Efficient Model Inference with Stein Density Ratio Estimation

Tractable Inference for Intractable Models

Song Liu (song.liu@bristol.ac.uk)[1,4]
with Wittawat Jitkrittum[2],
Takafumi Kanamori[3],
and Yu Chen[1]
Special thanks to Carl Henrik Ek[1]

[1]University of Bristol, [2]Max-Plank Institute,
[3]Tokyo Institute of Technology,[4]Alan Turing Institute

Paper/Code Available

- Paper: `https://arxiv.org/abs/1805.07454`
- Code: `https://github.com/lamfeeling/Stein-Density-Ratio-Estimation`

Stein Density Ratio Estimation (SDRE) and Its Applications

Reference:

Song Liu, Takafumi Kanamori, Wittawat Jitkrittum, Yu Chen, Fisher Efficient Inference of Intractable Models, E-print: arXiv:1805.07454, To appear NeurIPS2019, 2019,

Install the `sdre` package

If you plan to modify our code (very likely, you will want to do so), it is best to install by:

1. Clone this repository
2. `cd` to the folder that you get, and install our package by (notice the dot at the end)

```
pip install -e .
```

Table of contents

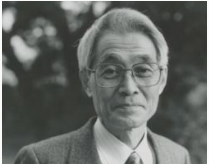
Intro

Problem: Model Inference

- Given **dataset** $X_q := \{\boldsymbol{x}_q^{(i)}\}_{i=1} \sim q_x$ and **model** $p(\boldsymbol{x}; \boldsymbol{\theta})$,
- Finding **model parameter** $\boldsymbol{\theta}$ that minimizes a statistical discrepancy (such as Kullback-Leibler (KL) divergence).

- Well studied problem: (**Wilk**'s Thoerem; **Fisher**'s Maximum Likelihood Estimation; **Akaike**'s Information Criterion).





Minimizing KL Divergence

Minimize a KL divergence

1. $\mathrm{KL}[q_X|p_{\boldsymbol{\theta}}] = \int q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x};\boldsymbol{\theta})} d\boldsymbol{x} = -\int q(\boldsymbol{x}) \log p(\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x} + C$
2. $\int q(\boldsymbol{x}) \log p(\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x} \approx \frac{1}{n} \sum_{i=1}^{n} \log p(\boldsymbol{x}_q^{(i)};\boldsymbol{\theta})$.
3. $\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \frac{1}{n} \sum_{i=1}^{n} \log p(\boldsymbol{x}_q^{(i)};\boldsymbol{\theta})$

3. is called **Maximum Likelihood Estimation (MLE)**

- **Normality**: $\hat{\theta}$ follows a normal distribution.
- **Efficiency**: $\hat{\theta}$ has the lowest asymptotic variance.

Intractable Model

$$p(\boldsymbol{x};\boldsymbol{\theta}) := \frac{\bar{p}(\boldsymbol{x};\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}$$

- $\bar{p}(\boldsymbol{x};\boldsymbol{\theta})$: unnormalized density, which has a parametric form and can be easily calculated.
- $Z(\boldsymbol{\theta}) := \int \bar{p}(\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x}$: normalization term.

- $p(\boldsymbol{x};\boldsymbol{\theta})$ may be **intractable**.
- $\bar{p}(\boldsymbol{x};\boldsymbol{\theta})$ is so complicated that $Z(\boldsymbol{\theta})$ **cannot be calculated**.
- We want to estimate $\boldsymbol{\theta}$; can we still minimize KL?

Tractability of Kullback-Leibler Divergence

KL is computationally intractable for an intractable model:
$\mathrm{KL}\left[q_X \| p_{\boldsymbol{\theta}}\right] := \mathbb{E}_q \left[\log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x};\boldsymbol{\theta})}\right]$.

- $p(\boldsymbol{x};\boldsymbol{\theta})$ cannot be evaluated.

Question
Can we find a *tractable* surrogate of KL from $p_{\boldsymbol{\theta}}$ to q?

- Fisher Divergence e.g. [Lyu, 2009, Hyvärinen, 2005]
- Kernel Stein Discrepancy (KSD) [Barp et al., 2019]

Can we find a surrogate that **still mimics behaviors of KL**?

Approximating KL

KL from p to q can be efficiently approximated

- $\mathrm{KL}[q\|p]$ is an expectation over a log ratio q/p.
- If we could approximate the ratio q/p, we can use sample to approximate the expecatation.

Density Ratio Estimation: Given two samples,

$$X_q := \{\boldsymbol{x}_q^{(i)}\}_{i=1}^{n_q} \sim q, X_p := \{\boldsymbol{x}_p^{(j)}\}_{j=1}^{n_p} \sim p,$$

Estimate $\hat{r}$ as the ratio between q and p using X_q and X_p.

$\mathrm{KL}\left[q|p\right] = \mathbb{E}_q\left[\log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right] \approx \frac{1}{n_q}\sum_{i=1}^{n_q} \log \hat{r}(\boldsymbol{x}_q^{(i)})$.

Density Ratio Estimation (KLIEP) (1)

Let us introduce a parametric model for density ratio: $r(\boldsymbol{x}; \boldsymbol{\delta})$.

$\boldsymbol{\delta}$ is fitted by minimizing a KL divergence $\mathrm{KL}[q \| r_{\boldsymbol{\delta}} p]$, i.e.,

$$\boldsymbol{\delta}^* := \operatorname*{argmin}_{\boldsymbol{\delta}} \mathrm{KL}[q \| r_{\boldsymbol{\delta}} p] \quad \text{s.t.} \int r(\boldsymbol{x}; \boldsymbol{\delta}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1, \tag{1}$$

$$\text{where} \quad \mathrm{KL}[q \| r_{\boldsymbol{\delta}} p] = \mathbb{E}_q[\log q(\boldsymbol{x})] - \mathbb{E}_q[\log r(\boldsymbol{x}; \boldsymbol{\delta})] - \mathbb{E}_q[\log p(\boldsymbol{x})]$$
$$\approx -\frac{1}{n_q} \sum_{i=1}^{n_q} \log r(\boldsymbol{x}_q^{(i)}; \boldsymbol{\delta}) + C,$$

and C is a constant [Sugiyama et al., 2008].

Density Ratio Estimation (2)

We can also approximate the equality constraint in (1) using X_p:

$$\int r(\boldsymbol{x}; \boldsymbol{\delta}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \approx \frac{1}{n_p} \sum_{j=1}^{n_p} r(\boldsymbol{x}_p^{(j)}; \boldsymbol{\delta}).$$

Thus, (1) can be re-written as:

$$\hat{\boldsymbol{\delta}} := \underset{\boldsymbol{\delta}}{\mathrm{argmin}} - \frac{1}{n_q} \sum_{i=1}^{n_q} \log r(\boldsymbol{x}_q^{(i)}; \boldsymbol{\delta}) \tag{2}$$

$$\text{s.t.} \frac{1}{n_p} \sum_{j=1}^{n_p} r(\boldsymbol{x}_p^{(j)}; \boldsymbol{\delta}) = 1.$$

Tractable, no need for MCMC sampling.

KLIEP Approximated KL for Model Inference?

Given X_q and $p(\boldsymbol{x}; \boldsymbol{\theta})$, how to estimate $q(\boldsymbol{x})/p_{\boldsymbol{\theta}}(\boldsymbol{x})$?.

We only have *one* sample X_q and a model $p_{\boldsymbol{\theta}}$.

Stein Density Ratio Estimation

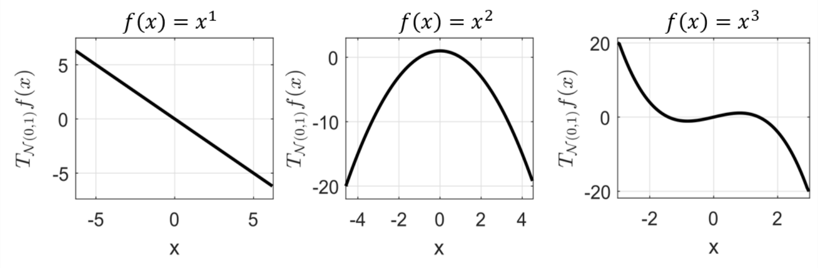
Stein Features

- Suppose $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^b$ is a feature function.
  - e.g. $\boldsymbol{f}(x) = [1, x, x^2 \dots x^{b-1}]$.
- Let $T_{\boldsymbol{\theta}}$ be a mapping:

$$T_{\boldsymbol{\theta}}\boldsymbol{f}(\boldsymbol{x}) := [T_{\boldsymbol{\theta}} f_1(\boldsymbol{x}), T_{\boldsymbol{\theta}} f_2(\boldsymbol{x}), \dots, T_{\boldsymbol{\theta}} f_b(\boldsymbol{x})],$$
$$T_{\boldsymbol{\theta}} f_i(\boldsymbol{x}) := \langle \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}), \nabla_{\boldsymbol{x}} f_i(\boldsymbol{x}) \rangle + \mathrm{tr}\left[\nabla^2_{\boldsymbol{x}} f_i(\boldsymbol{x})\right],$$

  where f_i is the i-th output of function $\boldsymbol{f}$.
  - $T_{\boldsymbol{\theta}}$ is called Stein Operator [Oates et al., 2017, Chwialkowski et al., 2016].
- Computing $T_{\boldsymbol{\theta}}\boldsymbol{f}(\boldsymbol{x})$ does **not** require the knowledge of $Z(\boldsymbol{\theta})$.

Stein Features: Visualization



Stein feature with respect to $p_\theta = \mathcal{N}(0,1)$ defined on $f_i(x) := x^i, i \in \{1,2,3\}$.

Stein Equality

- Stein Equality says, under **mild** regularity conditions:

$$\mathbb{E}_{p_{\boldsymbol{\theta}}}[\mathcal{T}_{\boldsymbol{\theta}} \boldsymbol{f}(\boldsymbol{x})] = \mathbf{0}.$$

  Proof see e.g., Lemma 5.1, [Chwialkowski et al., 2016]
- True for a large family of (continuously differentiable) $\boldsymbol{f}$,
  - kernel smoother
  - neural network, with some activation functions.
  - polynomial,

  w.r.t a wide range of $p_{\boldsymbol{\theta}}$.

Stein Density Ratio Model

We model density ratio function using Stein features:

$$r(\boldsymbol{x};\delta) := \boldsymbol{\delta}^\top T_{\boldsymbol{\theta}} \boldsymbol{f}(\boldsymbol{x}) + 1$$

It can be seen that

$$\begin{aligned}\int p_{\boldsymbol{\theta}}(\boldsymbol{x}) r(\boldsymbol{x};\boldsymbol{\delta}) \mathrm{d}\boldsymbol{x} &= \mathbb{E}_{p_{\boldsymbol{\theta}}}\left[\boldsymbol{\delta}^\top T_{\boldsymbol{\theta}} \boldsymbol{f}(\boldsymbol{x}) + 1\right] \\ &= \boldsymbol{\delta}^\top \mathbb{E}_{p_{\boldsymbol{\theta}}}\left[T_{\boldsymbol{\theta}} \boldsymbol{f}(\boldsymbol{x})\right] + 1 = 1.\end{aligned}$$

The last equality is due to **Stein equality**.

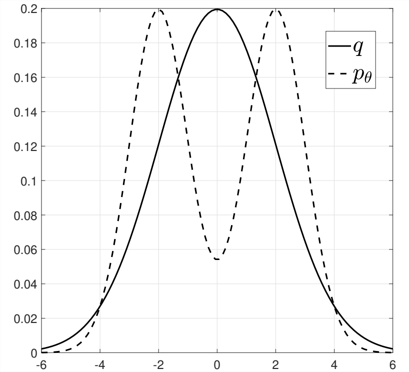
Stein Density Ratio Estimation (SDRE), 1

Now we have a **Stein Density Ratio Estimator**.

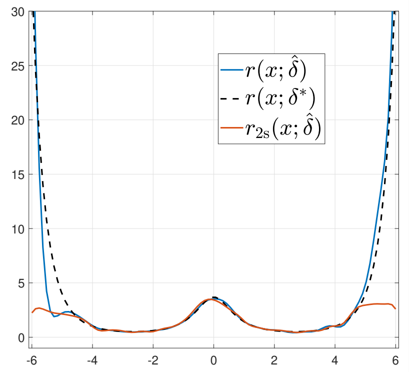
$$
\begin{aligned}
\hat{\boldsymbol{\delta}} &:= \operatorname*{argmax}_{\boldsymbol{\delta}} \frac{1}{n_q} \sum_{i=1}^{n_q} \log r(\boldsymbol{x}; \boldsymbol{\delta}) \\
&= \operatorname*{argmax}_{\boldsymbol{\delta}} \underbrace{\frac{1}{n_q} \sum_{i=1}^{n_q} \log \left[\boldsymbol{\delta}^\top T_{\boldsymbol{\theta}} \boldsymbol{f}(\boldsymbol{x}_q^{(i)}) + 1 \right]}_{\ell(\boldsymbol{\delta}; X_q, \boldsymbol{\theta})}
\end{aligned}
$$

- This is an unconstrained concave optimization.
- Do not need contraints, as it is automatically satisfied.

One Sample vs. Two Sample Density Ratio Estimation



Density function q and p_θ



$r(\boldsymbol{x};\hat{\boldsymbol{\delta}}), r(\boldsymbol{x};\boldsymbol{\delta}^*)$ and $r_{2s}(\boldsymbol{x};\hat{\boldsymbol{\delta}})$

Intractable Model Inference

Intractable Model Inference using "Discriminative-LE"

Ideally, MLE fits a $\boldsymbol{\theta}$ to X_q such that $\mathrm{KL}\left[q \| p_{\boldsymbol{\theta}}\right]$ is minimized.

- $\mathrm{KL}\left[q \| p_{\boldsymbol{\theta}}\right]$ is **intractable** in many cases.
- Now we have a **tractable** approximation:

$$\mathrm{KL}\left[q \| p_{\boldsymbol{\theta}}\right] \approx \mathbb{E}_q\left[\log r(\boldsymbol{x}; \hat{\boldsymbol{\delta}})\right] \approx \max_{\boldsymbol{\delta}} \ell(\boldsymbol{\delta}; X_q, \boldsymbol{\theta})$$

- Naturally, we can do a "Discriminative-LE (DLE)":

$$\hat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta}} \underbrace{\max_{\boldsymbol{\delta}} \ell(\boldsymbol{\delta}; X_q, \boldsymbol{\theta})}_{\text{discriminator}}$$

Consistent?

- **Q**: Under what conditions, DLE is consistent?
- Specifically, we study the following estimator:

$$(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\theta}}) := \arg \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \max_{\boldsymbol{\delta} \in \boldsymbol{\Delta}_{nq}} \ell(\boldsymbol{\delta}, \boldsymbol{\theta}).$$

- when $q(\boldsymbol{x}) \equiv p(\boldsymbol{x}; \boldsymbol{\theta}^*)$,
- when $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\delta}})$ is a saddle point.

Estimation Consistency

Theorem (Consistency)

Suppose $p(\boldsymbol{x}; \boldsymbol{\theta}^*) = q(\boldsymbol{x})$ and several regularity conditions hold.

$$(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\theta}}) \xrightarrow{\mathbb{P}} (\mathbf{0}, \boldsymbol{\theta}^*).$$

Estimation Consistency

MLE have many nice properties,

such as computable confidence interval on parameters.

However, computations is not possible when $p_{\boldsymbol{\theta}}$ is intractable.

Asymptotic Normality of $\hat{\theta}$

$\hat{\boldsymbol{\theta}}$ has a simple asymptotic distribution.

Theorem (Asymptotic Normality of $\hat{\theta}$)

Suppose $p(\boldsymbol{x};\boldsymbol{\theta}^*) = q(\boldsymbol{x})$ and several regularity conditions hold.

$$\sqrt{n_q}\left(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\right) \rightsquigarrow \mathcal{N}\left[0, \boldsymbol{V}\right],$$

where $\boldsymbol{V} := -\left(\mathbb{E}_q\left[\boldsymbol{H}^*\right]_{\boldsymbol{\theta},\boldsymbol{\delta}}\mathbb{E}_q\left[\boldsymbol{H}^*\right]^{-1}_{\boldsymbol{\delta},\boldsymbol{\delta}}\mathbb{E}_q\left[\boldsymbol{H}^*\right]_{\boldsymbol{\delta},\boldsymbol{\theta}}\right)^{-1}$,
$-\mathbb{E}_q\left[\boldsymbol{H}^*\right] := -\mathbb{E}_q\left[\nabla^2\ell(\boldsymbol{0},\boldsymbol{\theta}^*)\right]$.

Asymptotic Efficiency (Asymptotic Variance)

Theorem (Asymptotic Fisher Efficiency)
Given two Stein features $T_{\boldsymbol{\theta}^*}\boldsymbol{f}$, $T_{\boldsymbol{\theta}^*}\bar{\boldsymbol{f}}$, if $\operatorname{span} T_{\boldsymbol{\theta}^*}\boldsymbol{f} \subseteq \operatorname{span} T_{\boldsymbol{\theta}^*}\bar{\boldsymbol{f}}$, $\boldsymbol{V}_{\bar{\boldsymbol{f}}} \preceq \boldsymbol{V}_{\boldsymbol{f}}$. If $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{x};\boldsymbol{\theta}^*) \in \operatorname{span} T_{\boldsymbol{\theta}^*}\boldsymbol{f}$, $\boldsymbol{V}_{\boldsymbol{f}}$ attains a lowerbound

$$\mathbb{E}_q[\nabla_{\boldsymbol{\theta}} \log p(x;\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \log p(x;\boldsymbol{\theta}^*)^\top],$$

which is the Fisher information.

Note:
The smallest asymptotic variance an unbiased estimator can reach [Cramér, 1946, Rao, 1945].

Other asymptotic theorems can be found in our paper.

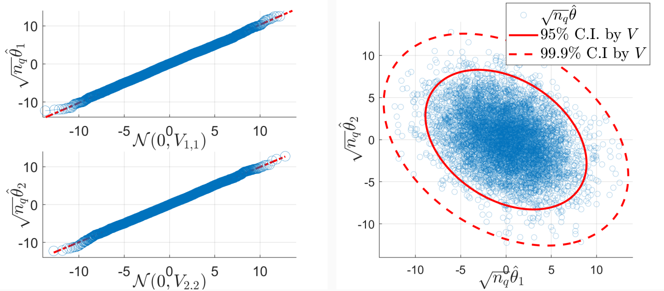
Experiments

Numerical Simulation: Distribution of $\hat{\theta}$

$$\bar{p}(\boldsymbol{x};\boldsymbol{\theta}) := \exp\left[\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{\psi}(\boldsymbol{x})\right], \boldsymbol{\eta}(\boldsymbol{\theta}) := [-.5, .2, .6, 0, 0, 0, \boldsymbol{\theta}]^\top,$$

$$\boldsymbol{\psi}(\boldsymbol{x}) := [\textstyle\sum_{i=1}^d x_i^2, \sum_{i=3}^d x_1 x_i, x_1 x_2, \tanh(\boldsymbol{x})]^\top, \ \boldsymbol{f}(\boldsymbol{x}) := [\tanh(\boldsymbol{x})]^\top.$$

$$q(\boldsymbol{x}) := p(\boldsymbol{x};\boldsymbol{0})$$



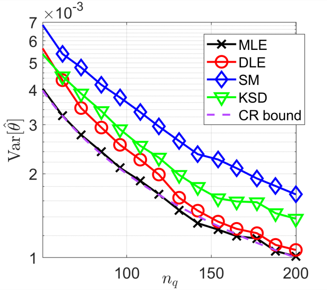
(a) Marginal: $\sqrt{n_q}\hat{\boldsymbol{\theta}}$ vs. Asymptotic prediction

(b) Joint: $\hat{\boldsymbol{\theta}}$ vs. Asymptotic prediction

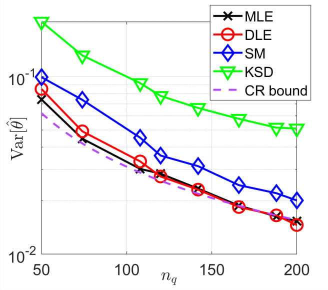
Numerical Simulation: $\mathrm{var}\left[\hat{\theta}\right]$

(Left): Gamma $p(x;\theta) = \Gamma(5,\theta), \theta^* = 1$

(Right): Gaussian mixture model
$p(x;\theta) = .5\mathcal{N}(\theta, 1) + .5\mathcal{N}(1, 1), \theta^* = -1$



(c) $\mathrm{Var}(\hat{\boldsymbol{\theta}})$, Gamma dist.

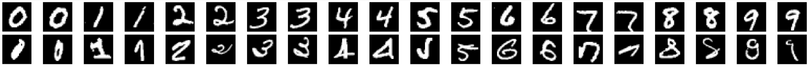


(d) $\mathrm{Var}(\hat{\boldsymbol{\theta}})$, Gaussian mix.

Note: we use tractable models so Fisher info. can be easily calculated.

Using Neural network as Sufficient Statistics

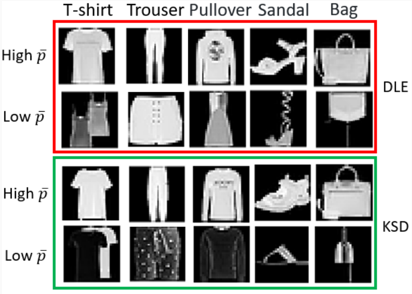
for $i = 0...9$, $\bar{p}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{i}}) := \exp[\boldsymbol{\theta}_{\mathrm{i}}^{\top} \boldsymbol{\psi}(\boldsymbol{x})], \boldsymbol{x} \in \mathbb{R}^{784}$, $\boldsymbol{\psi}(\boldsymbol{x}) \in \mathbb{R}^{20}$ is a pre-trained 3-layer **DNN** on the full MNIST dataset with all digits.



Images with highest (upper) and lowest $\log \bar{p}(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$ on each digit.

Using Neural network as Sufficient Statistics

for $i = 0...9$, $\bar{p}(\boldsymbol{x};\boldsymbol{\theta}_\mathrm{i}) := \exp[\boldsymbol{\theta}_\mathrm{i}^\top \psi(\boldsymbol{x})], \boldsymbol{x} \in \mathbb{R}^{784}$, $\psi(\boldsymbol{x}) \in \mathbb{R}^{20}$ is a pre-trained 3-layer **DNN** on the entire Fashion MNIST dataset.



Images with highest (upper) and lowest $\log \bar{p}(\boldsymbol{x};\hat{\theta})$ on each digit.

Conclusion

Conclusion

- With the help of Stein Operator, we can estimate a density ratio between a parametric p.d.f and data distribution.
- Use $\ell(\hat{\boldsymbol{\delta}}, X_q, \boldsymbol{\theta})$ as a “tractable replacement” to $\mathrm{KL}[q|p_{\boldsymbol{\theta}}]$ in model inference problems.
    - Discriminative Likelihood Estimation (DLE)
    - Asymptotic normality of $\hat{\boldsymbol{\theta}}$.
    - Asymptotic variance of $\hat{\boldsymbol{\theta}}$ reaches Camera-Rao bound.
- Experiments on toy and real datasets show promising results.

Thank you very much!

References i

Barp, A., Briol, F.-X., A.B., D., M., G., and L., M. (2019).
Minimum stein discrepancy estimators.

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

Cramér, H. (1946).
Mathematical methods of statistics.
Princeton university press.

- Hyvärinen, A. (2005).
  Estimation of non-normalized statistical models by score matching.
  Journal of Machine Learning Research, 6:695–709.
- Lyu, S. (2009).
  Interpretation and generalization of score matching.
  In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 359–366. AUAI Press.
- Oates, C. J., Girolami, M., and Chopin, N. (2017).
  Control functionals for monte carlo integration.
  Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3):695–718.

- Rao, C. R. (1945).
  Information and the accuracy attainable in the estimation of statistical parameters.
  Bulletin of the Calcutta Mathematical Society, 37:81–91.
- Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. (2008).
  Direct importance estimation for covariate shift adaptation.
  Annals of the Institute of Statistical Mathematics, 60(4):699–746.