Unified Perspective on Probability Divergence via the Density-Ratio Likelihood: Bridging KL-Divergence and Integral Probability Metrics

Masahiro Kato, CyberAgent, Inc.

Masaaki Imaizumi, The University of Tokyo

Kentaro Minami, Preferred Networks

1. Introduction

- How to measure probability divergences?
- **Kullback–Leibler (KL)–divergence**.
- **Integral Probability Metrics (IPMs)**. Ex. Wasserstein distance and MMD.
- We provide a unified view for the KL–divergence and IPMs from the perspective of maximum likelihood **density–ratio estimation** (DRE).
- We define several likelihoods based on different sampling schemes.

→ We derive KL–divergence and IPMs from the different likelihoods.

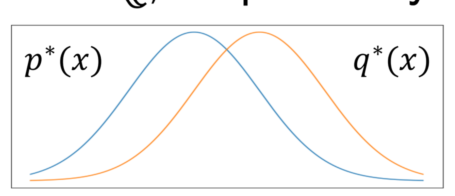
- Reveal the relationships between probability distances and density ratios.

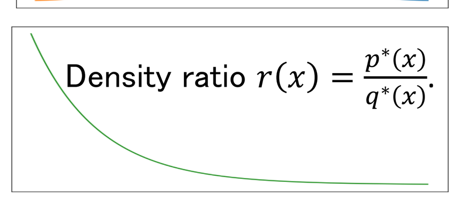
→ Understand the divergences and widen their applications.

2. Density Ratio

- Consider two distributions $\mathbb{P}$ and $\mathbb{Q}$ with a common support.
- We can relax the common support assumption later.
- Let p^* and q^* be the density functions of $\mathbb{P}$ and $\mathbb{Q}$, respectively.
- Define the **density ratio** (function) as

$$r^*(x) = \frac{p^*(x)}{q^*(x)}.$$





- We can estimate $r^*(x)$ by maximum likelihood estimation.

3. Maximum Likelihood Estimation

- Let $r(x)$ be a model of $r^*(x) = \frac{p^*(x)}{q^*(x)}$, and $\mathcal{R}$ be a class of r.

→ A model of $p^*(x)$ is given as $p(x) = r(x)q^*(x)$.

- For observations $\{X_i\}_{i=1}^n \sim p^*$, the likelihood of the model $p(x)$ is

$$\mathcal{L}(r) = \prod_{i=1}^n p(X_i) = \prod_{i=1}^n r(X_i)q^*(X_i).$$

- The log likelihood is given as $\ell(r) = \sum_{i=1}^n \log r(X_i) + \sum_{i=1}^n \log q^*(X_i)$.
- **Nonparametric maximum likelihood estimation of density ratios.**
- We can estimate r^* by solving

$$\max_{r\in\mathcal{R}} \frac{1}{n}\sum_{i=1}^n \log r(X_i) \quad \text{s.t.} \quad \int r(z)q^*(z)\mathrm{d}z = 1$$

- The constraint is based on $\int r^*(x)q^*(x)dx = \int p^*(x)dx = 1$.
- This formulation is equivalent to KLIEP (Sugiyama et al., 2008).

4. KL Divergence and IPMs

- **KL divergence (Kullback and Leibler, 1951):**

$$\mathrm{KL}(\mathbb{P} \parallel \mathbb{Q}) := \int p^*(x) \log \frac{p^*(x)}{q^*(x)} \mathrm{d}x.$$

- KL divergence can be approximated by the maximized log likelihood in DRE:

$$\widehat{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{Q}) = \max_{r\in\mathcal{R},\ \mathrm{s.t.} \int r(z)q^*(z)\mathrm{d}z=1} \frac{1}{n}\sum_{i=1}^n \log r(X_i).$$

- **IPMs with a function class $\mathcal{F}$ (Zolotarev, 1984; Müller, 1997):**

$$\mathrm{IPM}_{\mathcal{F}}(\mathbb{P} \parallel \mathbb{Q}) = \sup_{f\in\mathcal{F}} \left| \int f(x)p^*(x)\,\mathrm{d}x - \int f(z)q^*(z)\mathrm{d}z \right|.$$

- Ex. $\mathcal{F}$ is a class of Lipschitz continuous functions → Wasserstein distance.
- Likelihood under stratified samplings schemes:

Likelihood defined over both $\{X_i\}_{i=1}^n$ and $\{Z_j\}_{j=1}^m$.

- For observations $\{Z_j\}_{j=1}^m \sim q^*$, we can estimate $1/r^*$ by solving

$$\max_{r\in\mathcal{R}} \frac{1}{m}\sum_{j=1}^m \log\frac{1}{r(Z_j)} \quad \text{s.t.} \quad \int \frac{1}{r(x)}p^*(x)\mathrm{d}x = 1.$$

- By combining two likelihoods $\frac{1}{n}\sum_{i=1}^n \log r(X_i)$ and $-\frac{1}{m}\sum_{j=1}^m \log r(Z_j)$,

$$\max_{r\in\mathcal{R}} \frac{1}{n}\sum_{i=1}^n \log r(X_i) - \frac{1}{m}\sum_{j=1}^m \log r(Z_j) \quad \text{s.t.} \quad \int \frac{1}{r(x)}p^*(x)\mathrm{d}x = \int r(z)q^*(z)\mathrm{d}z = 1.$$

- This objective is a likelihood under stratified sampling scheme.
- Consider an exponential density ratio model, $r(x) = \exp(f(x))$. Then,

$$\max_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{m}\sum_{j=1}^m f(Z_j) \quad \text{s.t.} \int \exp(-f(x))\,p^*(x)\mathrm{d}x = \int \exp(f(z))\,q^*(z)\mathrm{d}z = 1$$

- **IPMs are maximized log likelihoods under stratified sampling scheme.**

5. Density Ratio Metrics

- Define the Density Ratio Metrics (DRMs):

$$\mathrm{DRM}^{\lambda}_{\mathcal{R}}(\mathbb{P} \parallel \mathbb{Q}) = \sup_{r\in C(\mathcal{R})} \left| \lambda \int \log r(x)\,p^*(x)\,\mathrm{d}x - (1-\lambda) \int \log r(x)\,q^*(x)\,\mathrm{d}x \right|,$$

$$C(\mathcal{R}) = \left\{ r \in \mathcal{R} \colon \int r(x)\,q^*(x)\mathrm{d}x = \int \frac{1}{r(x)}p^*(x)\mathrm{d}x = 1 \right\}.$$

- A distance based on the weighted average of maximum log likelihood of the density ratio under stratified sampling scheme ($\lambda \in [0,1]$).
- An empirical version of the DRMs is given as

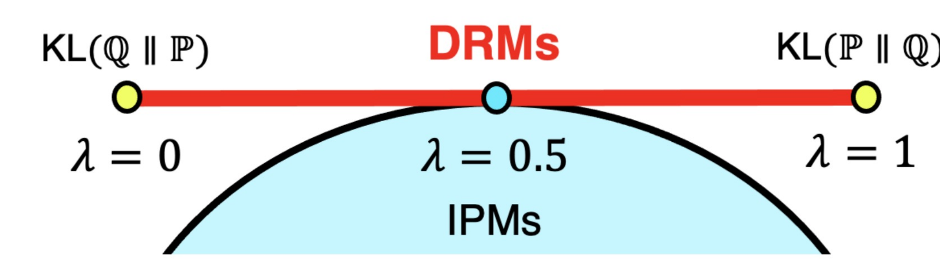
$$\widehat{\mathrm{DRM}}^{\lambda}_{\mathcal{R}}(\mathbb{P} \parallel \mathbb{Q}) = \max_{r\in C(\mathcal{R})} \left| \lambda \frac{1}{n}\sum_{i=1}^n \log r(X_i) - (1-\lambda)\frac{1}{m}\sum_{j=1}^m \log r(Z_j) \right|.$$

- DRMs include the KL divergence and IPMs.

6. Choice of λ and $\mathcal{R}$

- The hyper–parameter $\lambda \in [0,1]$ bridges the KL divergence and IPMs.
  - If $\lambda = 1/2$, $\mathrm{DRM}^{1/2}_{\mathcal{R}}(\mathbb{P} \parallel \mathbb{Q}) = \frac{1}{2}\mathrm{IPM}_{C(\mathcal{R})}(\mathbb{P} \parallel \mathbb{Q})$.
  - If $\lambda = 1$, $\mathrm{DRM}^{1}_{\mathcal{R}}(\mathbb{P} \parallel \mathbb{Q}) = \mathrm{KL}(\mathbb{P} \parallel \mathbb{Q})$.
  - If $\lambda = 0$, $\mathrm{DRM}^{0}_{\mathcal{R}}(\mathbb{P} \parallel \mathbb{Q}) = \mathrm{KL}(\mathbb{Q} \parallel \mathbb{P})$.

We can relax the common support assumption by defining the integral interval appropriately.



- Choice of λ.
- If we are interested in an accurate measure of probability distances, we can choose λ to minimize the asymptotic variance of $\widehat{\mathrm{DRM}}^{\lambda}_{\mathcal{R}}(\mathbb{P} \parallel \mathbb{Q})$.
- Under certain conditions, $\widehat{\mathrm{DRM}}^{\lambda}_{\mathcal{R}}(\mathbb{P} \parallel \mathbb{Q}) \xrightarrow{d} \mathcal{N}(\mathrm{DRM}^{\lambda}_{\mathcal{R}}(\mathbb{P} \parallel \mathbb{Q}), V(\lambda))$, where $V(\lambda) = \lambda \mathrm{Var}\left(\int \log r(x)\,p^*(x)\,\mathrm{d}x\right) + (1-\lambda)\mathrm{Var}\left(\int \log r(x)\,q^*(x)\,\mathrm{d}x\right)$.
- $\lambda = \frac{\mathrm{Var}_{\mathbb{P}}(\log r(x))}{\mathrm{Var}_{\mathbb{P}}(\log r(x)) + \mathrm{Var}_{\mathbb{Q}}(\log r(x))}$ minimizes the asymptotic variance.
- Choice of $\mathcal{R}$. → Smoothness of the model of the density ratio.

Ex. Spectral normalization (Miyato et al., 2016)

7. Applications and Experiments

- **DRE:** estimate the density ratio by maximizing the likelihoods.
- DRE from two Gaussian distributions with different dimensions.
- Compare the performances of DRMs–based DRE with the uLSIF (Kanamori et al., 2009) and RuLSIF (Yamada et al., 2011).
- We report the squared errors with different $\lambda \in \{0.1, 0.5, 0.9\}$.
- Tuning λ can lower the squared errors.

Table 1: Results of Section 6: means, medians, and stds of the squared error in DRE using synthetic datasets. The lowest mean and median (med) methods are highlighted in bold.

dim	uLSIF			RuLSIF ($\alpha = 0.1$)			WD			DRM ($\lambda = 0.5$)			DRM ($\lambda = 0.1$)			DRM ($\lambda = 0.9$)		
	mean	med	std	mean	med	std	mean	med	std	mean	med	std	mean	med	std	mean	med	std
2	11.216	9.997	4.547	11.655	10.481	4.566	7.962	4.536	13.838	1.781	1.107	2.025	**1.565**	**0.977**	1.762	3.039	2.194	2.959
10	15.512	14.673	3.218	15.285	14.292	3.178	12.164	13.033	4.879	3.638	3.222	1.861	**2.901**	**2.488**	1.583	5.577	4.885	2.151
50	15.578	15.045	3.326	15.586	15.055	3.326	14.307	13.922	3.614	6.654	6.032	2.340	**4.879**	**4.273**	1.856	10.117	9.451	2.728
100	16.479	15.356	4.194	16.318	15.054	4.220	9.040	6.976	6.849	**7.540**	**6.748**	2.812	8.785	8.305	2.266	12.163	11.074	3.552

- **Two sample test:** measure the divergences between $\mathbb{P}$ and $\mathbb{Q}$ by DRMs.
- Investigate whether $\mathbb{P} = \mathbb{Q}$ or $\mathbb{P} \neq \mathbb{Q}$.
- 95% confidence interval using DRMs, KLIEP, and MMD for two sample test.
- Report rejection rates.
- If the null–hypothesis is true, rates close to 0.05 are better.

	Beta 1	Normal 1	Beta 2	Normal 2	Normal 3
	null-hypothesis is true		null-hypothesis is not true		
DRM (optimal λ)	0.04	0.07	0.30	0.15	0.11
DRM ($\lambda = 0.5$)	0.04	0.07	0.30	0.15	0.11
KLIEP (DRM with $\lambda = 1$)	0.07	0.05	0.28	0.11	0.13
MMD	0.06	0.05	0.38	0.10	0.25

- If the null–hypothesis is not true, larger rates are better.
- We also show results on other applications, such as image generation.

Reference

Kullback, S. and Leibler, R. A. (1951) "On Information and Sufficiency." Zolotarev, V. M. (1984) "Probability Metrics."

Müller, A. (1997) "Integral Probability Metrics and Their Generating Classes of Functions."

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. (2008)

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018), "Spectral Normalization for Generative Adversarial Network."

Kanamori, T., Hido, S., and Sugiyama, M. (2009), "A Least-squares Approach to Direct Importance Estimation,"

Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., and Sugiyama, M. (2011), "Relative Density-Ratio Estimation for Robust Distribution Comparison."