Adaptive Learning of Density Ratios in RKHS

Werner Zellinger[1], Stefan Kindermann[2], Sergei V. Pereverzyev[1]

[1] Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences

[2] Industrial Mathematics Institute, Johannes Kepler University Linz, Austria

ÖAW RICAM

JKU

CONTRIBUTION

(a) Error rates for density ratio estimators
(b) Parameter choice method achieving rate
(c) Re-solving saturation issue by iteration

DENSITY RATIO ESTIMATION

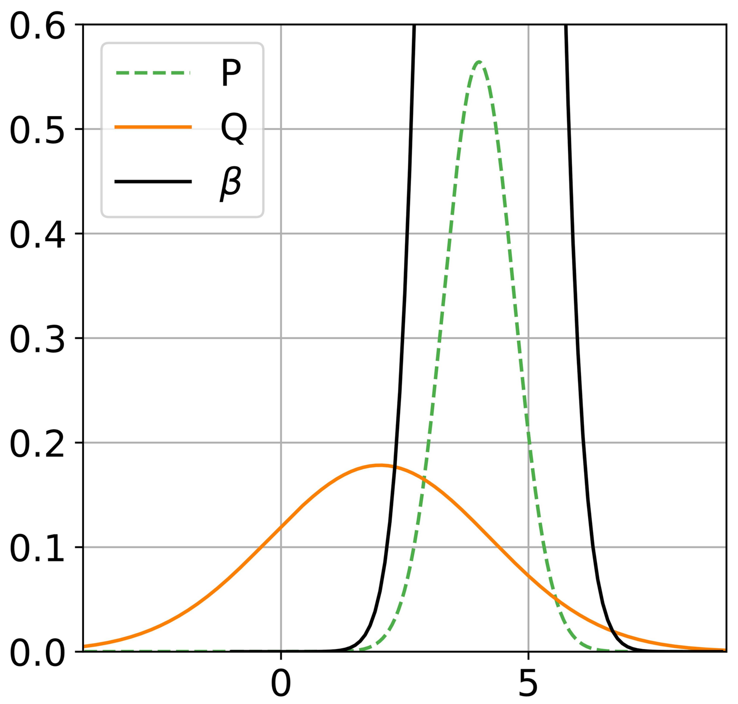

Figure 1: Density ratio β between P and Q.

Problem: Given $\{x_i\}_{i=1}^m \overset{\text{iid}}{\sim} P$ and $\{x'_i\}_{i=1}^n \overset{\text{iid}}{\sim} Q$, recover $\beta := \frac{dP}{dQ}$.

Applications:

- Importance weighting

$$\mathbb{E}_{x\sim P}[(f(x) - f_Q(x))^2] = \mathbb{E}_{x\sim Q}[\beta(x)(f(x) - f_Q(x))^2]$$

- Csiszárs ϕ-divergence estimation

$$D_\phi(P\|Q) = \int_{\mathcal{X}} \phi(\beta(x))\,\mathrm{d}P(x)$$

- Neyman-Pearson Lemma (1933)
  A simple statistical significance test of maximal power has unique representation as density ratio threshold decision.

BREGMAN DIVERGENCE METHODS

Methods [Sugiyama, Suzuki, Kanamori; 2012]:

Use estimator $\widehat{\beta}(x) := g(\widehat{f}_\lambda(x))$ with $\widehat{f}_\lambda$ solving empirical estimation of

$$\min_{f\in\mathcal{H}} B_F(\beta, g(f)) + \lambda\,\|f\|_{\mathcal{H}}^2 \qquad (1)$$

for RKHS $\mathcal{H}$, convex $F: L^1(Q) \to \mathbb{R}$, link function g, regularization parameter $\lambda > 0$ and Bregman divergence

$$B_F(\beta, \widehat{\beta}) = F(\beta) - F(\widehat{\beta}) - \nabla F(\beta)[\beta - \widehat{\beta}].$$

Examples:

- KuLSIF [Kanamori, Hidu, Sugiyama; 2009]

$$F_{\text{KuLSIF}}(h) = \frac{1}{2}\int_{\mathcal{X}} (h(x) - 1)^2\,\mathrm{d}Q(x)$$
$$B_F(\beta, g(f)) = \|\beta - f\|_{L^2(Q)}^2$$

- LogReg [Bickel, Brückner, Scheffer; 2010]

$$F_{\text{LR}}(h) = \int_{\mathcal{X}} h(x)\log(h(x)) - (1 + h(x))\log(1 + h(x))\,\mathrm{d}Q(x)$$
$$g_{\text{LR}}(f) = e^f$$

- Square [Menon and Ong; 2016]

$$F_{\text{SQ}}(h) = \int_{\mathcal{X}} 1/(2h(x) + 2)\,\mathrm{d}Q(x)$$
$$g_{\text{SQ}}(f) = \frac{-1 + 2f}{2 - 2f}$$

- Boosting [Menon and Ong; 2016]

$$F_{\text{Exp}}(h) = \int_{\mathcal{X}} h(x)^{-3/2}\,\mathrm{d}Q(x),\ g_{\text{Exp}}(f) = e^{2f}$$

RESULTS

Assumption 1 (Data generation model). *The data* $\mathbf{z} := (x_i, 1)_{i=1}^m \cup (x'_i, -1)_{i=1}^n$ *is independently drawn from* ρ, *where* $\rho(x|y=1) := P(x)$, $\rho(x|y=-1) := Q(x)$ *and* $\rho_{\{-1,1\}}(y=1) = \rho_{\{-1,1\}}(y=-1) = \frac{1}{2}$.

Lemma 1 (Menon and Ong, 2016). *For* $\circ \in \{\text{KuLSIF}, \text{LR}, \text{Exp}, \text{SQ}\}$ *it exists* $\ell_\circ : \{-1,1\}\times\mathbb{R} \to \mathbb{R}$ *with*

$$B_{F_\circ}(\beta, g_\circ(f)) = 2\left(\mathcal{R}(f) - \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f)\right)$$

where $\mathcal{R}(f) = \int_{\mathcal{X}\times\{-1,1\}} \ell_\circ(y, f(x))\,\mathrm{d}\rho(x,y)$.

Assumption 2 (Source condition). $\exists r \in (0, \frac{1}{2}], v \in \mathcal{H} : f_{\mathcal{H}} = \mathbf{H}(f_{\mathcal{H}})^r$, *where* $\mathbf{H}(f) := \mathbb{E}_{Z\sim\rho}[\nabla^2 \ell_Z(f)]$ *with* $\ell_{(x,y)}(f) := \ell(y, f(x))$ *and* $f_{\mathcal{H}} := \arg\min_{f\in\mathcal{H}} \mathcal{R}(f)$.

Assumption 3 (Capacity condition). $\exists \alpha \geq 1, Q \geq 0 : \mathrm{df}_\lambda \leq Q\lambda^{-\frac{1}{\alpha}}$ *with the effective dimension* $\mathrm{df}_\lambda := \mathbb{E}_{Z\sim\rho}\left[\left\|(\mathbf{H}(f_{\mathcal{H}}) + \lambda I)^{-\frac{1}{2}}\,\nabla\ell_Z(f_{\mathcal{H}})\right\|^2\right]$.

Theorem (Error rates result). *Let* $\circ \in \{\text{KuLSIF}, \text{LR}, \text{Exp}, \text{SQ}\}$, $\delta \in [\frac{2}{e^{1296}}, \frac{1}{2}]$ *and* m, n *be large enough. Further fix increasing sequence* $(\lambda_i)_{i=1}^l \in \mathbb{R}$ *and select*

$$\lambda^\circ := \max\left\{\lambda_i : \left\|\left(\widehat{\mathbf{H}}(\widehat{f}_{\lambda_j}) + \lambda_j I\right)^{\frac{1}{2}}\left(\widehat{f}_{\lambda_i} - \widehat{f}_{\lambda_j}\right)\right\|^2 \leq \frac{c}{\lambda^{1/\alpha}(m+n)}\log(2/\delta), j \in \{1,\ldots,i-1\}\right\}.$$

Then the following holds with probability $1 - 2\delta$:

$$B_{F_\circ}(\beta, g_\circ(\widehat{f}_{\lambda^\circ})) - B_{F_\circ}(\beta, g_\circ(f_{\mathcal{H}})) \leq c(m+n)^{-\frac{2r\alpha+\alpha}{2r\alpha+\alpha+1}}.$$

Rate is minimax optimal for SQ (and $r \leq \frac{1}{2}$). For $r > \frac{1}{2}$, Eq. (1) needs [2] to overcome saturation.

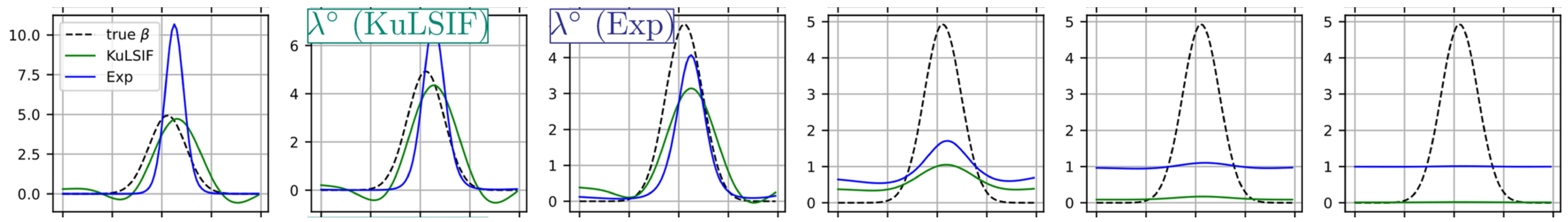

Figure 2: A posteriori Lepskii type parameter choice achieves error rate.

[1] W. Zellinger, S. Kindermann, and S.V. Pereverzyev. Adaptive learning of density ratios in RKHS. *Journal of Machine Learning Research* 24(395), 2023.

[2] L. Gruber, M. Holzleitner, J. Lehner, S. Hochreiter, and W. Zellinger. Overcoming saturation in density ratio estimation by iterated regularization. *International Conference on Machine Learning*, 2024.