(nmd^2\right)$ and $\mathcal{O}\left(m^3d^3\right)$ respectively. Although this reduces the cost dependence on sample size, the dependence on d is unaffected, hence the estimator remains unsuitable for high-dimensional data. Estimating a general exponential family model by either MLE or score matching also generally requires some form of Monte-Carlo sampling, such as MCMC or HMC, to perform inference, which significantly adds to the computation cost.

In summary, both the MLE and score matching based estimators for the infinite-dimensional exponential family incur significant computational overhead, particularly in high-dimensional applications.

In this paper, we revisit penalized MLE for the kernel exponential family and propose a new estimation strategy. Instead of solving the log-likelihood equation directly, as in existing MLE methods, we exploit a *doubly dual embedding* technique that leads to a novel saddle-point reformulation for the MLE (along with its conditional distribution generalization) in Section 3. We then propose a stochastic algorithm for the new view of penalized MLE in Section 4. Since the proposed estimator is based on penalized MLE, it does not require the first- and second-order derivatives of the kernel, as in score matching, greatly reducing the memory and time cost. Moreover, since the saddle point reformulation of penalized MLE avoids the intractable integral, the need for a Monte-Carlo step is also bypassed, thus accelerating the learning procedure. This approach also learns a flexibly parameterized sampler simultaneously, therefore, further reducing inference cost. We present the consistency and rate of the new estimation strategy in the well-specified case, *i.e.*, the true density belongs to the kernel exponential family, and an algorithm convergence guarantee in Section 5. We demonstrate the empirical advantages of the proposed algorithm in Section 6, comparing to state-of-the-art estimators for both the kernel exponential family and its conditional extension (Sutherland et al., 2017; Arbel and Gretton, 2017).

2 Preliminaries

We first provide a preliminary introduction to the exponential family and Fenchel duality, which will play vital roles in the derivation of the new estimator.

2.1 Exponential family

The natural form of the exponential family over Ω with the sufficient statistics $f \in \mathcal{F}$ is defined as

$$p_f(x) = p_0(x) \exp\left(\lambda f(x) - A\left(\lambda f\right)\right), \tag{1}$$

where $A\left(\lambda f\right) := \log \int_\Omega \exp\left(\lambda f\left(x\right)\right) p_0\left(x\right) dx$, $x \in \Omega \subset \mathbb{R}^d$, $\lambda \in \mathbb{R}$, and $\mathcal{F} := \{f \in \mathcal{H} : \exp\left(A\left(\lambda f\right)\right) < \infty\}$. In this paper, we mainly focus on the case where $\mathcal{H}$ is a reproducing Hilbert kernel space (RKHS) with kernel $k\left(x, x'\right)$, such that $f\left(x\right) = \langle f, k\left(x, \cdot\right)\rangle$. However, we emphasize that the proposed algorithm in Section 4 can be easily applied to arbitrary differentiable function parametrizations, such as deep neural networks.

Given samples $\mathcal{D} = [x_i]_{i=1}^N$, a model with a finite-dimensional parameterization can be learned via maximum log-likelihood estimation (MLE),

$$\max_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \log p_f(x_i) = \widehat{\mathbb{E}}_{\mathcal{D}}\left[\lambda f(x) + \log p_0\left(x\right)\right] - A(\lambda f),$$

where $\widehat{\mathbb{E}}_{\mathcal{D}}[\cdot]$ denotes the empirical expectation over $\mathcal{D}$. The MLE (2) is well-studied and has nice properties. However, the MLE is known to be "ill-posed" in the infinite-dimensional case, since the optimal solution might not be exactly achievable in the representable space. Therefore, penalized MLE has been introduced for the finite (Dudík et al., 2007) and infinite dimensional (Gu and Qiu, 1993; Altun and Smola, 2006) exponential families respectively. Such regularization essentially relaxes the moment matching constraints in terms of some norm, as shown later, guaranteeing the existence of a solution. In this paper, we will also focus on MLE with RKHS norm regularization.

One useful theoretical property of MLE for the exponential family is convexity w.r.t. f. With convex regularization, *e.g.*, $\|f\|_{\mathcal{H}}^2$, one can use stochastic gradient descent to recover a unique global optimum. Let $L(f)$ denote RKHS norm penalized log-likelihood, *i.e.*,

$$L(f) := \frac{1}{N}\sum_{i=1}^{N} \log p_f(x_i) - \frac{\eta}{2}\|f\|_{\mathcal{H}}^2 . \tag{2}$$

where $\eta > 0$ denotes the regularization parameter. The gradient of $L(f)$ w.r.t. f can be computed as

$$\nabla_f L = \widehat{\mathbb{E}}_{\mathcal{D}}[\lambda \nabla_f f(x)] - \nabla_f A(\lambda f) - \eta f. \tag{3}$$

To calculate the $\nabla_f A(\lambda f)$ in (3), we denote $Z(\lambda f) = \int_\Omega \exp(\lambda f(x)) p_0(x)\, dx$ and expand the partition function by definition,

$$\begin{aligned} \nabla_f A(\lambda f) &= \frac{1}{Z(\lambda f)} \nabla_f \int_\Omega \exp(\lambda f(x)) p_0(x)\, dx \\ &= \int_\Omega \frac{p_0(x)\exp(\lambda f(x))}{Z(\lambda f)} \nabla_f \lambda f(x) dx \\ &= \mathbb{E}_{p_f(x)}[\nabla_f \lambda f(x)] . \end{aligned} \tag{4}$$

One can approximate the $\mathbb{E}_{p_f(x)}[\nabla_f \lambda f(x)]$ by AIS (Neal, 2001) or MCMC samples (Vembu et al., 2009), which leads to the Contrastive Divergence (CD) algorithm (Hinton, 2002).

To avoid costly MCMC sampling in estimating the gradient, Sriperumbudur et al. (2017) construct an estimator based on score matching instead of MLE, which minimizes the penalized Fisher divergence. Plugging the kernel exponential family into the empirical Fisher divergence, the optimization reduces to

$$J(f) := \frac{\lambda^2}{2}\langle f, \widehat{C} f\rangle_{\mathcal{H}} + \lambda \langle f, \hat{\delta}\rangle_{\mathcal{H}} + \frac{\eta}{2}\|f\|_{\mathcal{H}}^2 , \tag{5}$$

where

$$\begin{aligned} \widehat{C} &:= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} \partial_j k(x_i, \cdot) \otimes \partial_j k(x_i, \cdot) , \\ \hat{\delta} &:= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} \partial_j k(x_i, \cdot)\left(\partial_j \log p_0(x_i) + \partial_j^2 k(x_i, \cdot)\right) . \end{aligned}$$

As we can see, the score matching objective (5) is convex and does not involve the intractable integral in $A(\lambda f)$. However, such an estimator requires the computation of first- and second-order of derivatives of kernel for each dimension on each data, leading to memory and time cost of $\mathcal{O}\left(n^2 d^2\right)$ and $\mathcal{O}\left(n^3 d^3\right)$ respectively. This quickly becomes prohibitive for even modest n and d.

The same difficulty also appears in the score matching based estimator in (Arbel and Gretton, 2017) for the *conditional exponential family*, which is defined as

$$p(y|x) = p_0(y)\exp(\lambda f(x, y) - A_x(\lambda f)) , \ f \in \mathcal{F} \tag{6}$$

where $y \in \Omega_y \subset \mathbb{R}^p$, $x \in \Omega_x \subset \mathbb{R}^d$, $\lambda \in \mathbb{R}$, and $A_x(\lambda f) := \log \int_{\Omega_y} p_0(y) \exp(\lambda f(x,y))\, dy$, $\mathcal{F}$ is defined as $\{f \in \mathcal{H}_y : \exp(A_x(\lambda f)) < \infty\}$. We consider $f : \Omega_x \times \Omega_y \to \mathbb{R}$ such that $f(x,\cdot)$ is in RKHS $\mathcal{H}_y$ for $\forall x \in \Omega_x$. Denoting $\mathcal{T} \in \mathcal{H}_{\Omega_x} : \Omega_x \to \mathcal{H}_y$ such that $\mathcal{T}_x(y) = f(x,y)$, we can derive its kernel function following Micchelli and Pontil (2005); Arbel and Gretton (2017). By the Riesz representation theorem, $\forall x \in \Omega_x$ and $h \in \mathcal{H}_y$, there exists a linear operator $\Gamma_x : \mathcal{H}_y \to \mathcal{H}_{\Omega_x}$ such that

$$\langle h, \mathcal{T}_x \rangle_{\mathcal{H}_y} = \langle \mathcal{T}, \Gamma_x h \rangle_{\mathcal{H}_{\Omega_x}}, \ \forall \mathcal{T} \in \mathcal{H}_{\Omega_x}.$$

Then, the kernel can be defined by composing Γ_x with its dual, *i.e.*, $k(x,x') = \Gamma_x^* \Gamma_{x'}$ and the function $f(x,y) = \mathcal{T}_x(y) = \langle \mathcal{T}, \Gamma_x k(y,\cdot) \rangle$. We follow such assumptions for kernel conditional exponential family.

2.2 Convex conjugate and Fenchel duality

Denote $h(\cdot)$ as a function $\mathbb{R}^d \to \mathbb{R}$, then its convex conjugate function is defined as

$$h^*(u) = \sup_{v \in \mathbb{R}^d} \{u^\top v - h(v)\}.$$

If $h(\cdot)$ is proper, convex and lower semicontinuous, the conjugate function, $h^*(\cdot)$, is also proper, convex and lower semicontinuous. Moreover, h and h^* are dual to each other, *i.e.*, $(h^*)^* = h$. Such a relationship is known as Fenchel duality (Rockafellar, 1970; Hiriart-Urruty and Lemaréchal, 2012). By the conjugate function, we can represent the h by as,

$$h(v) = \sup_{u \in \mathbb{R}^d} \{v^\top u - h^*(u)\}.$$

The supremum achieves if $v \in \partial h^*(u)$, or equivalently $u \in \partial h(v)$.

3 A Saddle-Point Formulation of Penalized MLE

As discussed in Section 2, the penalized MLE for the exponential family involves computing the log-partition functions, $A(\lambda f)$ and $A_x(\lambda f)$, which are intractable in general. In this section, we first introduce a saddle-point reformulation of the penalized MLE of the exponential family, using Fenchel duality to bypass the computation of the intractable log-partition function. This approach can also be generalized to the conditional exponential family. First, observe that we can rewrite the log-partition function $A(\lambda f)$ via Fenchel duality as follows.

Theorem 1 (Fenchel dual of log-partition)

$$A(\lambda f) = \max_{q \in \mathcal{P}} \ \lambda \langle q(x), f(x) \rangle_2 - KL(q||p_0), \tag{7}$$

$$p_f(x) = \operatorname*{argmax}_{q \in \mathcal{P}} \ \lambda \langle q(x), f(x) \rangle_2 - KL(q||p_0), \tag{8}$$

where $\langle f, g \rangle_2 := \int_\Omega f(x) g(x)\, dx$, $\mathcal{P}$ *denotes the space of distributions with bounded* L_2 *norm and* $KL(q||p_0) := \int_\Omega q(x) \log \frac{q(x)}{p_0(x)} dx$.

Proof Denote $l(q) := \lambda \langle q(x), f(x) \rangle - KL(q||p_0)$, which is strongly concave w.r.t. $q \in \mathcal{P}$, the optimal q^* can be obtained by setting the

$$\log q^*(x) \propto \lambda f(x) + \log p_0(x).$$

Since $q^* \in \mathcal{P}$, we have

$$q^*(x) = p_0(x) \exp(\lambda f(x) - A(\lambda f)) = p_f(x),$$

which leads to (8). Plugging q^* to $l(q)$, we obtain the maximum as $\log \int_\Omega \exp(\lambda f(x))\, p_0(x) dx$, which is exactly $A(\lambda f)$, leading to (7). ■

Therefore, invoking the Fenchel dual of $A(\lambda f)$ into the penalized MLE, we achieve a saddle-point optimization, *i.e.*,

$$\max_{f\in\mathcal{F}} L(f) \propto \min_{q\in\mathcal{P}} \underbrace{\widehat{\mathbb{E}}_{\mathcal{D}}[f(x)] - \mathbb{E}_{q(x)}[f(x)] - \frac{\eta}{2}\|f\|_{\mathcal{H}}^2 + \frac{1}{\lambda}KL(q||p_0)}_{\ell(f,q)}. \tag{9}$$

The min-max reformulation of the penalized MLE in (9) resembles the optimization in MMD GAN (Li et al., 2017; Bińkowski et al., 2018). In fact, the dual problem of the penalized MLE of the exponential family is a *KL-regularized* MMD GAN with a special design of the kernel family. Alternatively, if f is a Wasserstein-1 function, the optimization resembles the Wasserstein GAN (Arjovsky et al., 2017).

Next, we consider the duality properties.

Theorem 2 (weak and strong duality) *The weak duality holds in general,* i.e.*,*

$$\max_{f\in\mathcal{F}}\min_{q\in\mathcal{P}}\ \ell(f,q) \leqslant \min_{q\in\mathcal{P}}\max_{f\in\mathcal{F}}\ \ell(f,q).$$

The strong duality holds when $\mathcal{F}$ is a closed RKHS and $\mathcal{P}$ is the distributions with bounded L_2 norm, i.e.*,*

$$\max_{f\in\mathcal{F}}\min_{q\in\mathcal{P}}\ \ell(f,q) = \min_{q\in\mathcal{P}}\max_{f\in\mathcal{F}}\ \ell(f,q). \tag{10}$$

Theorem 2 can be obtained by directly applying the minimax theorem (Ekeland and Temam, 1999)[Proposition 2.1]. We refer to the max-min problem in (10) as the primal problem, while the min-max form as the its dual form.

Remark (connections to MMD GAN): Consider the dual problem with the kernel learning, *i.e.*,

$$\min_{q\in\mathcal{P}}\max_{f\in\mathcal{F}_\phi,\phi} \widehat{\mathbb{E}}_{\mathcal{D}}[f(x)] - \mathbb{E}_{q(x)}[f(x)] - \frac{\eta}{2}\|f\|_{\mathcal{H}}^2 + \frac{1}{\lambda}KL(q||p_0). \tag{11}$$

where we involve the parameters of the kernel ϕ to be learned in the optimization. By setting the gradient $\nabla_f \ell(f,q) = 0$, for fixed q, we obtain the optimal witness function in the RKHS, $f_q = \frac{1}{\eta}\left(\widehat{\mathbb{E}}[k_\phi(x,\cdot)] - \mathbb{E}_q[k_\phi(x,\cdot)]\right)$, which leads to

$$\min_{q\in\mathcal{P}}\max_{\phi} \underbrace{\widehat{\mathbb{E}}[k_\phi(x,x')] - 2\widehat{\mathbb{E}}\mathbb{E}_q[k_\phi(x,x')] + \mathbb{E}_q[k_\phi(x,x')]}_{MMD_\phi(\mathcal{D},q)} + \frac{2\eta}{\lambda}KL(q||p_0). \tag{12}$$

This can be regarded as the KL-divergence regularized MMD GAN. Thus, with KL-divergence regularization, the MMD GAN learns an infinite-dimension exponential family in an adaptive RKHS. Such a novel perspective bridges GAN and exponential family estimation, which appears to be of independent interest and potentially brings a new connection to the GAN literature for further theoretical development.

Remark (connections to Maximum Entropy Moment Matching): Altun and Smola (2006); Dudík et al. (2007) discuss the maximum entropy moment matching method for distribution estimation,

$$\begin{aligned} \min_{q\in\mathcal{P}} \quad & KL(q||p_0) \\ \text{s.t.} \quad & \left\|\mathbb{E}_q[k(x,\cdot)] - \widehat{\mathbb{E}}[k(x,\cdot)]\right\|_{\mathcal{H}_k} \leqslant \frac{\eta'}{2}, \end{aligned} \tag{13}$$

whose dual problem will be reduced to the penalized MLE (2) with proper choice of η' (Altun and Smola, 2006)[Lemma 6]. Interestingly, the proposed saddle-point formulation (9) shares the solution to (13). From the maximum entropy view, the penalty $\|f\|_{\mathcal{H}}$ relaxes the moment matching constraints. However, the

algorithms provided in Altun and Smola (2006); Dudík et al. (2007) simply ignore the difficulty in computing the expectation in $A(\lambda f)$, which is not practical, especially when f is infinite-dimensional.

Similar to Theorem 1, we can also represent $A_x(\lambda f)$ by its Fenchel dual,

$$A_x(\lambda f) = \max_{q(\cdot|x)\in\mathcal{P}} \lambda \langle q(y|x), f(x,y)\rangle - KL(q||p_0), \tag{14}$$

$$p_f(y|x) = \operatorname*{argmax}_{q(\cdot|x)\in\mathcal{P}} \lambda \langle q(x), f(x)\rangle - KL(q||p_0). \tag{15}$$

Then, we can recover the penalized MLE for the conditional exponential family as

$$\max_{f\in\mathcal{F}} \min_{q(\cdot|x)\in\mathcal{P}} \widehat{\mathbb{E}}_{\mathcal{D}}[f(x,y)] - \mathbb{E}_{q(y|x)}[f(x,y)] - \frac{\eta}{2}\|f\|_{\mathcal{H}}^2 + \frac{1}{\lambda} KL(q||p_0).$$

The saddle-point reformulations of penalized MLE in (9) and (16) bypass the difficulty in the partition function. Therefore, it is very natural to consider learning the exponential family by solving the saddle-point problem (9) with an appropriate parametrized dual distribution $q(x)$. This approach is referred to as the "dual embedding" technique in Dai et al. (2016), which requires:

i) the parametrization family should be flexible enough to reduce the extra approximation error;

ii) the parametrized representation should be able to provide density value.

As we will see in Section 5, the flexibility of the parametrization of the dual distribution will have a significant effect on the consistency of the estimator.

One can of course use the kernel density estimator (KDE) as the dual distribution parametrization, which preserves convex-concavity. However, KDE will easily fail when approximating high-dimensional data. Applying the reparametrization trick with a suitable class of probability distributions (Kingma and Welling, 2013; Rezende et al., 2014) is another alternative to parametrize the dual distribution. However, the class of such parametrized distributions is typically restricted to simple known distributions, which might not be able to approximate the true solution, potentially leading to a huge approximation bias. At the other end of the spectrum, the distribution family generated by transport mapping is sufficiently flexible to model smooth distributions (Goodfellow et al., 2014; Arjovsky et al., 2017). However, the density value of such distributions, *i.e.*, $q(x)$, is not available for KL-divergence computation, and thus, is not applicable for parameterizing the dual distribution. Recently, flow-based parametrized density functions (Rezende and Mohamed, 2015; Kingma et al., 2016; Dinh et al., 2016) have been proposed for a trade-off between the flexibility and tractability. However, the expressive power of existing flow-based models remains restrictive even in our synthetic example and cannot be directly applied to conditional models.

3.1 Doubly Dual Embedding for MLE

Transport mapping is very flexible for generating smooth distributions. However, a major difficulty is that it lacks the ability to obtain the density value $q(x)$, making the computation of the KL-divergence impossible. To retain the flexibility of transport mapping and avoid the calculation of $q(x)$, we introduce *doubly dual embedding*, which achieves a delicate balance between flexibility and tractability.

First, noting that KL-divergence is also a convex function, we consider the Fenchel dual of KL-divergence (Nguyen et al., 2008), *i.e.*,

$$KL(q||p_0) = \max_{\nu} \mathbb{E}_q[\nu(x)] - \mathbb{E}_{p_0}[\exp(\nu(x))] + 1, \tag{16}$$

$$\log\frac{q(x)}{p_0(x)} = \operatorname*{argmax}_{\nu} \mathbb{E}_q[\nu(x)] - \mathbb{E}_{p_0}[\exp(\nu(x))] + 1, \tag{17}$$

with $\nu(\cdot): \Omega \to \mathbb{R}$. One can see in the dual representation of the KL-divergence that the introduction of the auxiliary optimization variable ν eliminates the explicit appearance of $q(x)$ in (16), which makes the

Algorithm 1 Doubly Dual Embedding-SGD for saddle-point reformulation of MLE (18)

for $l = 1, \ldots, L$ **do**
  Compute (f, ν) by Algorithm 2.
  Sample $\{\xi_0 \sim p(\xi)\}_{b=1}^{B}$.
  Generate $x_b = g_{w_g}(\xi)$ for $b = 1, \ldots, B$.
  Compute stochastic approximation $\widehat{\nabla}_{w_g} \widehat{L}(w_g)$ by (22).
  Update $w_g^{l+1} = w_g^l - \rho_l \widehat{\nabla}_{w_g} L(w_g)$.
end for
Output w_g and f.

Algorithm 2 Stochastic Functional Gradients for f and ν

for $k = 1, \ldots, K$ **do**
  Sample $\xi \sim p(\xi)$.
  Generate $x = g(\xi)$.
  Sample $x' \sim p_0(x)$.
  Compute stochastic function gradient w.r.t. f and ν with (19) and (20).
  Update f_k and ν_k with (21) and (22).
end for
Output (f_K, ν_K) .

transport mapping parametrization for $q(x)$ applicable. Since there is no extra restriction on ν, we can use arbitrary smooth function approximation for $\nu(\cdot)$, such as kernel functions or neural networks.

Applying the dual representation of KL-divergence into the dual problem of the saddle-point reformulation (9), we obtain the ultimate saddle-point optimization for the estimation of the kernel exponential family,

$$\min_{q \in \mathcal{P}} \max_{f, \nu \in \mathcal{F}} \tilde{\ell}(f, \nu, q) := \widehat{\mathbb{E}}_{\mathcal{D}}[f] - \mathbb{E}_q[f] - \frac{\eta}{2}\|f\|_{\mathcal{H}}^2 + \frac{1}{\lambda}\left(\mathbb{E}_q[\nu] - \mathbb{E}_{p_0}[\exp(\nu)]\right). \tag{18}$$

Note that several other work also applied Fenchel duality to KL-divergence (Nguyen et al., 2008; Nowozin et al., 2016), but the approach taken here is different as it employs dual of $KL(q||p_0)$, rather than $KL(q||\mathcal{D})$.

Remark (Extension for kernel conditional exponential family): Similarly, we can apply the doubly dual embedding technique to the penalized MLE of the kernel conditional exponential family, which leads to

$$\min_{q \in \mathcal{P}_x} \max_{f, \nu \in \mathcal{F}} \widehat{\mathbb{E}}_{\mathcal{D}}[f] - \mathbb{E}_{q(y|x), x \sim \mathcal{D}}[f] - \frac{\eta}{2}\|f\|_{\mathcal{H}}^2 + \frac{1}{\lambda}\left(\mathbb{E}_{q(y|x), x \sim \mathcal{D}}[\nu] - \mathbb{E}_{p_0(y)}[\exp(\nu)]\right).$$

In summary, with the *doubly dual embedding* technique, we derive a saddle-point reformulation of penalized MLE that bypasses the difficulty of handling the intractable partition function while allowing great flexibility in parameterizing the dual distribution.

4 Practical Algorithm

In this section, we introduce the transport mapping parametrization for the dual distribution, $q(x)$, and apply the stochastic gradient descent for solving the optimization problems in (18) and (19). For simplicity of exposition, we only illustrate the algorithm for (18). The algorithm can be easily applied to (19).

Denote the parameters in the transport mapping for $q(x)$ as w_g, such that $\xi \sim p(\xi)$ and $x = g_{w_g}(\xi)$. We illustrate the algorithm with a kernel parametrize $\nu(\cdot)$. There are many alternative choices of parametrization of ν, *e.g.*, neural networks—the proposed algorithm is still applicable to the parametrization as long as

it is differentiable. We abuse notation somewhat by using $\tilde{\ell}(f,\nu,w_g)$ as $\tilde{\ell}(f,\nu,q)$ in (18). with such parametrization, the (18) is an upper bound of the penalty MLE in general by Theorem 2.

With the kernel parametrized (f,ν), the inner maximization over f and ν is a standard concave optimization. We can solve it using existing algorithms to achieve the global optimal solution. Due to the existence of the expectation, we will use the stochastic functional gradient descent for scalability. Given $(f,\nu)\in\mathcal{H}$, following the definition of functional gradients (Kivinen et al., 2004; Dai et al., 2014), we have

$$\zeta_f(\cdot) := \nabla_f \tilde{\ell}(f,\nu,w_g) = \widehat{\mathbb{E}}_{\mathcal{D}}\left[k(x,\cdot)\right] - \mathbb{E}_{\xi}\left[k\left(g_{w_g}(\xi_0),\cdot\right)\right] - \eta f(\cdot), \tag{19}$$

$$\zeta_\nu(\cdot) := \nabla_\nu \tilde{\ell}(f,\nu,w_g) = \frac{1}{\lambda}\left(\mathbb{E}_{\xi}\left[k\left(g_{w_g}(\xi),\cdot\right)\right] - \mathbb{E}_{p_0}\left[\exp\left(\nu(x)\right)k(x,\cdot)\right]\right). \tag{20}$$

In the k-th iteration, given sample $x\sim\mathcal{D}, \xi\sim p(\xi)$, and $x'\sim p_0(x)$, the update rule for f and ν will be

$$f_{k+1}(\cdot) = (1-\eta\tau_k) f_k(\cdot) + \tau_k\left(k(x,\cdot) - k\left(g_{w_g}(\xi),\cdot\right)\right), \tag{21}$$

$$\nu_{k+1}(\cdot) = \nu_k(\cdot) + \frac{\tau_k}{\lambda}\left(k\left(g_{w_g}(\xi),\cdot\right) - \exp\left(\nu_k(x')\right)k(x',\cdot)\right),$$

where τ_k denotes the step-size.

Then, we consider the update rule for the parameters in dual transport mapping embedding.

Theorem 3 (Dual gradient) *Denoting* $\left(f^*_{w_g},\nu^*_{w_g}\right) = \mathrm{argmax}_{(f,\nu)\in\mathcal{H}}\,\tilde{\ell}(f,\nu,w_g)$ *and* $\widehat{L}(w_g) = \tilde{\ell}\left(f^*_{w_g},\nu^*_{w_g},w_g\right)$, *we have*

$$\nabla_{w_g}\widehat{L}(w_g) = -\mathbb{E}_{\xi}\left[\nabla_{w_g} f^*_{w_g}\left(g_{w_g}(\xi)\right)\right] + \frac{1}{\lambda}\mathbb{E}_{\xi}\left[\nabla_{w_g}\nu^*_{w_g}\left(g_{w_g}(\xi)\right)\right]. \tag{22}$$

Proof details are given in Appendix A.1. With this gradient estimator, we can apply stochastic gradient descent to update w_g iteratively. We summarize the updates for (f,ν) and w_g in Algorithm 1 and 2.

Compared to score matching based estimators (Sriperumbudur et al., 2017; Sutherland et al., 2017), although the convexity no longer holds for the saddle-point estimator, the doubly dual embedding estimator avoids representing f with derivatives of the kernel, thus avoiding the memory cost dependence on the square of dimension. In particular, the proposed estimator for f based on (18) reduces the memory cost from $\mathcal{O}\left(n^2d^2\right)$ to $\mathcal{O}\left(K^2\right)$ where K denotes the number of iterations in Algorithm 2. In terms of time cost, we exploit the stochastic update, which is naturally suitable for large-scale datasets and avoids the matrix inverse computation in score matching estimator whose cost is $\mathcal{O}\left(n^3d^3\right)$. We also learn the dual distribution simultaneously, which can be easily used to generate samples from the exponential family for inference, thus saving the cost of Monte-Carlo sampling in the inference stage. For detailed discussion about the computation cost, please refer to Appendix C.

Remark (random feature extension): Memory cost is a well-known bottleneck for applying kernel methods to large-scale problems. When we set $K=n$, the memory cost will be $\mathcal{O}\left(n^2\right)$, which is prohibitive for millions of data points. Random feature approximation (Rahimi and Recht, 2008; Dai et al., 2014; Bach, 2015) can be utilized for scaling up kernel methods. The proposed Algorithm 1 and Algorithm 2 are also compatible with random feature approximation, and hence, applicable to large-scale problems in the same way. With r random features, we can further reduce the memory cost of storing f to $\mathcal{O}(rd)$. However, even with a random feature approximation, the score matching based estimator will still require $\mathcal{O}\left(rd^2\right)$ memory. One can also learn random features by back-propagation, which leads to the neural networks extension. Please refer to the details of this variant of Algorithm 1 in Appendix B.

5 Theoretical Analysis

In this section, we will first provide the analysis of consistency and the sample complexity of the proposed estimator based on the saddle-point reformulation of the penalized MLE in the well-specified case, where the true density is assumed to be in the kernel (conditional) exponential family, following Sutherland et al. (2017); Arbel and Gretton (2017). Then, we consider the convergence property of the proposed algorithm. We mainly focus on the analysis for $p(x)$. The results can be easily extended to kernel conditional exponential family $p(y|x)$.

5.1 Statistical Consistency

We explicitly consider approximation error from the dual embedding in the consistency of the proposed estimator. We first establish some notation that will be used in the analysis. For simplicity, we set $\lambda = 1$ and $p_0(x) = 1$ improperly in the exponential family expression (1). We denote $p^*(x) = \exp(f^*(x) - A(f^*))$ as the ground-true distribution with its true potential function f^* and $\left(\tilde{f}, \tilde{q}, \tilde{v}\right)$ as the optimal primal and dual solution to the saddle point reformulation of the penalized MLE (18). The parametrized dual space is denoted as $\mathcal{P}_w$. We denote $p_{\tilde{f}} := \exp\left(\lambda\tilde{f} - A\left(\lambda\tilde{f}\right)\right)$ as the exponential family generated by $\tilde{f}$. We have the consistency results as

Theorem 4 *Assume the spectrum of kernel $k(\cdot,\cdot)$ decays sufficiently homogeneously in rate l^{-r}. With some other mild assumptions listed in Appendix A.2, we have as $\eta \to 0$ and $n\eta^{\frac{1}{r}} \to \infty$,*

$$KL\left(p^*||p_{\tilde{f}}\right) + KL\left(p_{\tilde{f}}||p^*\right) = \mathcal{O}_{p^*}\left(n^{-1}\eta^{-\frac{1}{r}} + \eta + \epsilon_{approx}\right),$$

where $\epsilon_{approx} := \sup_{f\in\mathcal{F}} \inf_{q\in\mathcal{P}_w} KL(q||p_f)$ denotes the approximate error due to the parametrization of $\tilde{q}$ and $\tilde{\nu}$. Therefore, when setting $\eta = \mathcal{O}\left(n^{-\frac{r}{1+r}}\right)$, $p_{\tilde{f}}$ converge s to p^* in terms of Jensen-Shannon divergence at rate $\mathcal{O}_{p^*}\left(n^{-\frac{r}{1+r}} + \epsilon_{approx}\right)$.

For the details of the assumptions and the proof, please refer to Appendix A.2. Recall the connection between the proposed model and MMD GAN as discussed in Section 3. Theorem 4 also provides a learnability guarantee for a class of GAN models as a byproduct. The most significant difference of the bound provided above, compared to Gu and Qiu (1993); Altun and Smola (2006), is the explicit consideration of the bias from the dual parametrization. Moreover, instead of the Rademacher complexity used in the sample complexity results of Altun and Smola (2006), our result exploits the spectral decay of the kernel, which is more directly connected to properties of the RKHS.

From Theorem 4 we can clearly see the effect of the parametrization of the dual distribution: if the parametric family of the dual distribution is simple, the optimization for the saddle-point problem may become easy, however, ϵ_{approx} will dominate the error. The other extreme case is to also use the kernel exponential family to parametrize the dual distribution and the parametric family of ν contains $\log\frac{p_f(x)}{p_0(x)}$, then, ϵ_{approx} will reduce to 0, however, the optimization will be difficult to handle. The saddle-point reformulation provides us the opportunity to balance the difficulty of the optimization with approximation error.

The statistical consistency rate of our estimator and the score matching estimator (Sriperumbudur et al., 2017) are derived under different assumptions, therefore, they are not directly comparable. However, since the smoothness is not fully captured by the score matching based estimator in Sriperumbudur et al. (2017), it only achieves $\mathcal{O}\left(n^{-\frac{2}{3}}\right)$ even if f is infinitely smooth. While under the case that dual distribution parametrization is relative flexible, *i.e.*, ϵ_{approx} is negligible, and the the spectrum of the kernel decay rate $r \to \infty$, the proposed estimator will converge in rate $\mathcal{O}\left(n^{-1}\right)$, which is significantly more efficient than the score matching method.

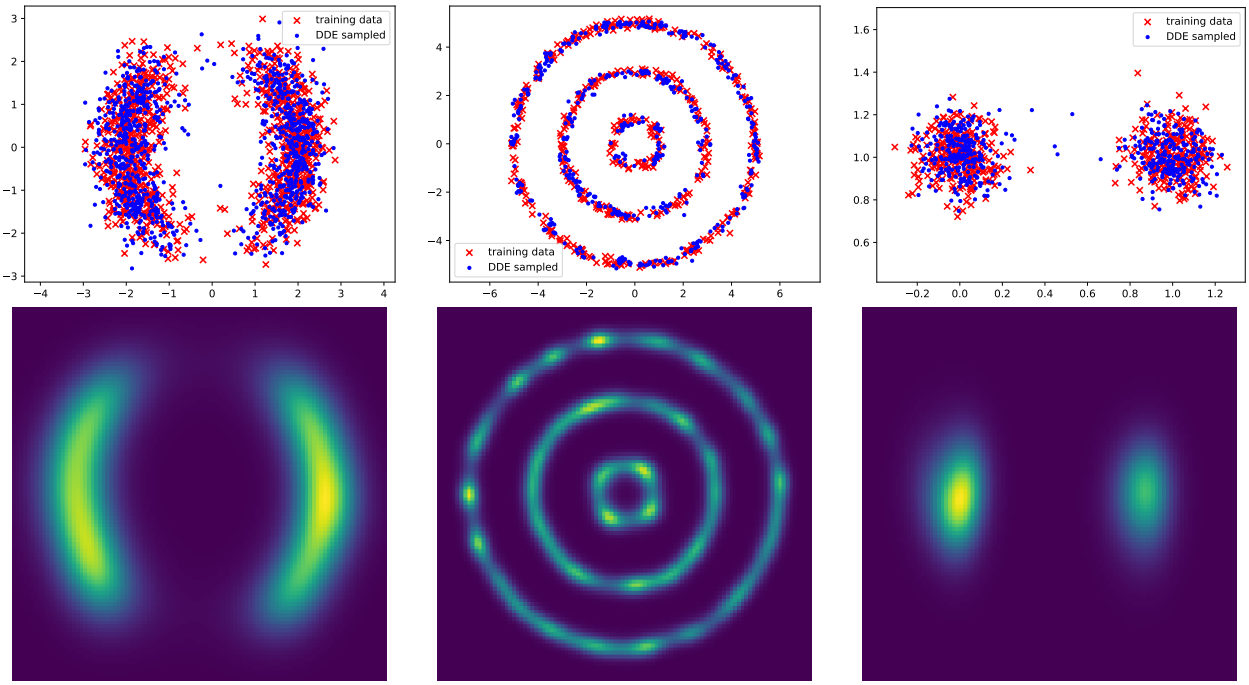


Figure 1: The blue points in the figures in first row show the generated samples by learned models, and the red points are the training samples. The learned f are illustrated in the second row.

5.2 Algorithm Convergence

It is well-known that the stochastic gradient descent converges for saddle-point problem with convex-concave property (Nemirovski et al., 2009). However, for better the dual parametrization to reduce ϵ_{approx} in Theorem 4, we parameterize the dual distribution with the nonlinear transport mapping, which breaks the convexity. In fact, by Theorem 3, we obtain the unbiased gradient w.r.t. w_g. Therefore, the proposed Algorithm 1 can be understood as applying the stochastic gradient descent for the non-convex dual minimization problem, *i.e.*, $\min_{w_g} \widehat{L}(w_g) := \tilde{\ell}\left(f^*_{w_g}, \nu^*_{w_g}, w_g\right)$. From such a view, we can prove the sublinearly convergence rate to a stationary point when stepsize is diminishing following Ghadimi and Lan (2013); Dai et al. (2017). We list the result below for completeness.

Theorem 5 *Assume that the parametrized objective $\widehat{L}(w_g)$ is C-Lipschitz and variance of its stochastic gradient is bounded by σ^2. Let the algorithm run for L iterations with stepsize $\rho_l = \min\{\frac{1}{L}, \frac{D'}{\sigma\sqrt{L}}\}$ for some $D' > 0$ and output $w_g^1, \ldots, w_g^L$. Setting the candidate solution to be $\widehat{w}_g$ randomly chosen from $w_g^1, \ldots, w_g^L$ such that $P(w = w_g^j) = \frac{2\rho_j - C\rho_j^2}{\sum_{j=1}^L (2\rho_j - C\rho_j^2)}$, then it holds that $\mathbb{E}\left[\left\|\nabla \widehat{L}(\widehat{w}_g)\right\|^2\right] \leqslant \frac{CD^2}{L} + (D' + \frac{D}{D'})\frac{\sigma}{\sqrt{L}}$ where $D := \sqrt{2(\widehat{L}(w_g^1) - \min \widehat{L}(w_g))/L}$ represents the distance of the initial solution to the optimal solution.*

The above result implies that under the choice of the parametrization of f, ν and g, the proposed Algorithm 1 converges sublinearly to a stationary point, whose rate will depend on the smoothing parameter.

6 Experiments

In this section, we compare the proposed doubly dual embedding (DDE) with the current state-of-the-art score matching estimators for kernel exponential family (KEF) Sutherland et al. (2017)[1] and its conditional extension (KCEF) Arbel and Gretton (2017)[2], respectively, as well as several competitors. We test the proposed estimator empirically following their setting. We use Gaussian RBF kernel for both exponential family and its conditional extension, $k(x, x') = \exp\left(-\|x - x'\|_2^2 / \sigma^2\right)$, with the bandwidth σ set by median-trick (Dai et al., 2014). For a fair comparison, we follow Sutherland et al. (2017) and Arbel and Gretton (2017) to set the $p_0(x)$ for kernel exponential family and its conditional extension, respectively. The dual variables are parametrized by MLP with 5 layers. More implementation details can be found in Appendix E and the code repository which is available at https://github.com/Hanjun-Dai/dde.

Density estimation We evaluate the DDE on the synthetic datasets, including `ring`, `grid` and `two moons`, where the first two are used in Sutherland et al. (2017), and the last one is from Rezende and Mohamed (2015). The `ring` dataset contains the points uniformly sampled along three circles with radii $(1, 3, 5) \in \mathbb{R}^2$ and $\mathcal{N}\left(0, 0.1^2\right)$ noise in the radial direction and extra dimensions. The d-dim `grid` dataset contains samples from mixture of d Gaussians. Each center lies on one dimension in the d-dimension hypercube. The `two moons` dataset is sampled from the exponential family with potential function as $\frac{1}{2}\left(\frac{\|x\|-2}{0.4}\right)^2 - \log\left(\exp\left(-\frac{1}{2}\left(\frac{x-2}{0.6}\right)^2\right) + \exp\left(-\frac{1}{2}\left(\frac{x+2}{0.6}\right)^2\right)\right)$. We use 500 samples for training, and for testing 1500 (`grid`) or 5000 (`ring, two moons`) samples, following Sutherland et al. (2017).

We visualize the samples generated by the learned sampler, and compare it with the training datasets in the first row in Figure 1. The learned f is also plotted in the second row in Figure 1. The DDE learned models generate samples that cover the training data, showing the ability of the DDE for estimating kernel exponential family on complicated data.

Then, we demonstrate the convergence of the DDE in Figure 2 on `rings` dataset. We initialize with a random dual distribution. As the algorithm iterates, the distribution converges to the target true distribution, justifying the convergence guarantees. More results on 2-dimensional `grid` and `two moons` can be found in Figure 3 in Appendix D. The DDE algorithm behaves similarly on these two datasets.

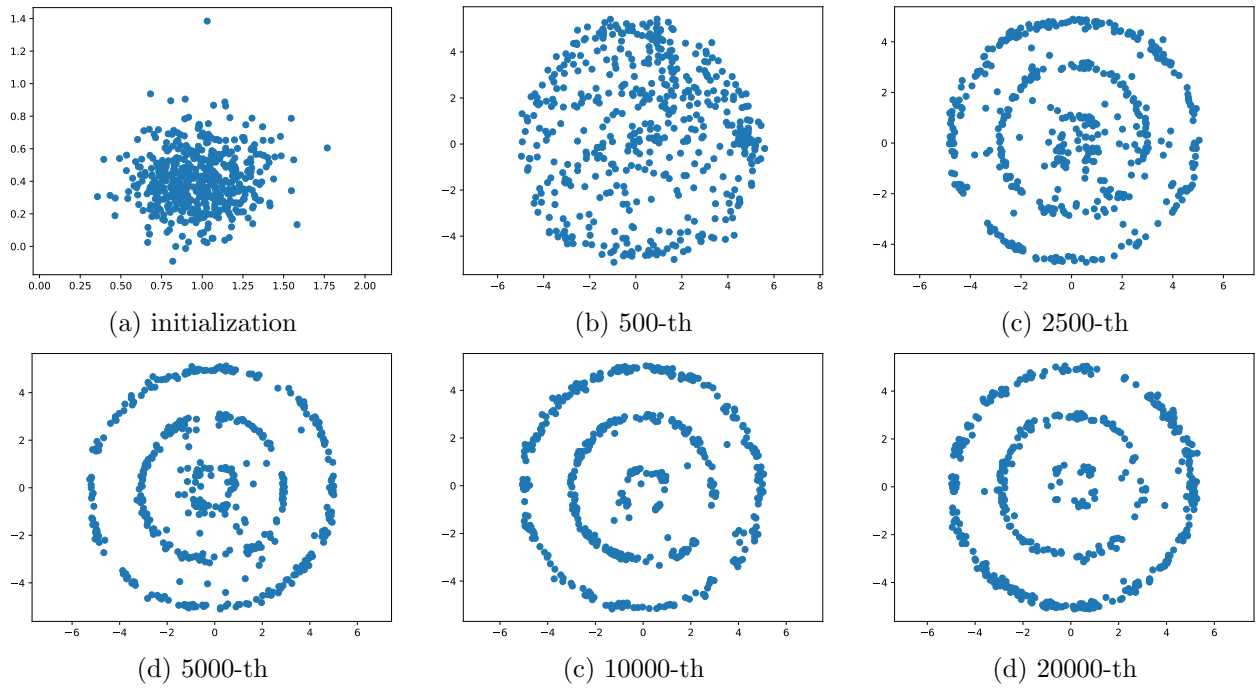
Table 1: Quantitative evaluation on synthetic data. MMD in $\times 10^{-3}$ scale on test set is reported.

	MMD		time (s)	
Datasets	DDE	KEF	DDE-sample	KEF+HMC
two moons	**0.11**	1.32	**0.07**	13.04
ring	**1.53**	3.19	**0.07**	14.12
grid	**1.32**	40.39	**0.04**	4.43

Finally, we compare the MMD between the generated samples from the learned model to the training data with the current state-of-the-art method for KEF (Sutherland et al., 2017), which performs better than KDE and other alternatives (Strathmann et al., 2015). We use HMC to generate the samples from the KEF learned model, while in the DDE, we can bypass the HMC step by the learned sampler. The computation time for inference is listed in Table 1. It shows the DDE improves the inference efficiency in orders. The MMD comparison is listed in Table 1. The proposed DDE estimator performs significantly better than the best performances by KEF in terms of MMD.

Conditional model extension. In this part of the experiment, models are trained to estimate the conditional distribution $p(y|x)$ on the benchmark datasets for studying these methods (Arbel and Gretton,

[1] https://github.com/karlnapf/nystrom-kexpfam
[2] https://github.com/MichaelArbel/KCEF



(a) initialization (b) 500-th (c) 2500-th

(d) 5000-th (c) 10000-th (d) 20000-th

Figure 2: The DDE estimators on `rings` dataset in each iteration. The blue points are sampled from the learned model. With the algorithm proceeds, the learned distribution converges to the ground-truth.

2017; Sugiyama et al., 2010). We centered and normalized the data and randomly split the datasets into a training and a testing set with equal size, as in Arbel and Gretton (2017). We evaluate the performances by the negative *log*-likelihood. Besides the score matching based KCEF, we also compared with LS-CDE and ϵ-KDE, introduced in (Sugiyama et al., 2010). The empirical results are summarized in Table 2.

Although these datasets are low-dimensional with few samples and the KCEF uses the anisotropic RBF kernel (*i.e.*, different bandwidth in each dimension, making the experiments preferable to the KCEF), the proposed DDE still outperforms the competitors on six datasets significantly, and achieves comparable performance on the rest, even though it uses a simple isotropic RBF kernel. This further demonstrates the statistical power of the proposed DDE, comparing to the score matching estimator.

7 Conclusion

In this paper, we exploit the *doubly dual embedding* to reformulate the penalized MLE to a novel saddle-point optimization, which bypasses the intractable integration and provides flexibility in parameterizing the dual distribution. The saddle point view reveals a unique understanding of GANs and leads to a practical algorithm, which achieves state-of-the-art performance. We also establish the statistical consistency and algorithm convergence guarantee for the proposed algorithm. Although the transport mapping parametrization is flexible enough, it requires extra optimization for the KL-divergence estimation. For the future work, we will exploit dynamic-based sampling methods to design new parametrization, which shares both flexibility and density tractability.

Acknowledgements

We thank Michael Arbel and the anonymous reviewers of AISTASTS 2019 for their insightful comments and suggestions. NH is supported in part by NSF-CRII-1755829, NSF-CMMI-1761699, and NCSA Faculty

Table 2: The negative log-likelihood comparison on benchmarks. Mean and std are calculated on 20 runs of different train/test splits. KCEF gets numerically unstable on two datasets, where we mark as N/A.

	DDE	KCEF	ϵ-KDE	LSCDE
geyser	**0.55 ± 0.07**	1.21 ± 0.04	1.11 ± 0.02	0.7 ± 0.01
caution	**0.95 ± 0.19**	**0.99 ± 0.01**	1.25 ± 0.19	1.19 ± 0.02
ftcollinssnow	**1.49 ± 0.14**	**1.46 ± 0.0**	1.53 ± 0.05	1.56 ± 0.01
highway	**1.18 ± 0.30**	**1.17 ± 0.01**	2.24 ± 0.64	1.98 ± 0.04
snowgeese	**0.42 ± 0.31**	**0.72 ± 0.02**	1.35 ± 0.17	1.39 ± 0.05
GAGurine	**0.43 ± 0.15**	**0.46 ± 0.0**	0.92 ± 0.05	0.7 ± 0.01
topo	1.02 ± 0.31	**0.67 ± 0.01**	1.18 ± 0.09	0.83 ± 0.0
CobarOre	**1.45 ± 0.23**	3.42 ± 0.03	**1.65 ± 0.09**	**1.61 ± 0.02**
mcycle	**0.60 ± 0.15**	**0.56 ± 0.01**	1.25 ± 0.23	0.93 ± 0.01
BigMac2003	**0.47 ± 0.36**	**0.59 ± 0.01**	1.29 ± 0.14	1.63 ± 0.03
cpus	**-0.63 ± 0.77**	N/A	1.01 ± 0.10	1.04 ± 0.07
crabs	**-0.60 ± 0.26**	N/A	0.99 ± 0.09	-0.07 ± 0.11
birthwt	**1.22 ± 0.15**	**1.18 ± 0.13**	1.48 ± 0.01	1.43 ± 0.01
gilgais	**0.61 ± 0.10**	**0.65 ± 0.08**	1.35 ± 0.03	0.73 ± 0.05
UN3	**1.03 ± 0.09**	1.15 ± 0.21	1.78 ± 0.14	1.42 ± 0.12
ufc	**1.03 ± 0.10**	**0.96 ± 0.14**	1.40 ± 0.02	**1.03 ± 0.01**

Fellowship.

References

Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *COLT*, pages 139–153, 2006.

Michael Arbel and Arthur Gretton. Kernel conditional exponential family. In *AISTATS*, pages 1337–1346, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. In *ICML*, 2017.

Francis R. Bach. On the equivalence between quadrature rules and random features. *Journal of Machine Learning Research*, 18:714–751, 2017.

Andrew R Barron and Chyong-Hwa Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, pages 1347–1369, 1991.

Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018.

Lawrence D. Brown. *Fundamentals of Statistical Exponential Families*, volume 9 of *Lecture notes-monograph series*. Institute of Mathematical Statistics, Hayward, Calif, 1986.

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

Bo Dai, Bo Xie, Niao He, Yingyu Liang, Anant Raj, Maria-Florina Balcan, and Le Song. Scalable kernel methods via doubly stochastic gradients. In *NeurIPS*, 2014.

Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual embeddings. In *AISTATS*, pages 1458–1467, 2017.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *ICML*, pages 1133–1142, 2018.

A. Devinatz. Integral representation of pd functions. *Trans. AMS*, 74(1):56–77, 1953.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017.

Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, 8:1217–1260, 2007.

Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*, volume 28. SIAM 1999.

Kenji Fukumizu. *Exponential manifold by reproducing kernel Hilbert spaces*, page 291306. Cambridge University Press, 2009.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014.

Chong Gu and Chunfu Qiu. Smoothing spline density estimation: Theory. *The Annals of Statistics*, pages 217–234, 1993.

Matthias Hein and Olivier Bousquet. Kernels, associated structures, and generalizations. Technical Report 127, Max Planck Institute for Biological Cybernetics, 2004.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

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

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *NeurIPS*, pages 4743–4751, 2016.

Jyrki Kivinen, Alex Smola, and Robert C. Williamson. Online learning with kernels. *IEEE Transactions on Signal Processing*, 52(8), Aug 2004.

Hermann König. *Eigenvalue Distribution of Compact Operators*. Birkhäuser, Basel, 1986.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabas Poczos. MMD GAN: Towards deeper understanding of moment matching network. In *NeurIPS*, pages 2203–2213, 2017.

Charles Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17(1): 177–204, 2005.

Radford Neal. Annealed Importance Sampling. *Statistics and computing*, 11(2):125–139, 2001.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization*, 19(4):1574–1609, January 2009. ISSN 1052-6234.

XuanLong Nguyen, Martin Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *NeurIPS*, pages 1089–1096, 2008.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *NeurIPS*, 2016.

Giovanni Pistone and Maria Piera Rogantin. The exponential statistical manifold: mean parameters, orthogonality and space transformations. *Bernoulli*, 5(4):721–760, 1999.

Ali Rahimi and Ben Recht. Random features for large-scale kernel machines. In *NeurIPS*, 2008.

Ali Rahimi and Ben Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *NeurIPS*, 2009.

Danilo J Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286, 2014.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, 2015.

R. T. Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematics Series*. Princeton University Press, Princeton, NJ, 1970.

Maurice Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *The Journal of Machine Learning Research*, 18(1): 1830–1888, 2017.

Heiko Strathmann, Dino Sejdinovic, Samuel Livingstone, Zoltan Szabo, and Arthur Gretton. Gradient-free hamiltonian monte carlo with efficient kernel exponential families. In *NeurIPS*, pages 955–963, 2015.

Masashi Sugiyama, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. Conditional density estimation via least-squares density ratio estimation. In *AISTATS*, pages 781–788, 2010.

Dougal J Sutherland, Heiko Strathmann, Michael Arbel, and Arthur Gretton. Efficient and principled score estimation with nyström kernel exponential families. In *AISTATS*, 2018.

Shankar Vembu, Thomas Gärtner, and Mario Boley. Probabilistic structured predictors. In *UAI*, pages 557–564, 2009.

Appendix

A Proof Details

A.1 Proof of Theorem 3

Theorem 3 (Dual gradient) *Denoted as* $(f^*, \nu^*) = \operatorname{argmax}_{(f,\nu)\in\mathcal{H}} \tilde{\ell}(f, \nu, w_g)$ *and* $\widehat{L}(w_g) = \tilde{\ell}(f^*, \nu^*, w_g)$, *we have*

$$\nabla_{w_g} \widehat{L}(w_g) = -\mathbb{E}_\xi \left[\nabla_{w_g} f^* \left(g_{w_g}(\xi)\right)\right] + \frac{1}{\lambda}\mathbb{E}_\xi \left[\nabla_{w_g} \nu^* \left(g_{w_g}(\xi)\right)\right].$$

Proof The conclusion can be proved by chain rule and the optimality conditions.

Specifically, notice that the $\left(f^*_{w_g}, \nu^*_{w_g}\right)$ are implicit functions of w_g, we can calculate the gradient of $\widehat{L}(w_g)$ w.r.t. w_g

$$\begin{aligned}
\nabla_{w_g} \widehat{L}(w_g) &= \widehat{\mathbb{E}}_{\mathcal{D}} \left[\nabla_f f^*_{w_g} \nabla_{w_g} f^*_{w_g}\right] - \mathbb{E}_\xi \left[\nabla_g f\left(g(\xi)\right) \nabla_{w_g} g\right] - \mathbb{E}_q \left[\nabla_f f^*_{w_g} \nabla_{w_g} f^*_{w_g}\right] - \frac{\eta}{2}\nabla_f \left\|f^*_{w_g}\right\|^2_{\mathcal{H}} \nabla_{w_g} f^*_{w_g} \\
&\quad + \frac{1}{\lambda}\left(\mathbb{E}_\xi \left[\nabla_g \nu^*_{w_g}\left(g(\xi)\right) \nabla_{w_g} g\right] + \mathbb{E}_q \left[\nabla_\nu \nu^*_{w_g} \nabla_{w_g} \nu^*_{w_g}\right] - \mathbb{E}_{p_0}\left[\exp\left(\nu^*_{w_g}\right) \nabla_\nu \nu^*_{w_g} \nabla_{w_g} \nu^*_{w_g}\right]\right) \\
&= \underbrace{\left(\widehat{\mathbb{E}}_{\mathcal{D}} \left[\nabla_f f^*_{w_g}\right] - \mathbb{E}_q \left[\nabla_f f^*_{w_g}\right] - \frac{\eta}{2}\nabla_f \left\|f^*_{w_g}\right\|^2_{\mathcal{H}}\right)}_{0} \nabla_{w_g} f^*_{w_g} - \mathbb{E}_\xi \left[\nabla_g f\left(g(\xi)\right) \nabla_{w_g} g\right] \\
&\quad + \frac{1}{\lambda}\mathbb{E}_\xi \left[\nabla_g \nu^*_{w_g}\left(g(\xi)\right) \nabla_{w_g} g\right] + \frac{1}{\lambda}\underbrace{\left(\mathbb{E}_q \left[\nabla_\nu \nu^*_{w_g}\right] - \mathbb{E}_{p_0}\left[\exp\left(\nu^*_{w_g}\right) \nabla_\nu \nu^*_{w_g}\right]\right)}_{0} \nabla_{w_g} \nu^*_{w_g} \\
&= -\mathbb{E}_\xi \left[\nabla_{w_g} f^*\left(g(\xi)\right)\right] + \frac{1}{\lambda}\mathbb{E}_\xi \left[\nabla_{w_g} \nu^*\left(g_{w_g}(\xi)\right)\right],
\end{aligned}$$

where the second equations come from the fact $\left(f^*_{w_g}, \nu^*_{w_g}\right)$ are optimal and $\left(\nabla_{w_g} f^*_{w_g}, \nabla_{w_g} \nu^*_{w_g}\right)$ are not functions of (x, ξ, x').

■

A.2 Proof of Theorem 4

The proof of Theorem 4 mainly follows the technique in Gu and Qiu (1993) with extra consideration of the approximation error from the dual embedding.

We first define some notations that will be used in the proof. We denote $\langle f, g\rangle_p = \int_\Omega f(x)\, g(x)\, p(x)\, dx$, which induces the norm denoted as $\|\cdot\|_p^2$. We introduce $\tilde{h}$ as the maximizer to $\tilde{L}(h)$ defined as

$$\tilde{L}(h) := \widehat{\mathbb{E}}_{\mathcal{D}}[h] - \mathbb{E}_{p^*}[h] - \frac{1}{2}\|h - f^*\|^2_{p^*} - \frac{\eta}{2}\|h\|^2_{\mathcal{H}}.$$

The proof relies on decomposing the error into two parts: **i)** the error between $\tilde{h}$ and f^*; and **ii)** the error between $\tilde{f}$ and $\tilde{h}$.

By Mercer decomposition (König, 1986), we can expand $k(\cdot, \cdot)$ as

$$k(x, x') = \sum_{l=1}^{\infty} \zeta_l \psi_l(x)\, \psi_l(x'),$$

With the eigen-decomposition, we can rewrite function $f \in \mathcal{H}$ as $f(\cdot) = \sum_{l=1}^{\infty} \langle f, \psi_l \rangle_{p^*} \psi(\cdot)$. Then, we have $\|f\|_{\mathcal{H}}^2 = \sum_{l=1}^{\infty} \zeta_l^{-1} \langle f, \psi_l \rangle_{p^*}^2$ and $\|f\|_{p^*}^2 = \sum_{l=1}^{\infty} \langle f, \psi_l \rangle_{p^*}^2$.

We make the following standard assumptions:

Assumption 1 *There exists $\kappa > 0$ such that $k(x, x') \leqslant \kappa$, $\forall x, x' \in \Omega$.*

Assumption 2 *The eigenvalues of the kernel $k(\cdot, \cdot)$ decay sufficiently homogeneously with rate r,* i.e.*, $\zeta_l = \mathcal{O}(l^{-r})$ where $r > 1$.*

Assumption 3 *There exists a distribution p_0 on the support Ω which is uniformly upper and lower bounded.*

Lemma 6 *Under Assumption 3, $\forall f \in \mathcal{H}_k$, $p_f(x) = \frac{\exp(f(x) - \log p_0(x))}{\int_\Omega \exp(f(x) - \log p_0(x)) p_0(x)} p_0(x)$, we have $2\exp(-\kappa C_{\mathcal{H}} - C_0) \leqslant p^*(x) \leqslant 2\exp(\kappa C_{\mathcal{H}} + C_0)$.*

Proof For $\forall f \in \mathcal{H}_k$ and the Assumption 3, $p_f(x) = \frac{\exp(f(x) - \log p_0(x))}{\int_\Omega \exp(f(x) - \log p_0(x)) p_0(x)} p_0(x)$ with $f \in \mathcal{H}_k$ and $\|f\|_{\mathcal{H}} \leqslant C_{\mathcal{H}}$ and $\|\log p_0(x)\|_\infty \leqslant C_0$, implies $2\exp(-\kappa C_{\mathcal{H}} - C_0) \leqslant p^*(x) \leqslant 2\exp(\kappa C_{\mathcal{H}} + C_0)$. ■

Therefore, it is reasonable to consider the parametrization of the dual distribution:

Assumption 4 *The parametric family of dual distribution in* (18) *is bounded above from infinity,* i.e.*, $\forall q(x) \in \mathcal{P}_w$, $q(x) \leqslant C_{\mathcal{P}_w} < \infty$, $\forall x \in \Omega$.*

To prove the Theorem 4, we first show the error between $\tilde{h}$ and f^* under Assumption 1, 2, 3, and 4.

Lemma 7 *Under Assumption 2, we have*

$$\mathbb{E}\left[\left\|\tilde{h} - f^*\right\|_{p^*}^2\right] = \mathcal{O}\left(n^{-1}\eta^{-\frac{1}{r}} + \eta\right),$$
$$\eta\mathbb{E}\left[\left\|\tilde{h} - f^*\right\|_{\mathcal{H}}^2\right] = \mathcal{O}\left(n^{-1}\eta^{-\frac{1}{r}} + \eta\right).$$

Proof Denote the $\tilde{h}(\cdot) = \sum_l^{\infty} \underbrace{\langle \tilde{h}, \psi_l \rangle_{p^*}}_{\tilde{h}_l} \psi_l(\cdot)$ and $f^*(\cdot) = \sum_l^{\infty} \underbrace{\langle f^*, \psi_l \rangle_{p^*}}_{f_l^*} \psi_l(\cdot)$, then, we can rewrite the $\tilde{L}(h)$ as

$$\tilde{L}(h) = \sum_{l=1}^{\infty} h_l \left[\widehat{\mathbb{E}}[\psi_l(x)] - \mathbb{E}_{p^*}[\psi_l(x)]\right] - \frac{1}{2}\sum_{l=1}^{\infty}(h_l - f_l^*)^2 - \frac{\eta}{2}\sum_{l=1}^{\infty}\zeta_l^{-1}h_l^2.$$

Setting the derivative of $\tilde{L}(h)$ w.r.t. $[h_l]$ equal to zero, we obtain the representation of $\tilde{h}_l$ as

$$\tilde{h}_l = \frac{f_l^* + \alpha_l}{1 + \eta\zeta_l^{-1}},$$

where $\alpha_l = \widehat{\mathbb{E}}[\psi_l(x)] - \mathbb{E}_{p^*}[\psi_l(x)]$. Then, we have

$$\left\|\tilde{h} - f^*\right\|_{p^*}^2 = \sum_{l=1}^{\infty}\left(\tilde{h}_l - f_l^*\right)^2 = \sum_{l=1}^{\infty}\frac{\alpha_l^2 - 2\alpha_l\eta\zeta_l^{-1}f_l^* + \eta^2\zeta_l^{-2}(f_l^*)^2}{\left(1 + \eta\zeta_l^{-1}\right)^2},$$
$$\eta\left\|\tilde{h} - f^*\right\|_{\mathcal{H}}^2 = \eta\sum_{l=1}^{\infty}\zeta_l^{-1}\left(\tilde{h}_l - f_l^*\right)^2 = \sum_{l=1}^{\infty}\eta\zeta_l^{-1}\frac{\alpha_l^2 - 2\alpha_l\eta\zeta_l^{-1}f_l^* + \eta^2\zeta_l^{-2}(f_l^*)^2}{\left(1 + \eta\zeta_l^{-1}\right)^2}.$$

Recall that $\mathbb{E}[a_l] = 0$ and $\mathbb{E}\left[a_l^2\right] = \frac{1}{n}$, then we have

$$\mathbb{E}\left[\left\|\tilde{h} - f^*\right\|_{p^*}^2\right] = \frac{1}{n}\sum_{l=1}^{\infty}\frac{1}{\left(1+\eta\zeta_l^{-1}\right)^2} + \eta\sum_{l=1}^{\infty}\frac{\eta\zeta_l^{-1}}{\left(1+\eta\zeta_l^{-1}\right)^2}\cdot\zeta_t^{-1}\left(f_l^*\right)^2, \tag{23}$$

$$\mathbb{E}\left[\eta\left\|\tilde{h} - f^*\right\|_{\mathcal{H}}^2\right] = \frac{1}{n}\sum_{l=1}^{\infty}\frac{\eta\zeta_l^{-1}}{\left(1+\eta\zeta_l^{-1}\right)^2} + \eta\sum_{l=1}^{\infty}\frac{\eta^2\zeta_l^{-2}}{\left(1+\eta\zeta_l^{-1}\right)^2}\cdot\zeta_l^{-1}\left(f_l^*\right)^2. \tag{24}$$

By calculation, we obtain that

$$\begin{aligned}\sum_{l=1}^{\infty}\frac{\eta\zeta_l^{-1}}{\left(1+\eta\zeta_l^{-1}\right)^2} &= \sum_{l<\eta^{-\frac{1}{r}}}\frac{\eta\zeta_l^{-1}}{\left(1+\eta\zeta_l^{-1}\right)^2} + \sum_{l\geqslant\eta^{-\frac{1}{r}}}\frac{\eta\zeta_l^{-1}}{\left(1+\eta\zeta_l^{-1}\right)^2}\\ &= \mathcal{O}\left(\eta^{-\frac{1}{r}}\right) + \mathcal{O}\left(\int_{\eta^{-\frac{1}{r}}}^{\infty}\frac{\eta t^r}{\left(1+\eta t^r\right)^2}dt\right)\\ &= \mathcal{O}\left(\eta^{-\frac{1}{r}}\right) + \eta^{-\frac{1}{r}}\mathcal{O}\left(\int_1^{\infty}\frac{t^r}{\left(1+t^r\right)^2}dt\right) = \mathcal{O}\left(\eta^{-\frac{1}{r}}\right).\end{aligned} \tag{25}$$

Similarly, we can achieve

$$\sum_{l=1}^{\infty}\frac{1}{\left(1+\eta\zeta_l^{-1}\right)^2} = \mathcal{O}\left(\eta^{-\frac{1}{r}}\right), \tag{26}$$

$$\sum_{l=1}^{\infty}\frac{1}{1+\eta\zeta_l^{-1}} = \mathcal{O}\left(\eta^{-\frac{1}{r}}\right). \tag{27}$$

Note also that $\sum_{l=1}^{\infty}\zeta_t^{-1}(f_l^*)^2 = \|f^*\|_{\mathcal{H}}^2 < \infty$. Hence, the second term in (23) is also finite. Plugging (25) and (26) into (23), we achieve the conclusion. ■

Next, we characterize the approximation error due to parametrization in L_2 norm,

Lemma 8 *Under Assumption 4, $\forall q \in \mathcal{P}_w$ and $f \in \mathcal{H}_k$, we have $\|p_f - q\|_2^2 \leqslant 4\exp\left(2\kappa C_{\mathcal{H}} + 2C_0\right) C_{\mathcal{P}_w} KL\left(q||p_f\right)$.*

Proof By Lemma 6, we have $t = \frac{q}{p_f} \leqslant 2\exp\left(\kappa C_{\mathcal{H}} + C_0\right) C_{\mathcal{P}_w} < \infty$. Denote $\Phi(t) = t\log t$, we have $\Phi''(t) = \frac{1}{t} \geqslant C_\Phi$ with $C_\Phi := \frac{1}{2\exp(\kappa C_{\mathcal{H}}+C_0)C_{\mathcal{P}_w}}$, and thus, $\Phi(t) - C_\Phi t^2$ is convex. Therefore, applying Jensen's inequality, we have

$$\begin{aligned}&\Phi\left(\mathbb{E}_{p_f}\left[\frac{q}{p_f}\right]\right) - C_\Phi\left(\mathbb{E}_{p_f}\left[\frac{q}{p_f}\right]\right)^2 \leqslant \mathbb{E}_{p_f}\left[\Phi\left(\frac{q}{p_f}\right) - C_\Phi\frac{q^2}{p_f^2}\right]\\ \Rightarrow\quad & C_\Phi\left(\int\frac{q^2(x)}{p_f(x)}dx - \left(\mathbb{E}_{p_f}\left[\frac{q}{p_f}\right]\right)^2\right) \leqslant \mathbb{E}_{p_f}\left[\Phi\left(\frac{q}{p_f}\right)\right] - \Phi\left(\mathbb{E}_{p_f}\left[\frac{q}{p_f}\right]\right)\\ \Rightarrow\quad & C_\Phi\underbrace{\left(\int\frac{q^2(x)}{p_f(x)}dx - 1\right)}_{\chi^2(q,p_f)} \leqslant KL\left(q||p_f\right) - \Phi(1) = KL\left(q||p_f\right).\end{aligned}$$

On the other hand,

$$\chi^2\left(q, p_f\right) = \int\frac{\left(q(x) - p_f(x)\right)^2}{p_f(x)}dx \geqslant 2\exp\left(-\kappa C_{\mathcal{H}} - C_0\right)\|q - p_f\|_2^2,$$

which leads to the conclusion.

■

We proceed the other part of the error, *i.e.*, between $\tilde{f}$ and $\tilde{h}$.

Lemma 9 *Under Assumption 1 and Assumption 2, we have as $n \to \infty$ and $\eta \to 0$,*

$$\begin{aligned}\left\|\tilde{f}-\tilde{h}\right\|_{p^*}^2 &= o_{p^*}\left(n^{-1}\eta^{-\frac{1}{r}}+\eta\right)+C\epsilon_{approx},\\ \eta\left\|\tilde{f}-\tilde{h}\right\|_{\mathcal{H}}^2 &= o_{p^*}\left(\left(n^{-1}\eta^{-\frac{1}{r}}+\eta\right)+\epsilon_{approx}\right)+o_{p^*}\left(n^{-1}\eta^{-\frac{1}{r}}+\eta\right)+C\epsilon_{approx}.\end{aligned}$$

Therefore,

$$\begin{aligned}\left\|\tilde{f}-f^*\right\|_{p^*}^2 &= \mathcal{O}_{p^*}\left(n^{-1}\eta^{-\frac{1}{r}}+\eta\right)+C\epsilon_{approx},\\ \eta\left\|\tilde{f}-f^*\right\|_{\mathcal{H}}^2 &= \mathcal{O}_{p^*}\left(n^{-1}\eta^{-\frac{1}{r}}+\eta\right)+o_{p^*}\left(\left(n^{-1}\eta^{-\frac{1}{r}}+\eta\right)+\epsilon_{approx}\right)+C\epsilon_{approx}.\end{aligned}$$

Proof Since $\left(\tilde{f},\tilde{\nu},\tilde{q}\right)$ are the optimal solutions to the primal-dual reformulation of the penalized MLE (18), we have the first-order optimality condition: $\nabla_f \ell(\tilde{f},\tilde{\nu},\tilde{q})=0$, which implies $\widehat{\mathbb{E}}\left[k\left(x,\cdot\right)\right]-\mathbb{E}_{\tilde{q}}\left[k\left(x,\cdot\right)\right]-\eta\tilde{f}=0$. Hence,

$$\widehat{\mathbb{E}}\left[\langle k\left(x,\cdot\right),\tilde{f}-\tilde{h}\rangle\right]-\mathbb{E}_{\tilde{q}}\left[\langle k\left(x,\cdot\right),\tilde{f}-\tilde{h}\rangle\right]-\eta\langle\tilde{f},\tilde{f}-\tilde{h}\rangle_{\mathcal{H}}=0. \tag{28}$$

Similarly, by the optimality of $\tilde{h}$ w.r.t. $\tilde{L}\left(h\right)$, we have

$$\widehat{\mathbb{E}}\left[\langle k\left(x,\cdot\right),\tilde{f}-\tilde{h}\rangle\right]-\mathbb{E}_{p^*}\left[\langle k\left(x,\cdot\right),\tilde{f}-\tilde{h}\rangle\right]-\langle\tilde{f}-\tilde{h},\tilde{h}-f^*\rangle_{p^*}-\eta\langle\tilde{h},\tilde{f}-\tilde{h}\rangle_{\mathcal{H}}=0. \tag{29}$$

Combining the (28) and (29), we further obtain

$$\begin{aligned}&\mathbb{E}_{\tilde{q}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]-\mathbb{E}_{p_{\tilde{h}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]+\eta\left\|\tilde{f}-\tilde{h}\right\|_{\mathcal{H}}^2 \\ &=\langle\tilde{f}-\tilde{h},\tilde{h}-f^*\rangle_{p^*}+\mathbb{E}_{p^*}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]-\mathbb{E}_{p_{\tilde{h}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right] \\ \Rightarrow\quad &\mathbb{E}_{p_{\tilde{f}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]-\mathbb{E}_{p_{\tilde{h}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]+\eta\left\|\tilde{f}-\tilde{h}\right\|_{\mathcal{H}}^2 \\ &=\langle\tilde{f}-\tilde{h},\tilde{h}-f^*\rangle_{p^*}+\underbrace{\mathbb{E}_{p^*}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]-\mathbb{E}_{p_{\tilde{h}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]}_{\epsilon_1} \\ &+\underbrace{\mathbb{E}_{p_{\tilde{f}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]-\mathbb{E}_{\tilde{q}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]}_{\epsilon_2}.\end{aligned} \tag{30}$$

For ϵ_1, denote $F\left(\theta\right)=\mathbb{E}_{p_{f^*+\theta(\tilde{h}-f^*)/\varsigma}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]-\mathbb{E}_{p^*}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]$ with $\varsigma=\left\|f^*-\tilde{h}\right\|_{p^*}=o_{p^*}\left(1\right)$, then, apply Taylor expansion to $F\left(\theta\right)$ will lead to

$$F\left(\theta\right)=\frac{\theta}{\varsigma}\left(1+o_p(1)\right)\langle\tilde{f}-\tilde{h},\tilde{h}-f^*\rangle_{p^*} \tag{31}$$

where $o_{p^*}\left(1\right)$ w.r.t. $\theta\to 0$. Therefore,

$$\mathbb{E}_{p_{\tilde{h}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]-\mathbb{E}_{p^*}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]=F\left(\varsigma\right)=\left(1+o_p(1)\right)\langle\tilde{f}-\tilde{h},\tilde{h}-f^*\rangle_{p^*}, \tag{32}$$

as $\eta \to 0$ and $n\eta^{\frac{1}{r}} \to \infty$.

For ϵ_2, by Hölder inequality,

$$\begin{aligned}\epsilon_2 &= \mathbb{E}_{p_{\tilde{f}}}\left[\tilde{f}(x)-\tilde{h}(x)\right]-\mathbb{E}_{\tilde{q}}\left[\tilde{f}(x)-\tilde{h}(x)\right]=\int_{\Omega}\frac{p_{\tilde{f}}(x)-\tilde{q}(x)}{p^*(x)}\left(\tilde{f}(x)-\tilde{h}(x)\right)p^*(x)\,dx \\ &\leqslant \left\|\tilde{f}-\tilde{h}\right\|_{p^*}\left\|\frac{p_{\tilde{f}}(x)-\tilde{q}(x)}{p^*(x)}\right\|_{p^*}.\end{aligned}$$

By Lemma 6, we have $p^*(x) = \exp(f^* - A(f^*)) = \frac{\exp(f^*(x)-\log p_0(x))}{\int_\Omega \exp(f^*(x)-\log p_0(x))p_0(x)}p_0(x)$ with $f^* \in \mathcal{H}_k$ and $\|f^*\|_{\mathcal{H}} \leqslant C_{f^*}$ implying $2\exp(-\kappa C_{f^*} - C_0) \leqslant p^*(x) \leqslant 2\exp(\kappa C_{f^*} + C_0)$. Therefore, we have

$$\epsilon_2 \leqslant 2\exp(\kappa C_{f^*} + C_0)\left\|p_{\tilde{f}} - \tilde{q}\right\|_2\left\|\tilde{f}-\tilde{h}\right\|_{p^*}. \tag{33}$$

Given $\forall f \in \mathcal{F}$ fixed, we have

$$\min_{q\in\mathcal{P}}\max_{\nu\in\mathcal{F}} -\mathbb{E}_q[f] + \frac{1}{\lambda}\left(\mathbb{E}_q[\nu] - \mathbb{E}_{p_0}[\exp(\nu)]\right) + 1 = \min_{q\in\mathcal{P}} -\mathbb{E}_q[f] + \frac{1}{\lambda}KL(q||p_0) = \min_{q\in\mathcal{P}}\frac{1}{\lambda}KL(q||p_f). \tag{34}$$

In ideal case where the $\mathcal{P}$ and the domain of ν is flexible enough, we have $KL(q||p_f) = 0$. However, due to parametrization, we introduce extra approximate error, denoting as $\epsilon_{approx} := \sup_{f\in\mathcal{H}_k}\inf_{q\in\mathcal{P}_w} KL(q||p_f)$. As we can see, as the parametrized family of q and ν becomes more and more flexible, the $\epsilon_{approx} \to 0$.

Therefore, by Lemma 8, we obtain

$$\epsilon_2 \leqslant 8C_{\mathcal{P}_w}\exp(3\kappa C_{\mathcal{H}} + 3C_0)\epsilon_{approx}^{\frac{1}{2}}\left\|\tilde{f}-\tilde{h}\right\|_{p^*}.$$

On the other hand, we define $D(\theta) = \mathbb{E}_{p_{\tilde{h}+\theta(\tilde{f}-\tilde{h})}}\left[\tilde{f}-\tilde{h}\right]$, notice that $D'(\theta) = \left\|\tilde{f}-\tilde{h}\right\|^2_{p_{\tilde{h}+\theta(\tilde{f}-\tilde{h})}}$, by the mean value theorem, we can obtain that

$$\mathbb{E}_{p_{\tilde{f}}}\left[\tilde{f}(x)-\tilde{h}(x)\right]-\mathbb{E}_{p_{\tilde{h}}}\left[\tilde{f}(x)-\tilde{h}(x)\right] = D(1)-D(0) = D'(\theta) = \left\|\tilde{f}-\tilde{h}\right\|^2_{p_{\tilde{h}+\theta(\tilde{f}-\tilde{h})}} \tag{35}$$

with $\theta \in [0,1]$. Gu and Qiu (1993) shows that when $\forall f \in \mathcal{H}_k$ is uniformly bounded, then, $c\|\cdot\|_{p^*} \leqslant \|\cdot\|_{p_{\tilde{h}+\theta(\tilde{f}-\tilde{h})}}$, $\theta \in [0,1]$, which is the true under the Assumption 1.

Plugging (32) and (33) into (30), we achieve

$$c\left\|\tilde{f}-\tilde{h}\right\|^2_{p^*} + \eta\left\|\tilde{f}-\tilde{h}\right\|^2_{\mathcal{H}} \leqslant o_p\left(\langle\tilde{f}-\tilde{h}, \tilde{h}-f^*\rangle_{p^*}\right) + 8C_{\mathcal{P}_w}\exp(3\kappa C_{\mathcal{H}} + 3C_0)\epsilon_{approx}^{\frac{1}{2}}\left\|\tilde{f}-\tilde{h}\right\|_{p^*},$$

which leads to the first part in the conclusion. Combining with the Lemma 7, we obtain the second part of the conclusion.

■

Now, we are ready for proving the main theorem about the statistical consistency.

Theorem 4 *Assume the spectrum of kernel $k(\cdot,\cdot)$ decays sufficiently homogeneously in rate l^{-r}. With some other mild assumptions listed in Appendix A.2, we have as $\eta \to 0$ and $n\eta^{\frac{1}{r}} \to \infty$,*

$$KL\left(p^*||p_{\tilde{f}}\right) + KL\left(p_{\tilde{f}}||p^*\right) = \mathcal{O}_{p^*}\left(n^{-1}\eta^{-\frac{1}{r}} + \eta + \epsilon_{approx}\right),$$

where $\epsilon_{approx} := \sup_{f\in\mathcal{F}}\inf_{q\in\mathcal{P}_w} KL(q||p_f)$ denotes the approximate error due to the parametrization of $\tilde{q}$ and $\tilde{\nu}$. Therefore, when setting $\eta = \mathcal{O}\left(n^{-\frac{r}{1+r}}\right)$, $p_{\tilde{f}}$ converge s to p^* in terms of Jensen-Shannon divergence at rate $\mathcal{O}_{p^*}\left(n^{-\frac{r}{1+r}} + \epsilon_{approx}\right)$.

Proof Recall the $\left(\tilde{f}, \tilde{q}\right)$ is the optimal solution to (9), we have the first-order optimality condition as

$$\widehat{\mathbb{E}}\left[\langle k\left(x,\cdot\right), \tilde{f} - f^*\rangle\right] - \mathbb{E}_{\tilde{q}}\left[\langle k\left(x,\cdot\right), \tilde{f} - f^*\rangle\right] - \eta\langle \tilde{f}, \tilde{f} - f^*\rangle_{\mathcal{H}} = 0, \tag{36}$$

which leads to

$$\begin{aligned}\mathbb{E}_{p_{\tilde{f}}}\left[\langle k\left(x,\cdot\right), \tilde{f} - f^*\rangle\right] &= \widehat{\mathbb{E}}\left[\langle k\left(x,\cdot\right), \tilde{f} - f^*\rangle\right] - \eta\langle \tilde{f}, \tilde{f} - f^*\rangle_{\mathcal{H}} \\ &\quad + \underbrace{\mathbb{E}_{p_{\tilde{f}}}\left[\langle k\left(x,\cdot\right), \tilde{f} - f^*\rangle\right] - \mathbb{E}_{\tilde{q}}\left[\langle k\left(x,\cdot\right), \tilde{f} - f^*\rangle\right]}_{\epsilon_3}.\end{aligned}$$

Then, we can rewrite the Jensen-Shannon divergence

$$\begin{aligned}KL\left(p^*||p_{\tilde{f}}\right) + KL\left(p_{\tilde{f}}||p^*\right) &= \mathbb{E}_{p_{\tilde{f}}}\left[\langle k\left(x,\cdot\right), \tilde{f} - f^*\rangle\right] - \mathbb{E}_{p^*}\left[\langle k\left(x,\cdot\right), \tilde{f} - f^*\rangle\right] \\ &= \epsilon_3 + \eta\langle \tilde{f}, f^* - \tilde{f}\rangle_{\mathcal{H}} + \widehat{\mathbb{E}}\left[\langle k\left(x,\cdot\right), \tilde{f} - f^*\rangle\right] - \mathbb{E}_{p^*}\left[\langle k\left(x,\cdot\right), \tilde{f} - f^*\rangle\right]\end{aligned}$$

Similar to the bound of ϵ_2, we have

$$\epsilon_3 \leqslant 8C_{\mathcal{P}_w}\exp\left(3\kappa C_{\mathcal{H}} + 3C_0\right)\epsilon_{approx}^{\frac{1}{2}}\left\|\tilde{f} - f^*\right\|_{p^*} = \mathcal{O}\left(\epsilon_{approx}^{\frac{1}{2}}\sqrt{n^{-1}\eta^{-\frac{1}{r}} + \eta}\right) = \mathcal{O}\left(\epsilon_{approx} + \left(n^{-1}\eta^{-\frac{1}{r}} + \eta\right)\right).$$

Moreover, with Cauchy-Schwarz inequality,

$$\eta\langle \tilde{f}, f^* - \tilde{f}\rangle_{\mathcal{H}} \leqslant \eta\left\|\tilde{f}\right\|_{\mathcal{H}}\left\|\tilde{f} - f^*\right\|_{\mathcal{H}},$$

$$\eta\left\|\tilde{f}\right\|_{\mathcal{H}} \leqslant 2\eta\left\|f^*\right\|_{\mathcal{H}} + 2\eta\left\|\tilde{f} - f^*\right\|_{\mathcal{H}}$$

Applying the conclusion in Lemma 9 and the fact that $\|f^*\|_{\mathcal{H}} \leqslant C_{f^*}$, we obtain that

$$\eta\langle \tilde{f}, f^* - \tilde{f}\rangle_{\mathcal{H}} = \mathcal{O}(\eta).$$

Finally, for the term

$$\widehat{\mathbb{E}}\left[\langle k\left(x,\cdot\right), \tilde{f} - f^*\rangle\right] - \mathbb{E}_{p^*}\left[\langle k\left(x,\cdot\right), \tilde{f} - f^*\rangle\right],$$

we rewrite $\tilde{f}$ and f^* in the form of ψ as $\sum_{l=1}^{\infty}\left(\tilde{f}_l - f_l^*\right)\alpha_l$. Then, apply Cauchy-Schwarz inequality,

$$\sum_{l=1}^{\infty}\left|\left(\tilde{f}_l - f_l^*\right)\alpha_l\right| \leqslant \left(\sum_{l=1}^{\infty} a_l^2\left(\tilde{f}_l - f_l^*\right)^2\right)^{\frac{1}{2}}\left(\sum_{l=1}^{\infty}\left(\frac{\alpha_l}{a_l}\right)^2\right)^{\frac{1}{2}}$$

where $a_l^2 = 1 + \eta\zeta_l^{-1}$. Then,

$$\sum_{l=1}^{\infty} a_l^2\left(\tilde{f}_l - f_l^*\right)^2 = \left\|\tilde{f} - f^*\right\|_{p^*}^2 + \eta\left\|\tilde{f} - f^*\right\|_{\mathcal{H}}^2 = \mathcal{O}_{p^*}\left(n^{-1}\eta^{-\frac{1}{r}} + \eta + \epsilon_{approx}\right) + o_{p^*}\left(\left(n^{-1}\eta^{-\frac{1}{r}} + \eta\right) + \epsilon_{approx}\right).$$

On the other hand, by Lemma 7

$$\mathbb{E}\left[\sum_{l=1}^{\infty}\left(\frac{\alpha_l}{a_l}\right)^2\right] = \mathcal{O}\left(n^{-1}\eta^{-\frac{1}{r}}\right).$$

Combining these bounds, we achieve the conclusion that

$$KL\left(p^*||p_{\tilde{f}}\right) + KL\left(p_{\tilde{f}}||p^*\right) = \mathcal{O}_{p^*}\left(n^{-1}\eta^{-\frac{1}{r}} + \eta + \epsilon_{approx}\right).$$

The second conclusion is straightforward by balancing η.

■

B MLE with Random Feature Approximation

The memory cost is the main bottleneck for applying the kernel methods to large-scale problems. The random feature (Rahimi and Recht, 2008; Dai et al., 2014; Bach, 2015) can be utilized for scaling up kernel methods. In this section, we will propose the variant of the proposed algorithm with random feature approximation.

For arbitrary positive definite kernel, $k(x,x)$, there exists a measure $\mathbb{P}$ on $\mathcal{X}$, such that $k(x,x') = \int \phi_w(x)\phi_w(x')d\mathbb{P}(w)$ Devinatz (1953); Hein and Bousquet (2004), where $\phi_w(x) : \mathcal{X} \to \mathbb{R}$ from $L_2(\mathcal{X},\mathbb{P})$. Therefore, we can approximate the function $f \in \mathcal{H}$ with Monte-Carlo approximation $\hat{f} \in \widehat{\mathcal{H}}^r = \{\sum_{i=1}^r \beta_i \phi_{\omega_i}(\cdot) | \|\beta\|_2 \leqslant C\}$ where $\{w_i\}_{i=1}^r$ sampled from $\mathbb{P}(\omega)$. The $\{\phi_{\omega_i}(\cdot)\}_{i=1}^r$ are called random features Rahimi and Recht (2009). With such approximation, we will apply the stochastic gradient to learn $\{\beta_i\}_{i=1}^r$. For simplicity, we still consider the saddle-point reformulation of MLE for exponential families. However, the algorithm applies to general flows and conditional models too.

Plug the approximation of $\widehat{f}(\cdot) = \sum_{i=1}^r \beta_i \phi_{\omega_i}(\cdot) = \beta_f^\top \Phi(\cdot)$ and $\widehat{\nu} = \beta_\nu^\top \Phi(\cdot)$ into the optimization (18) and denote $\beta = \{\beta_f, \beta_\nu\}$, we have

$$\min_{w_g} \bar{L}(w_g) := \max_{\beta_f,\beta_\nu} \underbrace{\tilde{\ell}(\beta_f,\beta_\nu,w_g) - \frac{\eta}{2}\|\beta_f\|^2}_{\bar{\ell}(\beta_f,\beta_\nu,w_g)}. \tag{37}$$

Therefore, we have the random feature variant of Algorithm 2

Algorithm 3 Stochastic Gradients for β_f^* and β_ν^*

1: **for** $k = 1,\ldots,K$ **do**
2: Sample $\xi \sim p(\xi)$, and generate $x = g(\xi)$.
3: Sample $x' \sim p_0(x)$.
4: Compute stochastic function gradient w.r.t. β_f and β_ν.
5: Decay the stepsize τ_k.
6: Update β_f^k and β_ν^k with the stochastic gradients.
7: **end for**
8: Output β_f^K, β_ν^K.

With the obtained $\left(\beta_f^K, \beta_\nu^K\right)$, the Algorithm 1 will keep almost the same, except in Step 2 call Algorithm 3 instead.

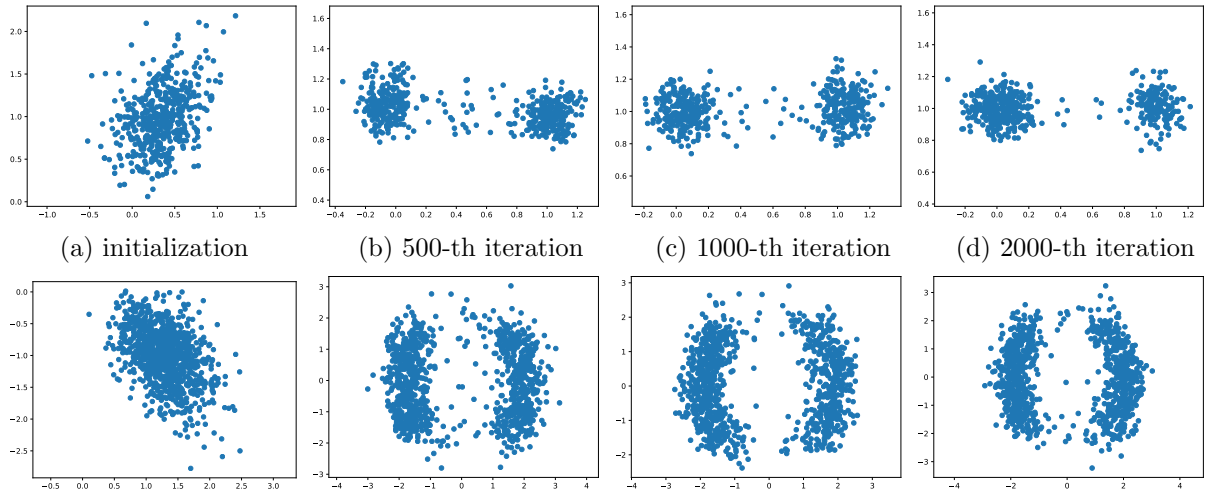
One can also adapt the random feature $\{\omega_i\}_{i=1}^r$ by stochastic gradient back-propagation (BP) too in Algorithm 3. Then, the ν is equivalent to parametrized by a two-layer MLP neural networks. Similarly, we can deepen the neural networks for ν, and the parameters can still be trained by BP.

C Computational Cost Analysis

Following the notations in the paper, the computational cost for Algorithm 2 will be $\mathcal{O}\left(K^2 d\right)$. Then, the total cost for Algorithm 1 will be $\mathcal{O}\left(L\left(K^2 d + BKd\right)\right)$ with B as batchsize and L as the number of iterations. If we stop the algorithm after scanning the dataset, i.e., $BL = N$, we have the cost as $\mathcal{O}\left(NK^2 d\right)$, which is more efficient comparing to score matching based estimator.

D More Experimental Results

We provide more empirical experimental results here. We further illustrate the convergence of the algorithm on the 2-dimensional `grid` and `two moons` in Figure 3.



(a) initialization (b) 500-th iteration (c) 1000-th iteration (d) 2000-th iteration

(a) initialization (b) 500-th iteration (c) 1000-th iteration (d) 2000-th iteration

Figure 3: The DDE estimators on 2-dimensional `grid` and `two moons` datasets in each iteration. The blue points are sampled from the learned dual distribution. The algorithm starts with random initialization. With the algorithm proceeds, the learned distribution converges to the ground-truth target distributions.

E Implementation Details

In this section, we will provide more details about algorithm implementation. Our implementation is based on PyTorch, and is open sourced at https://github.com/Hanjun-Dai/dde.

To optimize with the double min-max form, we adopt the following training schema. For every gradient update of the exponential family model f, 5 updates of sampler g_{w_g} will be performed. And for each update of g_{w_g}, 3 updates of ν will be performed. Generally, the inner terms of the objective function will get more updates.

For unconditional experiments on synthetic datasets, we use dimension 128 for both the hidden layers of MLP networks, as well as ξ. The number of layers for generator g_{w_g} and ν are tuned in the range of $\{3, 4, 5\}$. For conditional experiments on real-world datasets, we use 3 layers for both g_{w_g} and ν, since the dataset is relatively small. To make the training stable, we also clip the gradients of all updates by the norm of 5.

The hyperparameters, *e.g.*, stepsize, kernel parameters, and weights of the penalty, are tuned by cross-validation.