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 for 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 $\|\cdot\|^2_p$. 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 $\|f\|^2_{\mathcal{H}} = \sum_{l=1}^{\infty} \zeta_l^{-1} \langle f, \psi_l\rangle^2_{p^*}$ and $\|f\|^2_{p^*} = \sum_{l=1}^{\infty} \langle f, \psi_l\rangle^2_{p^*}$.

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.*

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

Lemma 6 *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 $\mathbb{E}[a_l^2] = \frac{1}{n}$, then we have

$$\mathbb{E}\left[\left\|\tilde{h} - f^*\right\|_{p^*}^2\right] = \frac{1}{n}\sum_{l=1}^\infty \frac{1}{\left(1+\eta\zeta_l^{-1}\right)^2} + \eta\sum_{l=1}^\infty \frac{\eta\zeta_l^{-1}}{\left(1+\eta\zeta_l^{-1}\right)^2}\cdot\zeta_t^{-1}(f_l^*)^2, \tag{24}$$
$$\mathbb{E}\left[\eta\left\|\tilde{h} - f^*\right\|_{\mathcal{H}}^2\right] = \frac{1}{n}\sum_{l=1}^\infty \frac{\eta\zeta_l^{-1}}{\left(1+\eta\zeta_l^{-1}\right)^2} + \eta\sum_{l=1}^\infty \frac{\eta^2\zeta_l^{-2}}{\left(1+\eta\zeta_l^{-1}\right)^2}\cdot\zeta_l^{-1}(f_l^*)^2. \tag{25}$$

By calculation, we obtain that

$$\begin{aligned}\sum_{l=1}^\infty \frac{\eta\zeta_l^{-1}}{\left(1+\eta\zeta_l^{-1}\right)^2} &= \sum_{l<\eta^{-\frac{1}{r}}} \frac{\eta\zeta_l^{-1}}{\left(1+\eta\zeta_l^{-1}\right)^2} + \sum_{l\geqslant\eta^{-\frac{1}{r}}} \frac{\eta\zeta_l^{-1}}{\left(1+\eta\zeta_l^{-1}\right)^2} \\ &= \mathcal{O}\left(\eta^{-\frac{1}{r}}\right) + \mathcal{O}\left(\int_{\eta^{-\frac{1}{r}}}^\infty \frac{\eta t^r}{(1+\eta t^r)^2}dt\right) \\ &= \mathcal{O}\left(\eta^{-\frac{1}{r}}\right) + \eta^{-\frac{1}{r}}\mathcal{O}\left(\int_1^\infty \frac{t^r}{(1+t^r)^2}dt\right) = \mathcal{O}\left(\eta^{-\frac{1}{r}}\right).\end{aligned} \tag{26}$$

Similarly, we can achieve

$$\sum_{l=1}^\infty \frac{1}{\left(1+\eta\zeta_l^{-1}\right)^2} = \mathcal{O}\left(\eta^{-\frac{1}{r}}\right), \tag{27}$$
$$\sum_{l=1}^\infty \frac{1}{1+\eta\zeta_l^{-1}} = \mathcal{O}\left(\eta^{-\frac{1}{r}}\right). \tag{28}$$

Note also that $\sum_{l=1}^\infty \zeta_t^{-1}(f_l^*)^2 = \|f^*\|_{\mathcal{H}}^2 < \infty$. Hence, the second term in (24) is also finite. Plugging (26) and (27) into (24), we achieve the conclusion. ■

Next, we proceed the second part of the error, *i.e.*, between $\tilde{f}$ and $\tilde{h}$.

Lemma 7 *Under Assumption 1 and Assumption 2, we have as $n \to \infty$ and $\eta \to 0$,*

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

Therefore,

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

Proof Since $\left(\tilde{f},\tilde{q}\right)$ are the optimal solutions to the primal-dual reformulation of the penalized MLE (9), we have the first-order optimality condition: $\nabla_f \ell(\tilde{f},\tilde{q})=0$, 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,

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

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

$$\widehat{\mathbb{E}}\left[\langle k\left(x,\cdot\right),\tilde{f}-\tilde{h}\rangle\right]-\mathbb{E}_{p^*}\left[\langle k\left(x,\cdot\right),\tilde{f}-\tilde{h}\rangle\right]-\langle\tilde{f}-\tilde{h},\tilde{h}-f^*\rangle_{p^*}-\eta\langle\tilde{h},\tilde{f}-\tilde{h}\rangle_{\mathcal{H}}=0. \tag{30}$$

Combining the (29) and (30), we further obtain

$$\begin{aligned}&\mathbb{E}_{\tilde{q}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]-\mathbb{E}_{p_{\tilde{h}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]+\eta\left\|\tilde{f}-\tilde{h}\right\|_{\mathcal{H}}^2 \\ &=\langle\tilde{f}-\tilde{h},\tilde{h}-f^*\rangle_{p^*}+\mathbb{E}_{p^*}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]-\mathbb{E}_{p_{\tilde{h}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right] \\ \Rightarrow\quad &\mathbb{E}_{p_{\tilde{f}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]-\mathbb{E}_{p_{\tilde{h}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]+\eta\left\|\tilde{f}-\tilde{h}\right\|_{\mathcal{H}}^2 \\ &=\langle\tilde{f}-\tilde{h},\tilde{h}-f^*\rangle_{p^*}+\underbrace{\mathbb{E}_{p^*}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]-\mathbb{E}_{p_{\tilde{h}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]}_{\epsilon_1} \\ &+\underbrace{\mathbb{E}_{p_{\tilde{f}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]-\mathbb{E}_{\tilde{q}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]}_{\epsilon_2}.\end{aligned} \tag{31}$$

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

$$F\left(\theta\right)=\frac{\theta}{\varsigma}\left(1+o_p(1)\right)\langle\tilde{f}-\tilde{h},\tilde{h}-f^*\rangle_{p^*} \tag{32}$$

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

$$\mathbb{E}_{p_{\tilde{h}}}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]-\mathbb{E}_{p^*}\left[\tilde{f}\left(x\right)-\tilde{h}\left(x\right)\right]=F\left(\varsigma\right)=\left(1+o_p(1)\right)\langle\tilde{f}-\tilde{h},\tilde{h}-f^*\rangle_{p^*}, \tag{33}$$

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

For ϵ_2, by Hölder inequality,

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

Due to the Assumption 3, $p^*\left(x\right)=\exp\left(f^*-A\left(f^*\right)\right)=\frac{\exp(f^*(x)-\log p_0(x))}{\int_\Omega \exp(f^*(x)-\log p_0(x))p_0(x)}p_0\left(x\right)$ with $f^*\in\mathcal{H}_k$ and $\left\|f^*\right\|_{\mathcal{H}}\leqslant C_{f^*}$ and $\left\|\log p_0\left(x\right)\right\|_\infty\leqslant C_0$, implies $2\exp\left(-\kappa C_{f^*}-C_0\right)\leqslant p^*\left(x\right)\leqslant 2\exp\left(\kappa C_{f^*}+C_0\right)$. Therefore, we have

$$\epsilon_2\leqslant 2\exp\left(\kappa C_{f^*}+C_0\right)\epsilon_{approx}\left\|\tilde{f}-\tilde{h}\right\|_{p^*}. \tag{34}$$

On the other hand, we define $D\left(\theta\right)=\mathbb{E}_{p_{\tilde{h}+\theta(\tilde{f}-\tilde{h})}}\left[\tilde{f}-\tilde{h}\right]$, notice that $D'\left(\theta\right)=\left\|\tilde{f}-\tilde{h}\right\|_{p_{\tilde{h}+\theta(\tilde{f}-\tilde{h})}}^2$, 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 (33) and (34) into (31), we achieve

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

which leads to the first part in the conclusion. Combining with the Lemma 6, 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 $n\eta^{\frac{1}{r}}\to\infty$,*

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

where $\epsilon_{approx}:=\sup_{f\in\mathcal{F}}\inf_{q\in\mathcal{P}_w}\|p_f-q\|_{p^*}$. 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 in rate $\mathcal{O}_{p^*}\left(n^{-\frac{r}{1+r}}+\epsilon^2_{approx}\right)$.

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

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

which leads to

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

Then, we can rewrite the Jensen-Shannon divergence

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

Similar to the bound of ϵ_2, we have

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

Moreover, with Cauchy-Schwarz inequality,

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

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

Applying the conclusion in Lemma 7 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

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

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,

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

On the other hand, by Lemma 6

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

Combining these bounds, we achieve the conclusion that

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

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 (19) 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^*

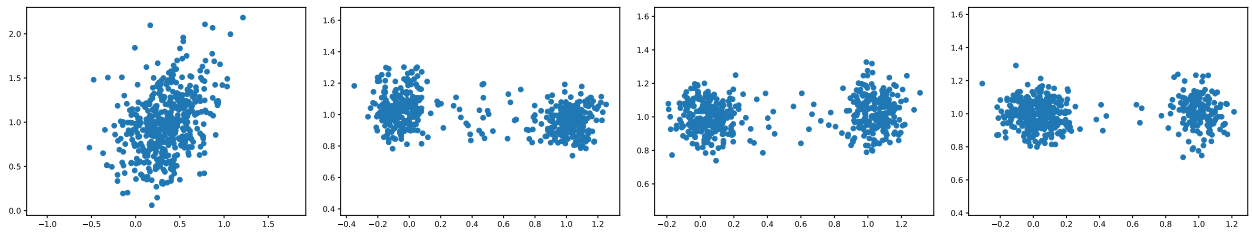
1: **for** $k = 1, \ldots, K$ **do**
2: Sample $\xi \sim p(\xi)$, and generate $x = g(\xi)$.
3: Sample $x' \sim p_0(x)$.
4: Compute stochastic function gradient w.r.t. β_f and β_ν.
5: Decay the stepsize τ_k.
6: Update β_f^k and β_ν^k with the stochastic gradients.
7: **end for**
8: Output β_f^K, β_ν^K.

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.

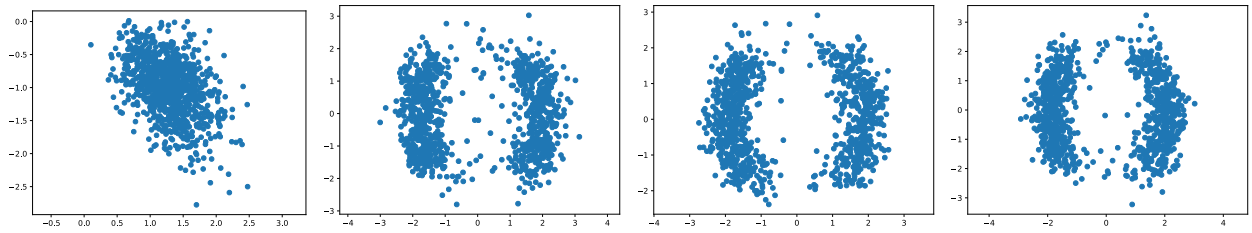
One can also adapt the random feature $\{\omega_i\}_{i=1}^r$ by stochastic gradient back-propagation (BP) too in Algorithm 3. Then, the ν is equivalent to parametrized by a two-layer MLP neural networks. Similarly, we can deepen the neural networks for ν, and the parameters can still be trained by BP.

C 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.

D 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.

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.