Lecture 3: Covariate Shift

Yang Feng[1], Ye Tian[2]

[1]Department of Biostatistics, School of Global Public Health, New York University
[2]Department of Statistics, Columbia University

Overview

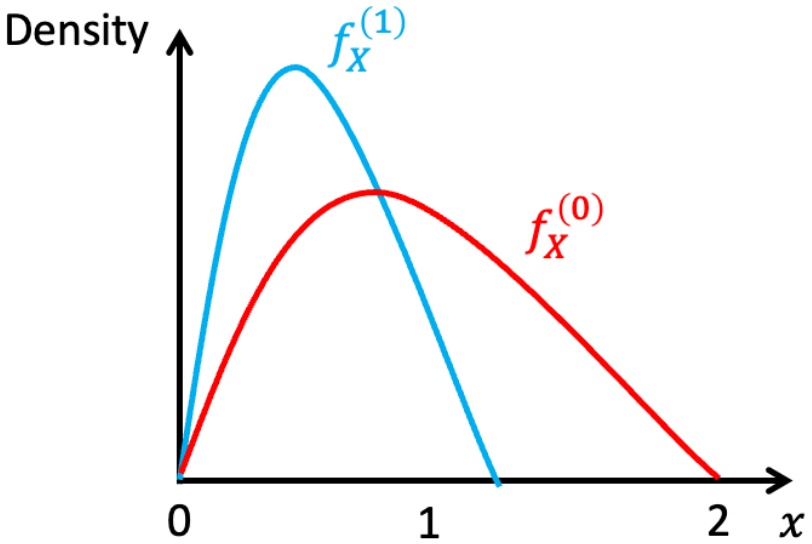
1. §3.1: Covariate shift
2. §3.2: Adaptivity to covariate shift for free?
3. §3.3: The reweighting method
4. §3.4: Density ratio estimation
   - §3.4.1 A naive method: separate density estimation
   - §3.4.2 Histogram-based method
   - §3.4.3 Kernel mean matching
   - §3.4.4 Discriminative learning
   - §3.4.5 Kullback-Leibler method
   - §3.4.6 Semi-parametric method
   - §3.4.7 Least square method
5. References

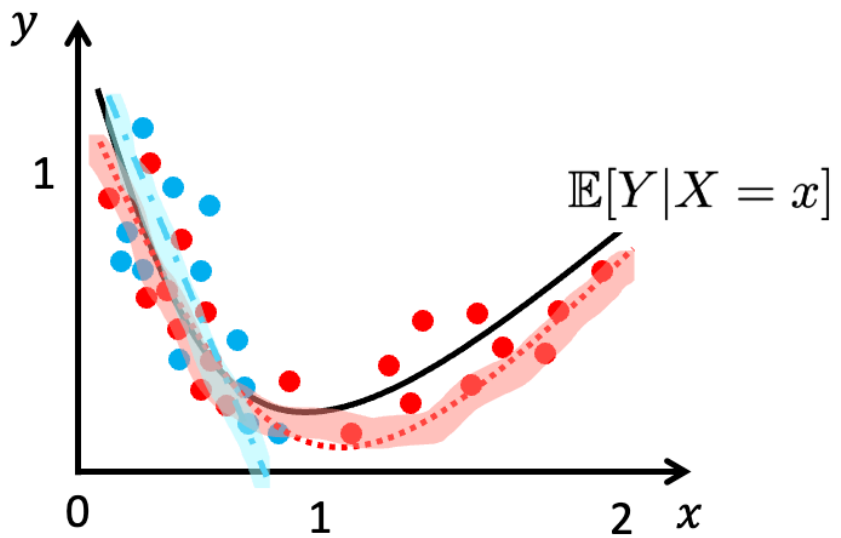
§3.1: Covariate shift

Covariate shift

For simplicity, we consider the target and one single source in this section.

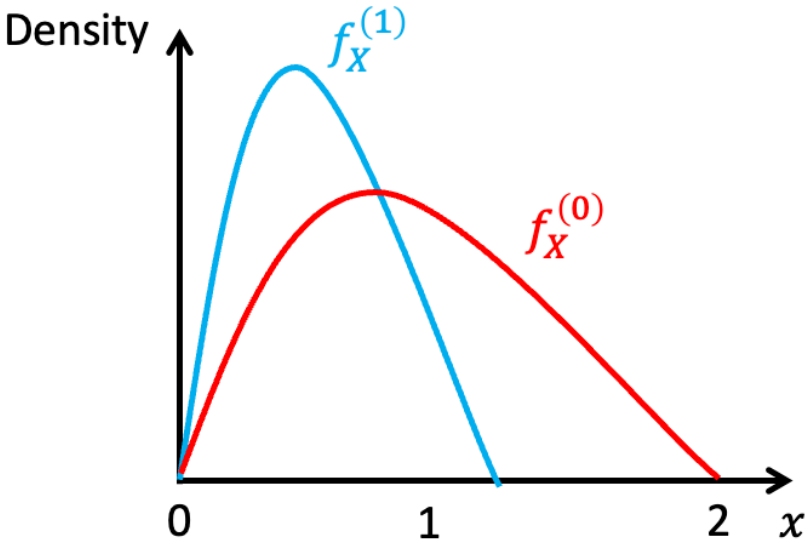
- Target distribution $(X^{(0)}, Y^{(0)}) \sim \mathbb{P}^{(0)}$
- Source distribution $(X^{(1)}, Y^{(1)}) \sim \mathbb{P}^{(1)}$
- **Covariate shift**: $\mathbb{P}_X^{(0)} \neq \mathbb{P}_X^{(1)}$, $\mathbb{P}_{Y|X}^{(0)} = \mathbb{P}_{Y|X}^{(1)}$ (in the $\mathbb{P}_X^{(0)}$-a.s. sense)
- **Goal:** learn $\mathbb{E}[Y|X = x]$ or make prediction on target domain

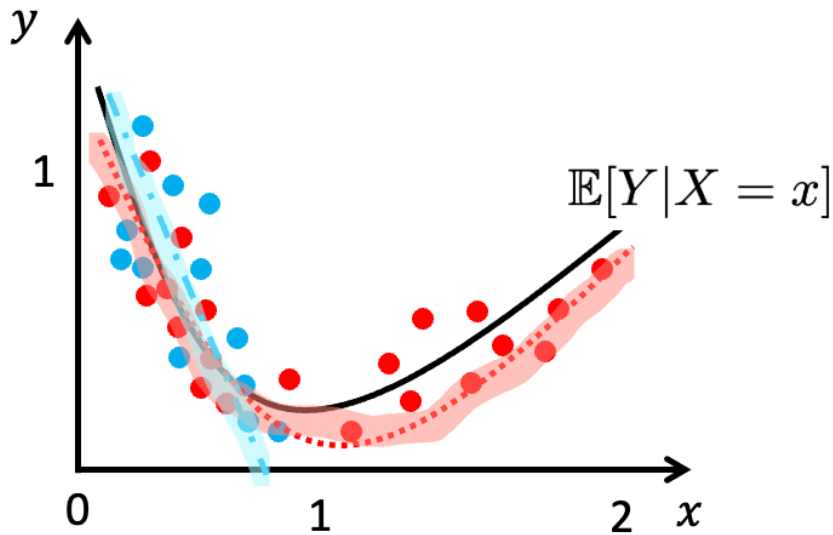




§3.2: Adaptivity to covariate shift for free?

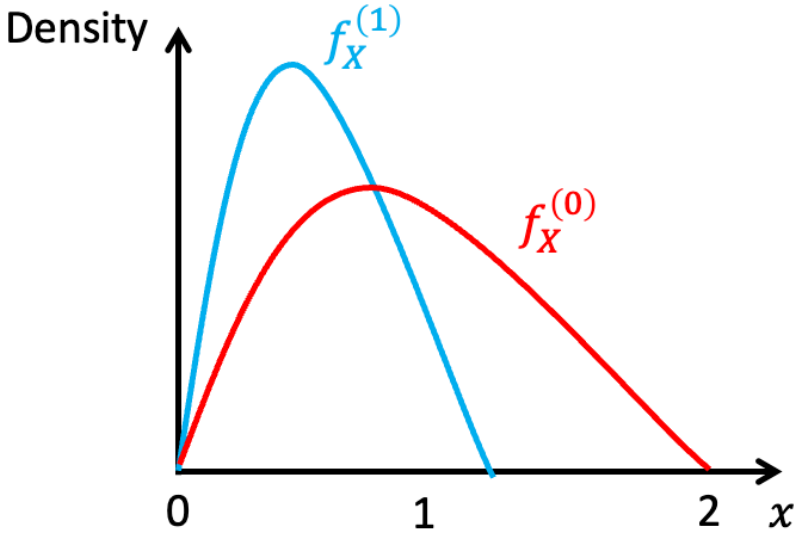
What we need

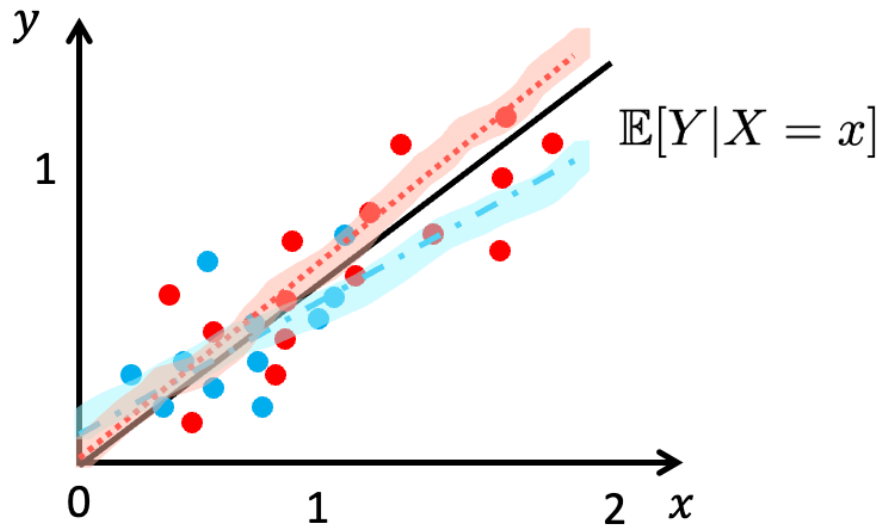




- What do we need to resolve this covariate shift issue?
  - Full and enough (X, Y) data from both the source and the target?
    ✓ (just learn from the target)
  - Full (X, Y) data from the source, only X from the target?
    ✓ (in many cases)
  - Full (X, Y) data from the source, no data from the target?
    ✗ (in general)

A second example





- It seems that in this case, fitting a linear regression model on source data only is enough
- **How it works:** (Suppose $Y = X^\top \boldsymbol{\theta}^* + N(0, \sigma^2)$ for both domains)
  (1) OLS on the full source data of size n $\quad\Rightarrow\quad$ $\widehat{\boldsymbol{\theta}}$ with $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \lesssim_{\mathbb{P}} n^{-1/2}$;
  (2) Prediction error on the target domain:

$$\begin{aligned}\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}}[(X^\top\widehat{\boldsymbol{\theta}} - Y)^2] &\le \sigma^2 + \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}}[(X^\top\widehat{\boldsymbol{\theta}} - X^\top\boldsymbol{\theta}^*)^2] \\ &\lesssim \sigma^2 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 \\ &\lesssim_{\mathbb{P}} \underbrace{\sigma^2}_{\text{irreducible}} + n^{-1}.\end{aligned}$$

Understand this phenomenon from MLE

Consider $Y|X = \boldsymbol{x} \sim$ density $p_{\boldsymbol{\theta}^*}(y|\boldsymbol{x})$ and negative log-likelihood as the loss

$$\ell_{\boldsymbol{\theta}}(\boldsymbol{x}, y) = -\log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$$

and the population-level risk

$$R^{(k)}(\boldsymbol{\theta}) = -\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(k)}}[\log p_{\boldsymbol{\theta}}(Y|X)] = -\mathbb{E}_{X\sim\mathbb{P}^{(k)}}\mathbb{E}_{Y|X\sim p_{\boldsymbol{\theta}^*}}[\log p_{\boldsymbol{\theta}}(Y|X)].$$

It is easy to see that

$$\begin{aligned}\mathbb{E}_{Y|X=\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}[\log p_{\boldsymbol{\theta}^*}(Y|\boldsymbol{x})] - \mathbb{E}_{Y|X=\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}[\log p_{\boldsymbol{\theta}}(Y|\boldsymbol{x})] &= d_{\mathsf{KL}}\big(p_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})\big) \\ &\geq 0,\end{aligned}$$

for any $\boldsymbol{x}$. Hence

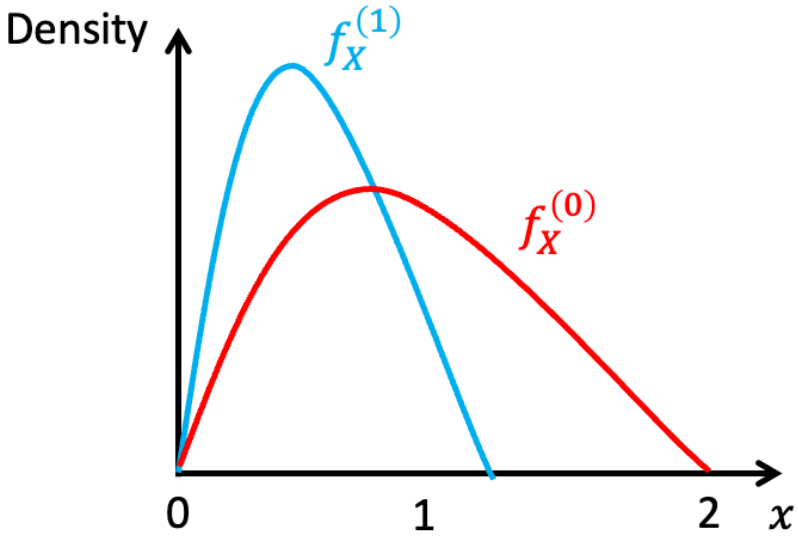
$$R^{(k)}(\boldsymbol{\theta}) \geq R^{(k)}(\boldsymbol{\theta}^*), \quad k = 0, 1.$$

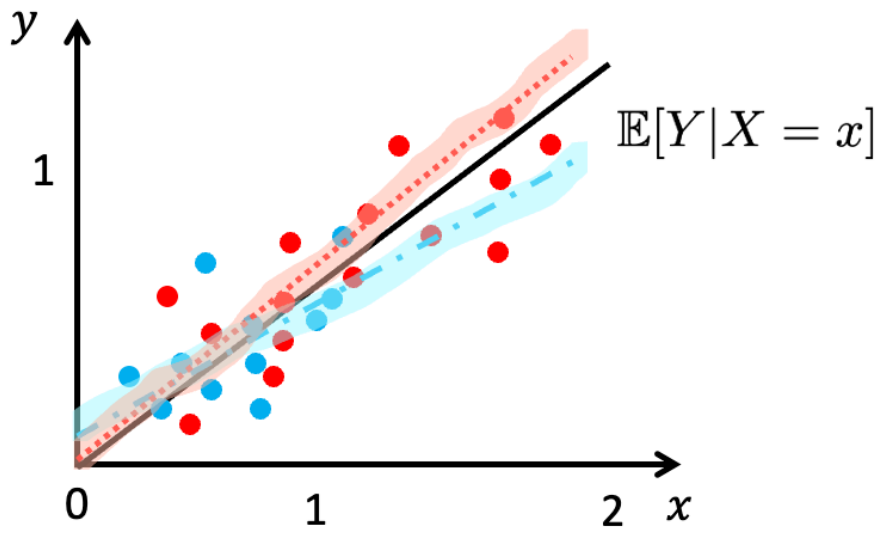
The initial analysis of this phenomenon was conducted in Shimodaira (2000).

[1] Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of statistical planning and inference, 90(2), 227-244.

Intuition

Recall our previous example:





How it works: (Suppose $Y = X^\top \boldsymbol{\theta}^* + N(0, \sigma^2)$ for both domains)

(1) OLS on the full source data of size n $\quad\Rightarrow\quad$ $\widehat{\boldsymbol{\theta}}$ with $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \lesssim_{\mathbb{P}} n^{-1/2}$;

(2) Prediction error on target domain: $\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}}[(X^\top \widehat{\boldsymbol{\theta}} - Y)^2] \lesssim_{\mathbb{P}} \sigma^2 + n^{-1}$.

Necessary ingredients:

- Well-specified parametric models for $Y|X = \boldsymbol{x}$: $p(y|\boldsymbol{x}) = p_{\boldsymbol{\theta}^*}(y|\boldsymbol{x})$ for some $\boldsymbol{\theta}^*$
- Good ``curvature" for source risk around $\boldsymbol{\theta}^*$ (to guarantee recovery of $\boldsymbol{\theta}^*$)

Adaptivity to covariate shift for free

- Consider an ERM problem:
  - Negative log-likelihood as loss $\ell(\boldsymbol{\theta}; \boldsymbol{x}, y) = -\log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ and population-level risk $R^{(k)}(\boldsymbol{\theta}) = \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(k)}}[\ell(\boldsymbol{\theta}; X, Y)]$
  - $Y|X = \boldsymbol{x} \sim p_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{x})$ for both domains
  - Data $\{(\boldsymbol{x}_i^{(1)}, y_i^{(1)})\}_{i=1}^{n_1}$ from the source domain
  - Empirical risk $\hat{R}^{(1)}(\boldsymbol{\theta}) = \frac{1}{n_1}\sum_{i=1}^{n_1} \ell(\boldsymbol{\theta}; \boldsymbol{x}_i^{(1)}, y_i^{(1)})$
  - Source ERM estimator: $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \hat{R}^{(1)}(\boldsymbol{\theta})$

Theorem 3.2.1 (Ge et al. (2024))

Under certain conditions, $R^{(0)}(\widehat{\boldsymbol{\theta}}) - R^{(0)}(\boldsymbol{\theta}^*) \lesssim_{\mathbb{P}} \frac{\mathsf{Tr}\left(\mathcal{I}^{(0)}(\mathcal{I}^{(1)})^{-1}\right)}{n_1}$, *where Fisher information* $\mathcal{I}^{(k)} = \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(k)}}[\nabla^2 \ell(\boldsymbol{\theta}^*; X, Y)]$.

Recall our previous intuitions:

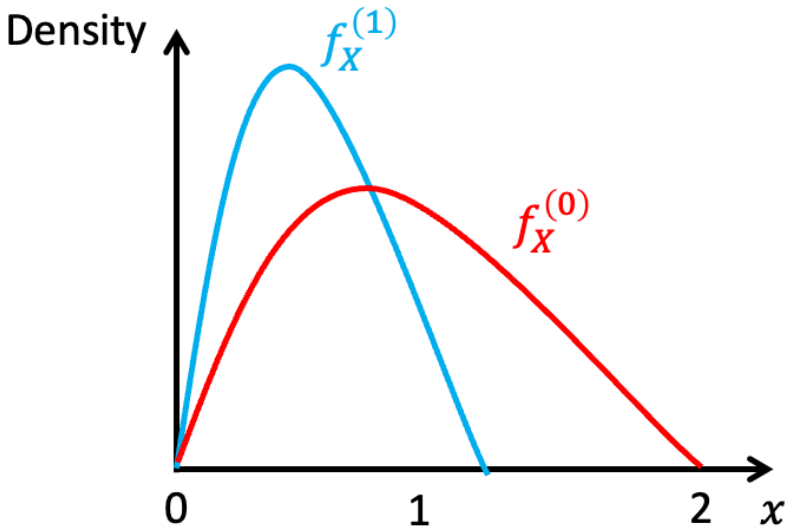
- Well-specified models for $Y|X = \boldsymbol{x}$: ✓
- Good ``curvature" for source risk around $\boldsymbol{\theta}^*$ (to guarantee recovery of $\boldsymbol{\theta}^*$): ✓, see $(\mathcal{I}^{(1)})^{-1}$ term

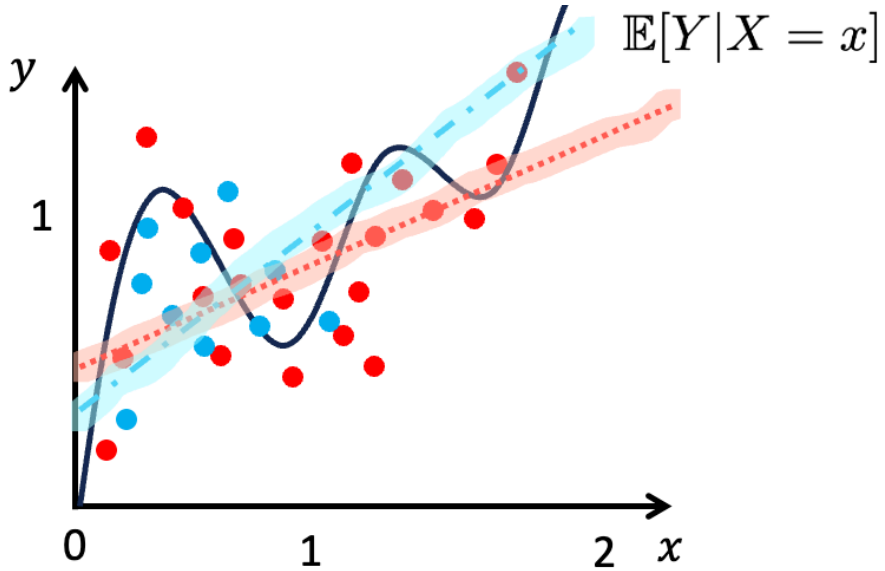
  ``*MLE is all you need for well-specified covariate shift.*" -- Ge et al. (2024)

[1] Ge, J., Tang, S., Fan, J., Ma, C., & Jin, C. (2024). Maximum Likelihood Estimation is All You Need for Well-Specified Covariate Shift. In The Twelfth International Conference on Learning Representations.

Adaptivity to covariate shift for free

- The idea can be extended from MLE with well-specified models to ERM with general parametric models





Necessary ingredients:

- Good ``curvature" for source risk around optimal ``parameter" $\boldsymbol{\theta}^*$: guarantees a good estimation error of $\boldsymbol{\theta}^*$, e.g. $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \lesssim_{\mathbb{P}} \ldots$
- The target excess risk can be bounded by the estimation error of $\boldsymbol{\theta}^*$, e.g. $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$

Beyond parametric models

There are some recent works on kernel ridge regression (KRR) where similar phenomenons are found. With Mercer kernels, the induced RKHS $\mathcal{H}$ admits a countable orthonormal basis of $L^2(\mathcal{X}; \mathbb{P}^{(0)})$.

- Ma et al. (2023) showed that if the regression function $f^* \in \mathcal{H}$ and the density ratio $\frac{\mathrm{d}\mathbb{P}_X^{(0)}}{\mathrm{d}\mathbb{P}_X^{(1)}}$ is bounded, $\mathbb{E}_{X\sim\mathbb{P}^{(0)}}|\hat{f}^{\mathsf{KRR}}(X) - f^*(X)|^2$ can be well controlled
- Wang (2023) showed that if the regression function $f^* \in \mathcal{H}$, even when the density ratio $\frac{\mathrm{d}\mathbb{P}_X^{(0)}}{\mathrm{d}\mathbb{P}_X^{(1)}}$ is unbounded, a strategy called ``pseudo-labeling" can lead to a well-controlled $\mathbb{E}_{X\sim\mathbb{P}^{(0)}}|\hat{f}(X) - f^*(X)|^2$

Surprising?

- **Yes:** The same phenomenon extends to non-parametric models in the ∞-dimensional space
- **No:** With Mercer kernels, RKHS $\mathcal{H}$ can be embedded into $\ell_2(\mathbb{N})$, hence our previous intuitions (estimation error $\to$ prediction error) can still work

[1] Ma, C., Pathak, R., & Wainwright, M. J. (2023). Optimally tackling covariate shift in RKHS-based nonparametric regression. The Annals of Statistics, 51(2), 738-761.

[2] Wang, K. (2023). Pseudo-labeling for kernel ridge regression under covariate shift. arXiv preprint arXiv:2302.10160.

§3.3: The reweighting method

Motivation

Consider an ERM setup:

- Target domain: $(X, Y) \sim \mathbb{P}^{(0)}$, source domain: $(X, Y) \sim \mathbb{P}^{(1)}$
- Covariate shift: $\mathbb{P}_X^{(0)} \neq \mathbb{P}_X^{(1)}$, $\mathbb{P}_{Y|X}^{(0)} = \mathbb{P}_{Y|X}^{(1)}$ with $\mathbb{P}_X^{(0)}$-a.s. X
- Loss function: $\ell(y, y')$, risk $R^{(k)}(h) = \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(k)}}[\ell(h(X), Y)]$
- Goal: find an $h \in \mathcal{H}$ that minimizes target risk $R^{(0)}(h)$

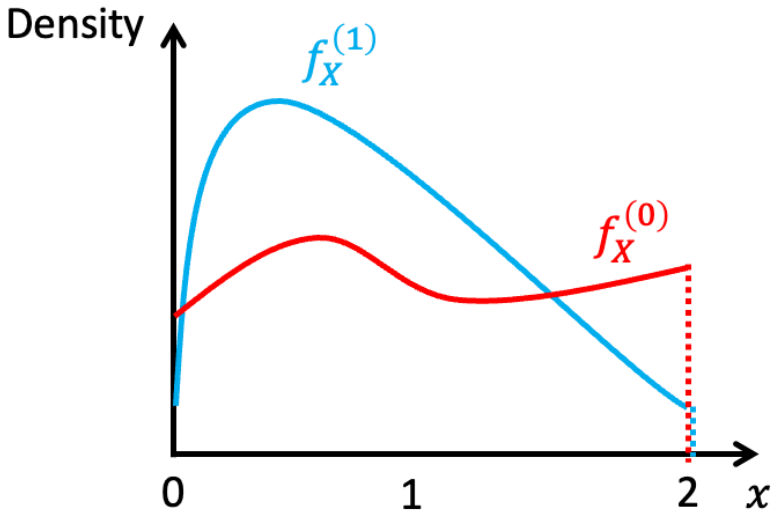
With labeled source data $\{(\boldsymbol{x}_i^{(1)}, y_i^{(1)})\}_{i=1}^{n_1}$, we are able to conduct ERM:

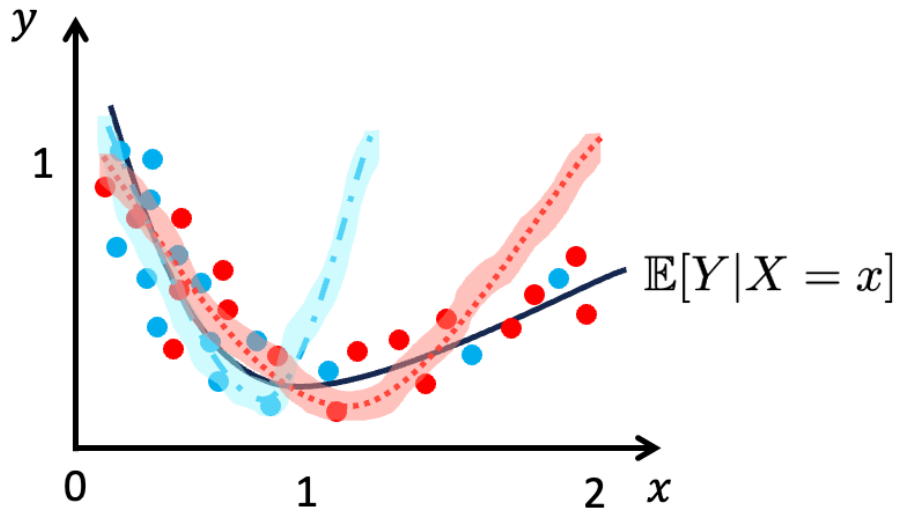
$$\hat{h} \in \arg\min_{h\in\mathcal{H}} \hat{R}^{(1)}(h),$$

with $\hat{R}^{(1)}(h) = {n_1}^{-1} \sum_{i=1}^{n_1} \ell(h(\boldsymbol{x}_i^{(1)}), y_i^{(1)})$.

Warning: Source ERM could suffer from severe biases because $R^{(1)}(h) \neq R^{(0)}(h)$ in general.

Motivation





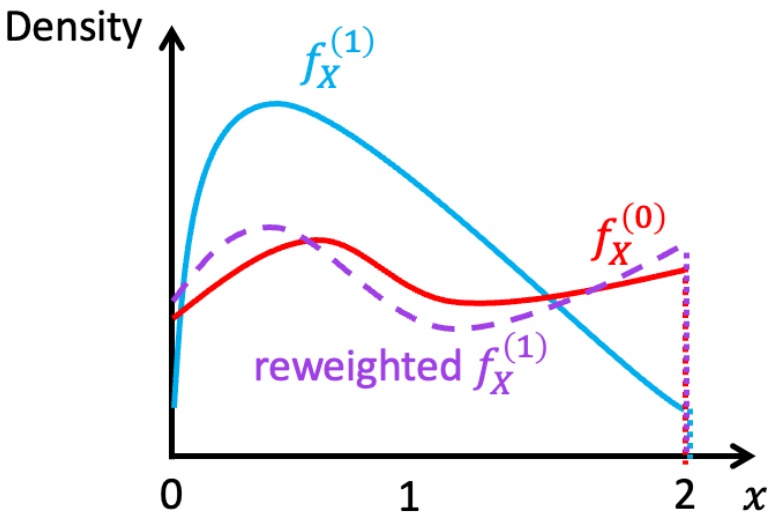
The **density ratio** plays a critical rule here.

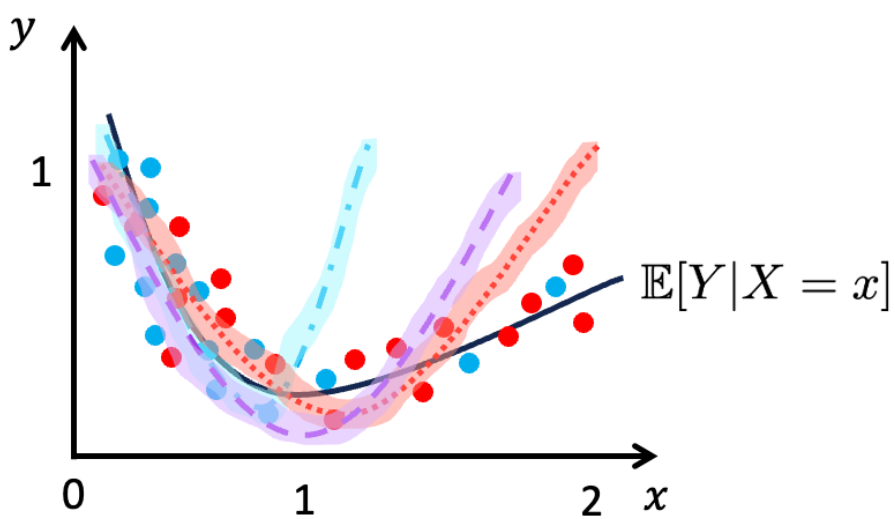
$$\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}}[\ell(h(X),Y)] = \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(1)}}\left[\frac{\mathrm{d}\mathbb{P}^{(0)}}{\mathrm{d}\mathbb{P}^{(1)}}(X,Y)\cdot\ell(h(X),Y)\right]$$

$$= \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(1)}}\left[\frac{\mathrm{d}\mathbb{P}_X^{(0)}}{\mathrm{d}\mathbb{P}_X^{(1)}}\cdot\frac{\cancel{\mathrm{d}\mathbb{P}_{Y|X}^{(0)}}}{\cancel{\mathrm{d}\mathbb{P}_{Y|X}^{(1)}}}\cdot\ell(h(X),Y)\right]$$

$$= \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(1)}}\left[\underbrace{\frac{\mathrm{d}\mathbb{P}_X^{(0)}}{\mathrm{d}\mathbb{P}_X^{(1)}}(X)}_{:=w(X)}\cdot\ell(h(X),Y)\right].$$

Solution: reweighting the loss by density ratio





$$\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}}[\ell(h(X),Y)] = \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(1)}}\Bigg[\underbrace{\frac{\mathrm{d}\mathbb{P}_X^{(0)}}{\mathrm{d}\mathbb{P}_X^{(1)}}(X)}_{w(X)} \cdot \ell(h(X),Y)\Bigg].$$

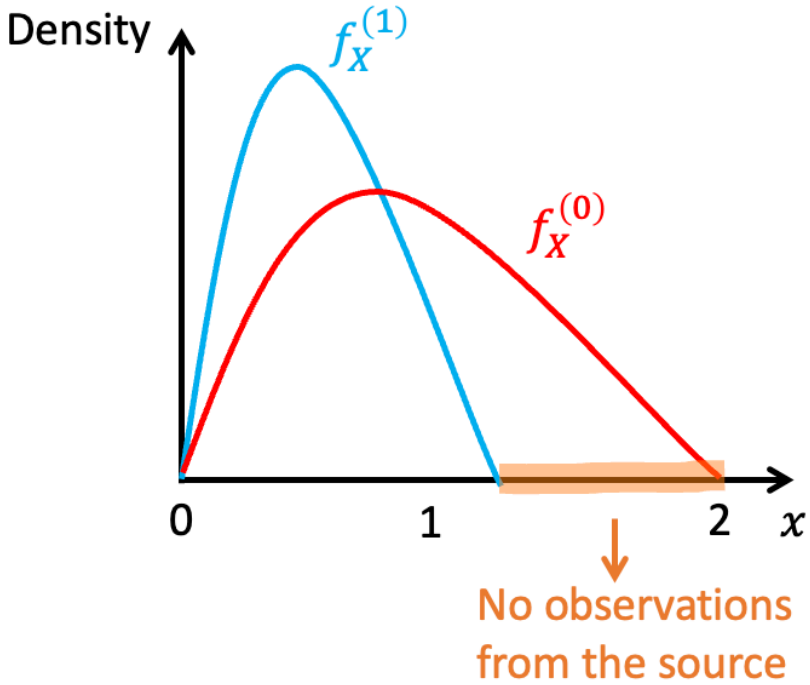
Reweighted ERM: ERM on source data with reweighted loss function

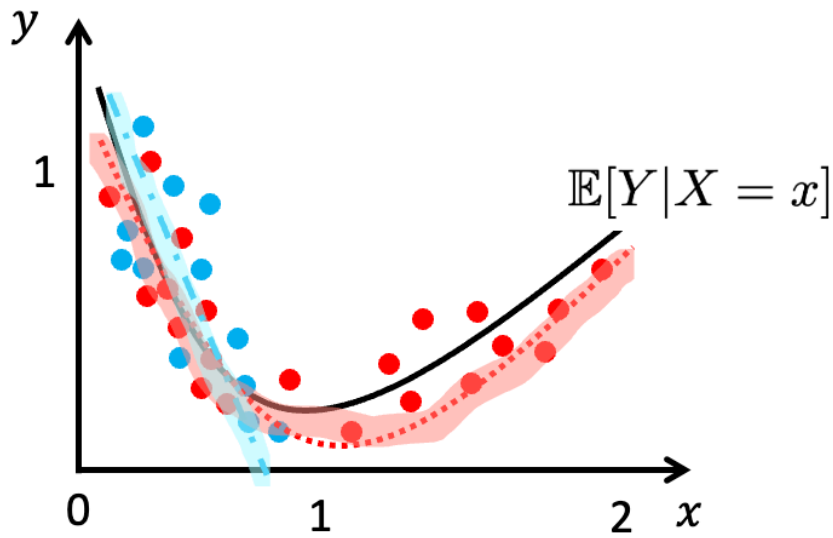
$$\hat{h} \in \arg\min_{h\in\mathcal{H}} \left\{ \frac{1}{n_1}\sum_{i=1}^{n_1} w(\boldsymbol{x}_i^{(1)})\ell(h(\boldsymbol{x}_i^{(1)}), y_i^{(1)}) \right\}.$$

Solution: reweighting the loss by density ratio

Problems:

- The weight function (i.e. density ratio) w is usually **unknown** in practice
  - Solution: estimate it
  - We only need some unlabeled target data, which is usually easier to get
- Unbounded/$+\infty$ density ratio





  - Not fixable in general, especially the $+\infty$ case
  - We will assume the existence of prob. measure σ (e.g., Lebesgue) s.t. $\mathbb{P}^{(0)}, \mathbb{P}^{(1)} \ll \sigma$ and the density ratio $w(X) = \frac{\mathrm{d}\mathbb{P}^{(0)}/\mathrm{d}\sigma}{\mathrm{d}\mathbb{P}^{(1)}/\mathrm{d}\sigma}$ is **bounded**

Why reweighting works: a simple theoretical justification

Consider the reweighted ERM on the source data

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \left\{ \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{w}(\boldsymbol{x}_i^{(1)}) \ell(h(\boldsymbol{x}_i^{(1)}), y_i^{(1)}) \right\}.$$

Then if the loss function ℓ is bounded:

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}}[\ell(\hat{h}(X), Y)] \\
&= \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(1)}}[w(X) \cdot \ell(\hat{h}(X), Y)] \\
&\le [\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(1)}} - \mathbb{E}_{(X,Y)\sim\widehat{\mathbb{P}}^{(1)}}][w(X) \cdot \ell(\hat{h}(X), Y)] + \mathbb{E}_{(X,Y)\sim\widehat{\mathbb{P}}^{(1)}}[w(X) \cdot \ell(\hat{h}(X), Y)] \\
&\le \mathcal{O}_{\mathbb{P}}(1) + \mathbb{E}_{(X,Y)\sim\widehat{\mathbb{P}}^{(1)}}[\hat{w}(X) \cdot \ell(\hat{h}(X), Y)] + C\mathbb{E}_{(X,Y)\sim\widehat{\mathbb{P}}^{(1)}}|\hat{w}(X) - w(X)| \\
&\le \mathbb{E}_{(X,Y)\sim\widehat{\mathbb{P}}^{(1)}}[\hat{w}(X)\ell(h^*(X), Y)] + \mathcal{O}_{\mathbb{P}}(1) + C\mathbb{E}_{(X,Y)\sim\widehat{\mathbb{P}}^{(1)}}|\hat{w}(X) - w(X)| \\
&\le \underbrace{\min_{h \in \mathcal{H}} \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(1)}}[w(X)\ell(h^*(X), Y)]}_{\text{oracle}} + \underbrace{\mathcal{O}_{\mathbb{P}}(1)}_{\text{uniform convergence}} + \underbrace{2C\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}|\hat{w}(X) - w(X)|}_{\text{cost of estimating the weight}}.
\end{aligned}$$

This bound might be loose but it is good enough to justify the reweighted ERM here.

§3.4: Density ratio estimation

- §3.4.1 A naive method: separate density estimation
- §3.4.2 Histogram-based method
- §3.4.3 Kernel mean matching
- §3.4.4 Discriminative learning
- §3.4.5 Kullback-Leibler method
- §3.4.6 Semi-parametric method
- §3.4.7 Least square method

Note: In this section, depending on the context, $\mathbb{P}^{(k)}$ can be either the joint distribution of (X, Y) or the marginal distribution of X

§3.4: Density ratio estimation

- §3.4.1 A naive method: separate density estimation
- §3.4.2 Histogram-based method
- §3.4.3 Kernel mean matching
- §3.4.4 Discriminative learning
- §3.4.5 Kullback-Leibler method
- §3.4.6 Semi-parametric method
- §3.4.7 Least square method

A naive method: separate density estimation

Recall that the density ratio $w(\boldsymbol{x}) = \frac{\mathrm{d}\mathbb{P}^{(0)}/\mathrm{d}\sigma}{\mathrm{d}\mathbb{P}^{(1)}/\mathrm{d}\sigma}(\boldsymbol{x}) := \frac{f^{(0)}}{f^{(1)}}(\boldsymbol{x})$.

We can estimate two densities $f^{(0)}$, $f^{(1)}$ separately by $\hat{f}^{(0)}$, $\hat{f}^{(1)}$, then plug in $\hat{w}(\boldsymbol{x}) = \frac{\hat{f}^{(0)}}{\hat{f}^{(1)}}(\boldsymbol{x})$. Let us review some commonly used density estimation methods.

Common density estimation methods:

- Histogram
  - Divide the space into bins $B_1, \cdots, B_m$
  - Estimate the density as $\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{m} \frac{\#\{i:\boldsymbol{x}_i \in B_j\}}{n \cdot \mathrm{Vol}(B_j)} \mathbb{1}(\boldsymbol{x} \in B_j)$
- Kernel density estimation (KDE)
  - Choose a kernel function $K : \mathbb{R}^p \to \mathbb{R}^+$
  - Estimate the density as $\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \frac{1}{nh} K(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h})$
  - Examples of kernels ($p = 1$):
    - Box kernel: $K(x) = \mathbb{1}(|x| \le 1)$
    - Gaussian kernel: $K(x) = (2\pi)^{-1} \exp\{-x^2/2\}$
    - Epanechnikov kernel: $K(x) = \frac{3}{4}(1 - x^2)\mathbb{1}(|x| \le 1)$

See Wasserman (2006) for more details.

[1] Wasserman, L. (2006). All of nonparametric statistics. Springer Science & Business Media.

A naive method: separate density estimation

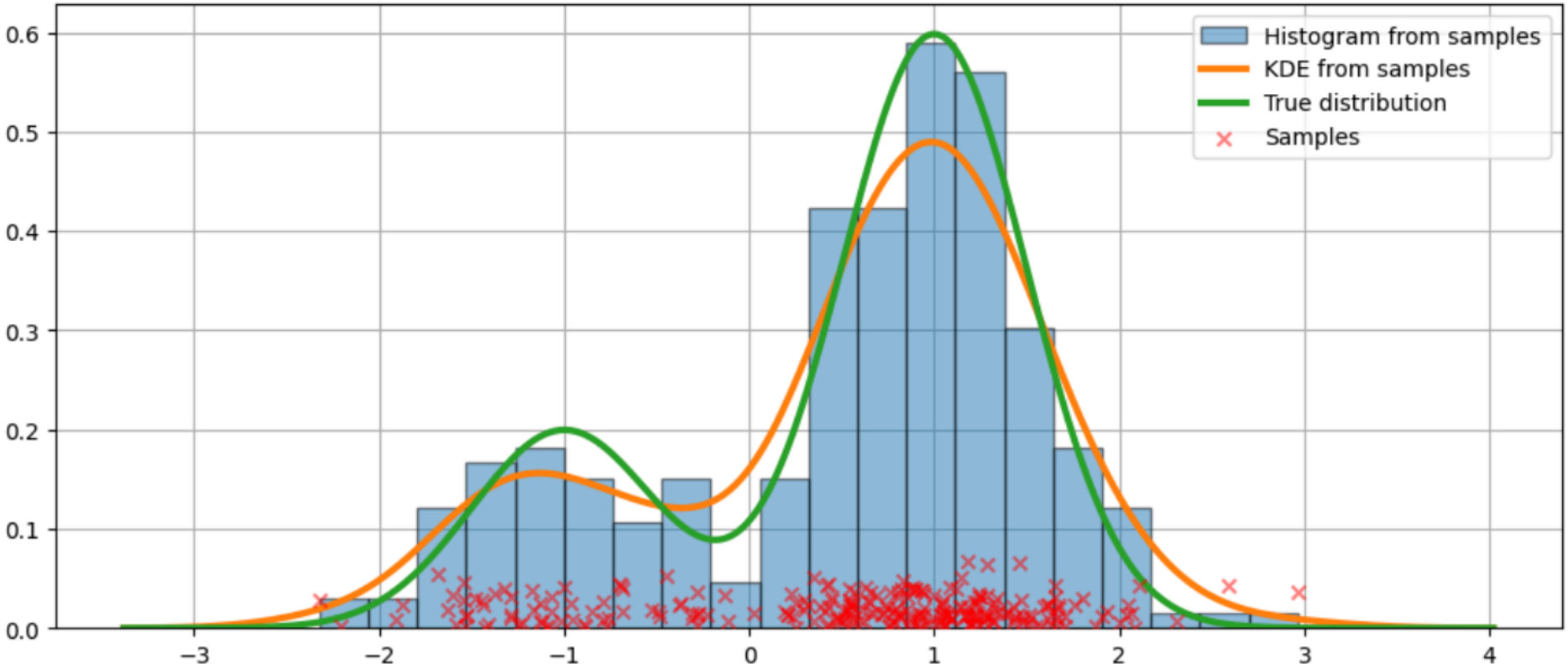


Image URL: `https://www.statsmodels.org/stable/examples/notebooks/generated/kernel_density.html`

A naive method: separate density estimation

Common density estimation methods: (continued)

- Parametric method: MLE
  - Assume the density f belongs to a parametric family $p(\cdot|\boldsymbol{\theta})$
  - $\widehat{\boldsymbol{\theta}} \leftarrow$ MLE (equivalent to minimize the KL divergence or cross-entropy)
  - Use $p(\cdot|\widehat{\boldsymbol{\theta}})$ as the density
  - Can be combined with information criterion like AIC and BIC for model selection $\rightarrow$ model mis-specification is allowed

Other methods:

- Nearest neighbors
- Bayesian method
- ...

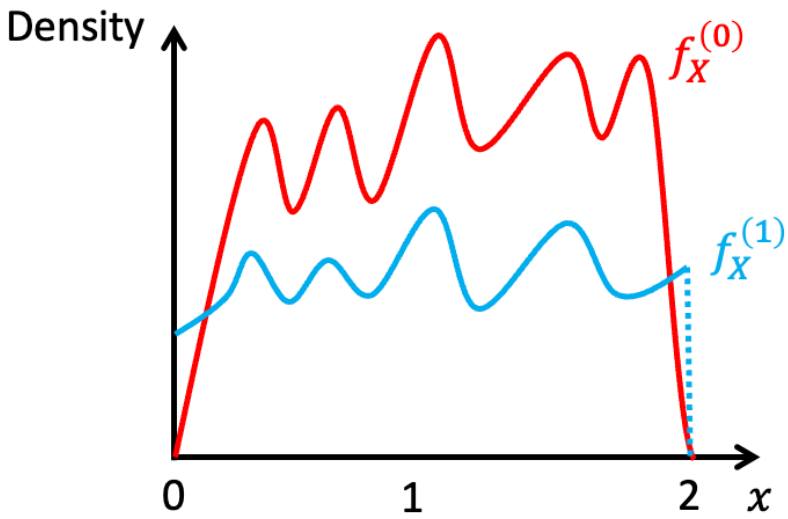
See Sugiyama et al. (2012) for more details.

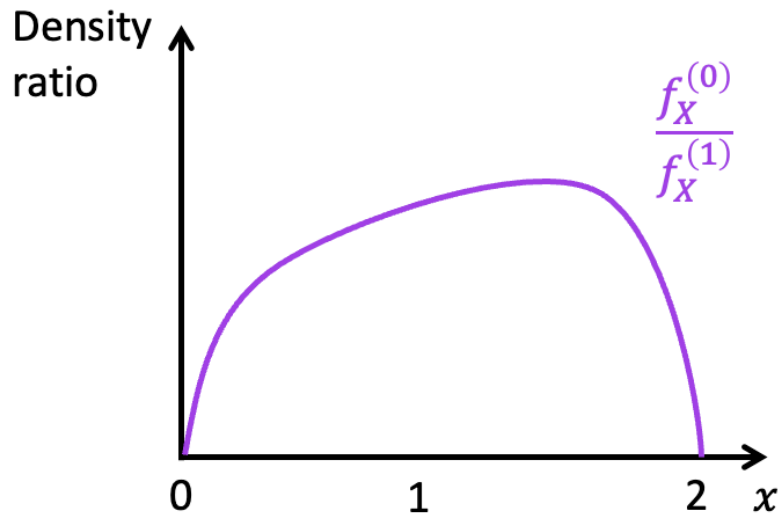
[1] Sugiyama, M., Suzuki, T., & Kanamori, T. (2012). Density ratio estimation in machine learning. Cambridge University Press.

A naive method: separate density estimation

Issues:

- Dividing by a density could inflate the estimation error on the numarator
- Two densities might be very rough themselves but the ratio could be smooth





We know that smoother functions are usually easier to estimate. But the density ratio estimation error can depend on the **worse** smoothness between $f_X^{(0)}$ and $f_X^{(1)}$.

Heuristics: If $f_X^{(1)}$ is bounded away from 0 and ∞, and $f_X^{(0)}$ is bounded away from ∞, then by triangle inequalities:

$$\left|\frac{\hat{f}_X^{(0)}}{\hat{f}_X^{(1)}}(X) - \frac{f_X^{(0)}}{f_X^{(1)}}(X)\right| \lesssim |\hat{f}_X^{(0)}(X) - f_X^{(0)}(X)| + |\hat{f}_X^{(1)}(X) - f_X^{(1)}(X)|$$

A naive method: separate density estimation

Issues: (continued)

- Estimation error of each density depends on the **full dimension**, while the density ratio could live in a much smaller subspace with a smaller **intrinsic dimension**
  Heuristics: $f^{(k)}(\boldsymbol{x}) = \prod_{j=1}^{d} f_j^{(k)}(x_j)$, with $f_j^{(1)} - f_j^{(0)} \equiv 0$ for $j = 2 : d$. Then $f^{(0)}/f^{(1)}(\boldsymbol{x}) = f_1^{(0)}/f_1^{(1)}(x_1)$, where the intrinsic dimension $= 1$.

Solution:

- In practice, for non-parametric methods like KDE, we may choose different kernels and different bandwidth h for $\hat{f}_X^{(0)}$ and $\hat{f}_X^{(1)}$, which usually depend on the smoothness of $\hat{f}_X^{(0)}$ and $\hat{f}_X^{(1)}$
- Can we estimate the density ratio as a whole with the same tuning parameters shared by the numerator and denominator?

§3.4: Density ratio estimation

- §3.4.1 A naive method: separate density estimation
- §3.4.2 Histogram-based method
- §3.4.3 Kernel mean matching
- §3.4.4 Discriminative learning
- §3.4.5 Kullback-Leibler method
- §3.4.6 Semi-parametric method
- §3.4.7 Least square method

Histogram-based method

Recall KDE $\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \frac{1}{nh} \mathbb{1}(\|\boldsymbol{x} - \boldsymbol{x}_i\|_2 \le h)$ with kernel $K(\boldsymbol{x}) = \mathbb{1}(\|x\|_2 \le 1)$.

Goal: estimating $w = f^{(0)}/f^{(1)}$ from $\{\boldsymbol{x}_i^{(k)}\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} f^{(k)}$.

Estimator (Kpotufe, 2017): $\tilde{w}(\boldsymbol{x}) = \frac{\widehat{\mathbb{P}}^{(0)}(B(\boldsymbol{x},r))}{\widehat{\mathbb{P}}^{(1)}(B(\boldsymbol{x},r))} = \frac{n_0^{-1}\sum_{i=1}^{n_0} \mathbb{1}(\|\boldsymbol{x}-\boldsymbol{x}_i^{(0)}\|_2 \le r)}{n_1^{-1}\sum_{i=1}^{n_1} \mathbb{1}(\|\boldsymbol{x}-\boldsymbol{x}_i^{(1)}\|_2 \le r)}$, and

$\hat{w}(\boldsymbol{x}) = \tilde{w}(\boldsymbol{x}) \mathbb{1}\left(\widehat{\mathbb{P}}^{(1)}(B(\boldsymbol{x}, r)) \ge \alpha\right)$

The estimator $\tilde{w}$ is analyzed in Cortes et al. (2008) with discrete $\mathbb{P}_X^{(0)}$ and $\mathbb{P}_X^{(1)}$. Here we present the results in Kpotufe (2017) as it allows general $\mathbb{P}_X^{(0)}$ and $\mathbb{P}_X^{(1)}$ with smoothness conditions.

Assumptions:

- Compact support: $\text{supp}(\mathbb{P}^{(1)})$ is compact in $\mathbb{R}^d$ [1]
- Bounded density ratio: $\|f\|_\infty \le B < \infty$
- β-Hölder class: $|w(\boldsymbol{x}) - w(\boldsymbol{x}')| \le L\|\boldsymbol{x} - \boldsymbol{x}'\|_2^\beta$ with $\beta \in (0, 1]$.

[1] d can be relaxed to a covering number which adapts to the ``intrinsic dimension" of $\text{supp}(\mathbb{P}^{(1)})$

[1] Kpotufe, S. (2017, April). Lipschitz density-ratios, structured data, and data-driven tuning. In Artificial Intelligence and Statistics (pp. 1320-1328). PMLR.

[2] Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008, October). Sample selection bias correction theory. In International conference on algorithmic learning theory (pp. 38-53). Berlin, Heidelberg: Springer Berlin Heidelberg.

Histogram-based method

Theorem 3.4.1 (Kpotufe, 2017)

With $r \asymp \left[\frac{\log(n_0 \wedge n_1)}{n_0 \wedge n_1}\right]^{\beta/(2\beta+d)}$, $\alpha \asymp \log(n_0)/n_0$, *we have*

$$\mathbb{E}_{X \sim \mathbb{P}^{(1)}} |\hat{w}(X) - w(X)| \lesssim_{\mathbb{P}} \left[\frac{\log(n_0 \wedge n_1)}{n_0 \wedge n_1}\right]^{\frac{\beta}{2\beta+d}}.$$

- The rate depends on the choice $r \asymp \left[\frac{\log(n_0 \wedge n_1)}{n_0 \wedge n_1}\right]^{\beta/(2\beta+d)}$
- When β is unknown, a validation-based model selection procedure (akin to the Lepski's method, e.g. Lepski and Spokoiny, 1997) can be used to pick r. See Section 4.2 of Kpotufe (2017). Such a validation-based model selection is commonly used in non-parametrics. See more discussions in Section 5.3 of Wasserman (2006).
- For w with higher-order Hölder conditions, current kernel $K(\boldsymbol{x}) = \mathbb{1}(\|\boldsymbol{x}\|_2 \leq 1)$ is too simple. Local polynomial estimates can be considered.
- Do we have other methods that do not suffer from curse of dimensionality?

[1] Lepski, O. V., & Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. The Annals of Statistics, 25(6), 2512-2546.

[2] Kpotufe, S. (2017, April). Lipschitz density-ratios, structured data, and data-driven tuning. In Artificial Intelligence and Statistics (pp. 1320-1328). PMLR.

[3] Wasserman, L. (2006). All of nonparametric statistics. Springer Science & Business Media.

§3.4: Density ratio estimation

- §3.4.1 A naive method: separate density estimation
- §3.4.2 Histogram-based method
- §3.4.3 Kernel mean matching
- §3.4.4 Discriminative learning
- §3.4.5 Kullback-Leibler method
- §3.4.6 Semi-parametric method
- §3.4.7 Least square method

Kernel mean matching

Moment matching (Methods of moments):

- Goal: estimating the distribution law $\mu \in \mathcal{P}$ [2] from the sample $x_i \overset{\text{i.i.d.}}{\sim} \mu$
- Method: Let $\mathbb{E}_{X\sim\mu} X^r = \mathbb{E}_{X\sim\hat{\mu}} X^r = n^{-1}\sum_{i=1}^n x_i^r$, $r = 1, 2, \cdots$
- Requirement: $\mu \in \mathcal{P}$ can be uniquely identified by the moments

For our density ratio estimation problem: $\{\boldsymbol{x}_i^{(k)}\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} \mathbb{P}^{(k)}$

- $\mathbb{E}_{X\sim\mathbb{P}^{(0)}}[\Phi(X)] = \mathbb{E}_{X\sim\mathbb{P}^{(1)}}[\Phi(X)w(X)]$ for any appropriate Φ
- Maybe we can let $\mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(0)}} X^r = \mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(1)}}[w(X)X^r]$ then solve w?
- Issues:
  - Consider finite $r = 1 : R$: Computable but not able to uniquely identify w
  - Consider $r = 1, 2, \ldots$: Not computable, still cannot uniquely identify w
  - In fact, there exist counter-examples where two different distributions share all the moments (e.g., Chapter 3.15 of Siegel, 2017, Chapter 11 of Stoyanov (2013))

[2] $\mu \in \mathcal{P}$ is usually indexed by a finite-dimensional parameter.

[1] Siegel, A. F. (2017). Counterexamples in probability and statistics. Routledge.
[2] Stoyanov, J. M. (2013). Counterexamples in probability. Courier Corporation.

Kernel mean matching

Goal: Find an ``anchor" functional Φ such that $\mu : \mathbb{P} \mapsto \mathbb{E}_{X\sim\mathbb{P}}\Phi(X)$ is **injective**, i.e. $\mathbb{P} \neq \mathbb{Q} \;\Rightarrow\; \mu(\mathbb{P}) \neq \mu(\mathbb{Q})$

- Intuitively, such a Φ has to be ∞-dimensional
- This can be used to estimate w

Consequence of the injection: If $\tilde{w} : \mathcal{X} \to \mathbb{R}_+$ satisfies $\mathbb{E}_{X\sim\mathbb{P}^{(1)}}\tilde{w}(X) = 1$ and $\mathbb{E}_{X\sim\mathbb{P}^{(1)}}[\tilde{w}(X)\Phi(X)] = \mathbb{E}_{X\sim\mathbb{P}^{(0)}}[\Phi(X)]$, then $\tilde{w} = w$.

In practice, we can replace $\mathbb{P}^{(k)}$ with $\widehat{\mathbb{P}}^{(k)}$ for $k = 0, 1$ and estimate w.

Proof: $\mathbb{E}_{X\sim\tilde{w}\mathbb{P}^{(1)}}[\Phi(X)] = \mathbb{E}_{X\sim\mathbb{P}^{(0)}}[\Phi(X)]$

$\Rightarrow \tilde{w}\mathbb{P}^{(1)} = \mathbb{P}^{(0)} = w\mathbb{P}^{(1)}$

$\Rightarrow \tilde{w} = w$ $\square$

Question: What ``anchor" functional Φ can we use?

Kernel mean matching

Before we talk about the choice of Φ, let us review some basics of *Reproducing Kernel Hilbert Space (RKHS)*.

Definition 3.4.1

- A **kernel** [3] $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric and PSD, where PSD means $\boldsymbol{K}_{n\times n} = \{K(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{i,j=1}^n$ is positive semidefinite for any $\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathcal{X}$.
- For a Hilbert space $\mathcal{H}$ containing functions mapping from $\mathcal{X}$ to $\mathbb{R}$, if $K(\cdot, \boldsymbol{x}) \in \mathcal{H}$ and $\mathcal{H}$ satisfies the following **kernel reproducing property**:

$$\langle h, K(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}} = h(\boldsymbol{x}), \quad \forall h \in \mathcal{H} \text{ and } \forall \boldsymbol{x} \in \mathcal{X},$$

  then we call $\mathcal{H}$ as an **RKHS** induced by K.
- $\Phi(\boldsymbol{x}) = K(\cdot, \boldsymbol{x})$ is often called a **feature map** $(\mathcal{X} \to \mathbb{R})$
- Such an RKHS $\mathcal{H}$ is unique and $\mathcal{H} = \overline{\text{span}(\{\Phi(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\})}$

[3] This is different from the "kernel" in KDE we discussed before.

Kernel mean matching

Definition 3.4.2 (Steinwart, 2001)

A kernel K is called **universal** if both of the followings hold:

- $\mathcal{X}$ is compact
- The associated RKHS $\mathcal{H}$ is **dense** in the space of continuous functions on $\mathcal{X}$, i.e., for any $f : \mathcal{X} \to \mathbb{R}$ continuous and $\epsilon > 0$, $\exists h \in \mathcal{H}$ s.t.

$$\sup_{\boldsymbol{x} \in \mathcal{X}} |f(\boldsymbol{x}) - h(\boldsymbol{x})| \leq \epsilon.$$

It turns out that a universal kernel leads to a valid ``anchor" functional Φ.

Theorem 3.4.3 (Huang et al., 2006; Gretton et al., 2008)

When $\Phi(\boldsymbol{x}) = K(\cdot, \boldsymbol{x})$ is a feature map of the RKHS induced by a universal kernel K, $\mu : \mathbb{P} \mapsto \mathbb{E}_{X \sim \mathbb{P}} \Phi(X)$ is injective, i.e. $\mathbb{P} \neq \mathbb{Q} \Rightarrow \mu(\mathbb{P}) \neq \mu(\mathbb{Q})$.

---

[1] Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. Journal of machine learning research, 2(Nov), 67-93.

[2] Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., & Smola, A. (2006). Correcting sample selection bias by unlabeled data. Advances in neural information processing systems, 19.

[3] Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., & Schölkopf, B. (2008). Covariate shift by kernel mean matching.

Kernel mean matching

Main conclusion: The feature map Φ associated with universal kernels is a good ``anchor" functional.

Examples of universal kernels: $\mathcal{X}$ compact $\subset \mathbb{R}^d$

- Gaussian kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{ - \frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{2\sigma^2} \right\}$
- Exponential (Laplace) kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{ - \frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2}{\sigma} \right\}$
- $K(\boldsymbol{x}, \boldsymbol{x}') = \exp\{\langle \boldsymbol{x}, \boldsymbol{x}' \rangle\}$
- ..... (see Steinwart (2001))

Remark: Universal kernels are defined on compact $\mathcal{X}$. There are relaxations of the current universality definition which allow non-compact $\mathcal{X}$. E.g., see Sriperumbudur et al. (2011).

---

[1] Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. Journal of machine learning research, 2(Nov), 67-93.

[2] Sriperumbudur, B. K., Fukumizu, K., & Lanckriet, G. R. (2011). Universality, Characteristic Kernels and RKHS Embedding of Measures. Journal of Machine Learning Research, 12(7).

Kernel mean matching

We will use the fact that $\mathbb{E}_{X\sim\mathbb{P}^{(1)}}[\tilde{w}(X)\Phi(X)] = \mathbb{E}_{X\sim\mathbb{P}^{(0)}}[\Phi(X)]$ and $\mathbb{E}_{X\sim\mathbb{P}^{(1)}}\tilde{w}(X) = 1 \ \Rightarrow \tilde{w} = w$ to estimate w in practice. We call this procedure **kernel mean matching (KMM)**.

Practical algorithm: Requires $\{\boldsymbol{x}_i^{(k)}\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} \mathbb{P}^{(k)}$, $\epsilon, B > 0$ as input. Solve

$$\min_{\boldsymbol{w}} \left\| \frac{1}{n_1}\sum_{i=1}^{n_1} w_i \Phi(\boldsymbol{x}_i^{(1)}) - \frac{1}{n_0}\sum_{i=1}^{n_0} \Phi(\boldsymbol{x}_i^{(0)}) \right\|_{\mathcal{H}}^2$$

$$\text{s.t. } w_i \in [0, B], \left| \frac{1}{n_1}\sum_{i=1}^{n_1} w_i - 1 \right| \le \epsilon.$$

- By reproducing property, $\| \frac{1}{n_1}\sum_{i=1}^{n_1} w_i \Phi(\boldsymbol{x}_i^{(1)}) - \frac{1}{n_0}\sum_{i=1}^{n_0} \Phi(\boldsymbol{x}_i^{(0)}) \|_{\mathcal{H}}^2 = \frac{1}{n_1^2}\boldsymbol{w}^\top \boldsymbol{K}\boldsymbol{w} - \frac{2}{n_1}\boldsymbol{\kappa}^\top \boldsymbol{w} + \text{constant}$, where $\kappa_i = \frac{1}{n_0}\sum_{j=1}^{n_0} K(\boldsymbol{x}_j^{(0)}, \boldsymbol{x}_i^{(1)})$. Hence it is a **quadratic programming (QP)** problem.

Kernel mean matching: theory

Recall that our goal is to use $\{\hat{w}_i\}_{i=1}^{n_1}$ to do ERM on source data

$$\hat{h} = \arg\min_{h \in \mathcal{H}'} \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{w}_i \ell(h(\boldsymbol{x}_i^{(1)}), y_i^{(1)}),$$

and then bound the excess risk

$$\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}} \ell(\hat{h}(X), Y) - \min_{h \in \mathcal{H}'} \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}} \ell(h(X), Y),$$

where $\mathcal{H}'$ is some hypothesis class.

Gap: However, the initial theoretical study of KMM either tries to bound $\| \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{w}_i \Phi(\boldsymbol{x}_i^{(1)}) - \frac{1}{n_0} \sum_{i=1}^{n_0} \Phi(\boldsymbol{x}_i^{(0)}) \|_{\mathcal{H}}$ (Gretton et al., 2008) or bound $| \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{w}_i y_i^{(1)} - \mathbb{E} Y^{(0)} |$ (Yu and Szepesvári, 2012).

[1] Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., & Schölkopf, B. (2008). Covariate shift by kernel mean matching.

[2] Yu, Y., & Szepesvári, C. (2012). Analysis of kernel mean matching under covariate shift. arXiv preprint arXiv:1206.4650.

Kernel mean matching: theory

Li et al. (2020) provides a study on the excess risk, but uses a variant of KMM and requires a bit more conditions on the loss function ℓ. However, we can follow their analysis to obtain the following result for KMM.

Goal: Do ERM on source data

$$\hat{h} = \arg\min_{h \in \mathcal{H}'} \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{w}_i \ell(h(\boldsymbol{x}_i^{(1)}), y_i^{(1)}),$$

and then bound the excess risk

$$\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}} \ell(\hat{h}(X), Y) - \min_{h \in \mathcal{H}'} \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}} \ell(h(X), Y).$$

Assumptions:

- Any $h \in \mathcal{H}'$ is indexed by $\theta \in \Theta$ with Θ a bounded subset of $\mathcal{H}$ (the RKHS we used to estimate w)
- $\ell(h(\boldsymbol{x}), y) := \ell(\boldsymbol{x}, y, \theta) = \langle \Upsilon(\boldsymbol{x}, y), \Lambda \rangle_{\mathcal{H}}$ with $\|\Lambda\|_{\mathcal{H}} \leq C < \infty$, for any $\theta \in \Theta$
- $\mathbb{E}_{Y|X=\boldsymbol{x}}[\ell(\boldsymbol{x}, y, \theta)] = \langle \Phi(\boldsymbol{x}), \theta \rangle_{\mathcal{H}}$ for any $\theta \in \Theta$

---

[1] Li, F., Lam, H., & Prusty, S. (2020, June). Robust importance weighting for covariate shift. In International conference on artificial intelligence and statistics (pp. 352-362). PMLR.

Kernel mean matching: theory

Theorem 3.4.4 (Adapted from Theorem 2 in Li et al., 2020)

$$\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}}\ell(\hat{h}(X),Y)-\min_{h\in\mathcal{H}'}\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}}\ell(h(X),Y)\lesssim_{\mathbb{P}} n_0^{-1/2}+n_1^{-1/2}.$$

- Relaxations on the condition that $\ell(\boldsymbol{x},y,\theta)=\langle\Upsilon(\boldsymbol{x},y),\Lambda\rangle_{\mathcal{H}}$ and $\mathbb{E}_{Y|X=\boldsymbol{x}}[\ell(\boldsymbol{x},y,\theta)]=\langle\Phi(\boldsymbol{x}),\theta\rangle_{\mathcal{H}}$ are possible. Then we would introduce an additional **approximation error** on the RHS. This phenomenon is common in kernel ridge regression. E.g., see Yu and Szepesvári (2012) and Smale and Zhou (2007).
- Cortes et al. (2008) also analyzes how KMM affects the target excess risk, under a penalized kernel regression setting. Their analysis depends on the so-called **stability** of the regression algorithm and some stringent conditions on the kernel (strictly definite positive etc.)

[1] Li, F., Lam, H., & Prusty, S. (2020, June). Robust importance weighting for covariate shift. In International conference on artificial intelligence and statistics (pp. 352-362). PMLR.

[2] Yu, Y., & Szepesvári, C. (2012). Analysis of kernel mean matching under covariate shift. arXiv preprint arXiv:1206.4650.

[3] Smale, S., & Zhou, D. X. (2007). Learning theory estimates via integral operators and their approximations. Constructive approximation, 26(2), 153-172.

[4] Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008, October). Sample selection bias correction theory. In International conference on algorithmic learning theory (pp. 38-53). Berlin, Heidelberg: Springer Berlin Heidelberg.

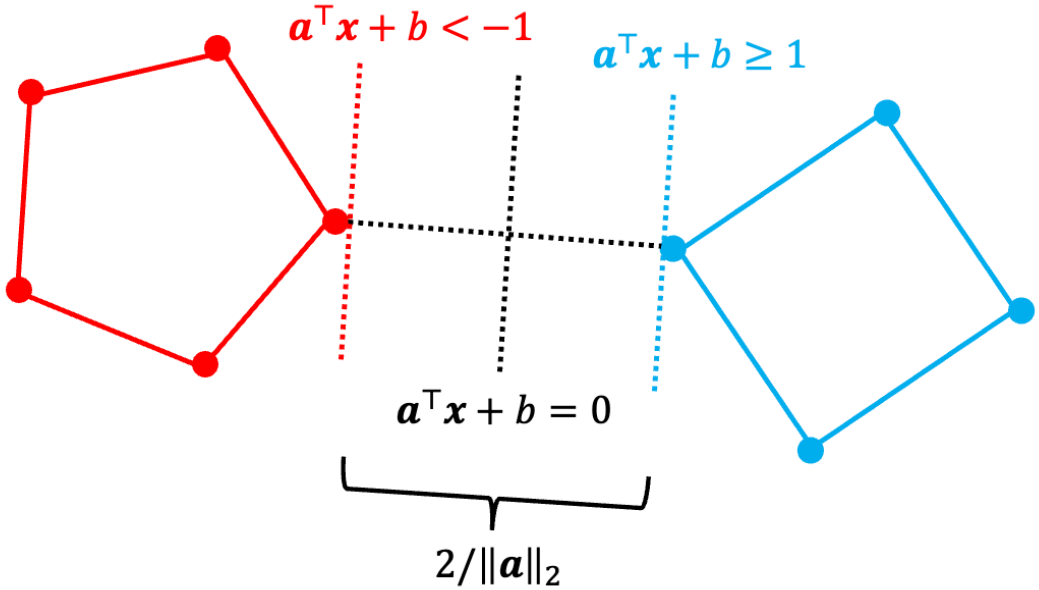
Kernel mean matching: connection to SVM

KMM is connected to **kernelized SVM** and this connection motivates our next density ratio estimation method. Let's review some basics of SVM first.

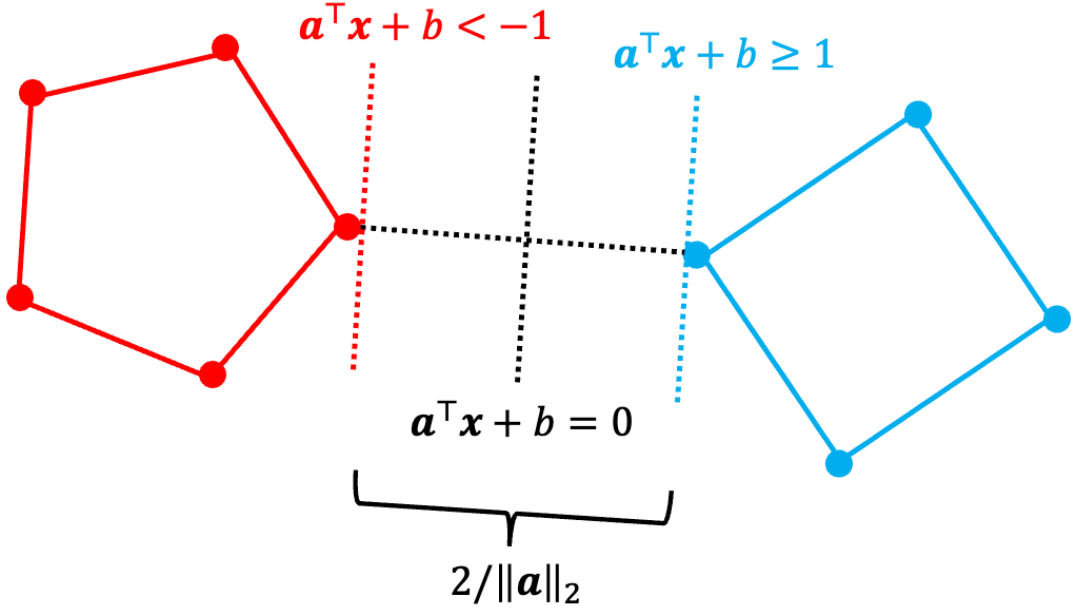
Support vector machine (SVM)

- **Goal:** Use a linear classifier to classify $\{\boldsymbol{x}_i^{(0)}, y_i^{(0)}\}_{i=1}^{n_0}$ and $\{\boldsymbol{x}_i^{(1)}, y_i^{(1)}\}_{i=1}^{n_1}$ with $y_i^{(0)} \equiv -1$ and $y_i^{(0)} \equiv 1$, where $\boldsymbol{x}_i^{(k)} \in \mathbb{R}^d$
- **Linear classifier (separating hyperplane):**

$$h_{\boldsymbol{a},b}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \boldsymbol{a}^\top \boldsymbol{x} + b \geq 0; \\ -1, & \text{if } \boldsymbol{a}^\top \boldsymbol{x} + b < 0. \end{cases}$$



Kernel mean matching: connection to SVM



- **Soft-margin SVM:** $C > 0$ is a constant chosen by users.

$$\begin{aligned}
&\min_{\boldsymbol{a},b,\boldsymbol{u},\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{a}\|_2^2 + C(\mathbf{1}^\top \boldsymbol{u} + \mathbf{1}^\top \boldsymbol{v}) \\
&\text{s.t. } \boldsymbol{a}^\top \boldsymbol{x}_i^{(1)} + b \geq 1 - u_i, \quad i = 1 : n_1 \\
&\qquad \boldsymbol{a}^\top \boldsymbol{x}_i^{(0)} + b \leq -1 + v_i, \quad i = 1 : n_0 \\
&\qquad \boldsymbol{u} \geq \mathbf{0}, \boldsymbol{v} \geq \mathbf{0}, a \in \mathbb{R}^d, b \in \mathbb{R}
\end{aligned}$$

Kernel mean matching: connection to SVM

- **Kernelized soft-margin SVM:** Consider an RKHS $\mathcal{H}$ with kernel K and feature embedding $\Phi : \mathcal{X} \to \mathcal{H}$.

$$\begin{aligned}
\min_{\boldsymbol{a},b,\boldsymbol{u},\boldsymbol{v}} \ & \frac{1}{2}\|\boldsymbol{a}\|_{\mathcal{H}}^2 + C(\mathbf{1}^\top \boldsymbol{u} + \mathbf{1}^\top \boldsymbol{v}) \\
\text{s.t.} \ & \langle \boldsymbol{a}, \Phi(\boldsymbol{x}_i^{(1)}) \rangle_{\mathcal{H}} + b \geq 1 - u_i, \quad i = 1 : n_1 \\
& \langle \boldsymbol{a}, \Phi(\boldsymbol{x}_i^{(0)}) \rangle_{\mathcal{H}} + b \leq -1 + v_i, \quad i = 1 : n_0 \\
& \boldsymbol{u} \geq \mathbf{0}, \boldsymbol{v} \geq \mathbf{0}, a \in \mathcal{H}, b \in \mathbb{R}
\end{aligned}$$

Dual form: $\boldsymbol{K}_{kk} = \{K(\boldsymbol{x}_i^{(k)}, \boldsymbol{x}_j^{(k)})\}_{i,j=1}^{n_k}$, $\boldsymbol{K}_{10} = \{K(\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_j^{(0)})\}_{i=1:n_1, j=1:n_0}$

$$\begin{aligned}
\min_{\boldsymbol{\mu},\boldsymbol{\lambda}} \ & -\mathbf{1}^\top \boldsymbol{\mu} - \mathbf{1}^\top \boldsymbol{\lambda} + \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{K}_{11} \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\lambda}^\top \boldsymbol{K}_{00} \boldsymbol{\lambda} - \boldsymbol{\mu}^\top \boldsymbol{K}_{10} \boldsymbol{\lambda} \\
\text{s.t.} \ & \mathbf{0}_{n_1} \leq \boldsymbol{\mu} \leq C\mathbf{1}_{n_1} \\
& \mathbf{0}_{n_0} \leq \boldsymbol{\lambda} \leq C\mathbf{1}_{n_0}, \\
& \mathbf{1}^\top \boldsymbol{\mu} = \mathbf{1}^\top \boldsymbol{\lambda}.
\end{aligned}$$

Let us connect the dual form to the KMM.

Kernel mean matching: connection to SVM

KMM:

$$\begin{aligned}&\min_{\boldsymbol{w}} \tfrac{1}{n_1^2}\boldsymbol{w}^\top \boldsymbol{K}_{11}\boldsymbol{w} - \tfrac{2}{n_0 n_1}\mathbf{1}_{n_0}^\top \boldsymbol{K}_{01}\boldsymbol{w}\\ &\text{s.t. } w_i \in [0, B], \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n_1} w_i = 1\end{aligned}$$

SVM:

$$\begin{aligned}&\min_{\boldsymbol{\mu},\boldsymbol{\lambda}} -\mathbf{1}_{n_1}^\top \boldsymbol{\mu} - \mathbf{1}_{n_0}^\top \boldsymbol{\lambda} + \tfrac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{K}_{11}\boldsymbol{\mu} + \tfrac{1}{2}\boldsymbol{\lambda}^\top \boldsymbol{K}_{00}\boldsymbol{\lambda} - \boldsymbol{\mu}^\top \boldsymbol{K}_{10}\boldsymbol{\lambda}\\ &\text{s.t. } \mathbf{0}_{n_1} \le \boldsymbol{\mu} \le C\mathbf{1}_{n_1};\\ &\qquad \mathbf{0}_{n_0} \le \boldsymbol{\lambda} \le C\mathbf{1}_{n_0};\\ &\qquad \mathbf{1}_{n_1}^\top \boldsymbol{\mu} = \mathbf{1}_{n_0}^\top \boldsymbol{\lambda}.\end{aligned}$$

- Let $\boldsymbol{\lambda} = \frac{1}{n_0}\mathbf{1}_{n_0}$, $\boldsymbol{\mu} = \frac{1}{n_1}\boldsymbol{w}$, $C \to +\infty$: two problems are equivalent
- This implies that KMM tends to distinguish the target and source domains by minimizing the violation of linear constraints in the embedding space
- More discussions can be found in Gretton et al. (2008) and Bickel et al. (2009).
- **Question:** Can we generalize this idea of ``discriminatve learning"?

[1] Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., & Schölkopf, B. (2008). Covariate shift by kernel mean matching.

[2] Bickel, S., Brückner, M., & Scheffer, T. (2009). Discriminative learning under covariate shift. Journal of Machine Learning Research, 10(9).

§3.4: Density ratio estimation

- §3.4.1 A naive method: separate density estimation
- §3.4.2 Histogram-based method
- §3.4.3 Kernel mean matching
- §3.4.4 Discriminative learning
- §3.4.5 Kullback-Leibler method
- §3.4.6 Semi-parametric method
- §3.4.7 Least square method

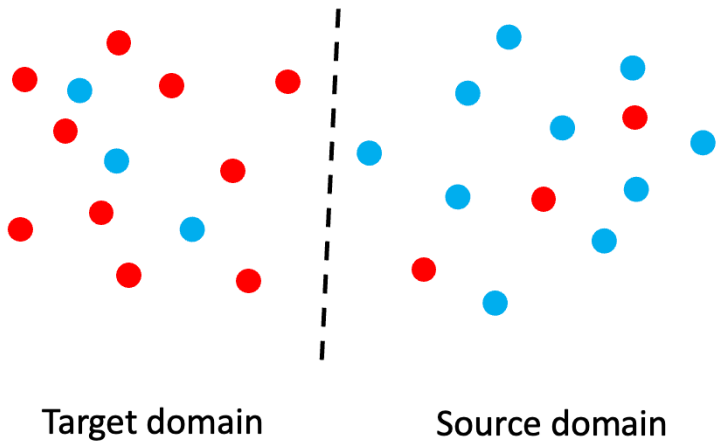
Discriminative learning: a Bayes view

Suppose the data we observe is from a mixture model $\pi_0 \mathbb{P}^{(0)}_{X,Y} + \pi_1 \mathbb{P}^{(1)}_{X,Y}$, with $\pi_k = \mathbb{P}(Z = k)$ and $Z \in \{0, 1\}$ as a latent label of each observation.

- Since $\mathbb{P}^{(0)}_{Y|X} = \mathbb{P}^{(1)}_{Y|X}$: we can view $X \sim \pi_0 \mathbb{P}^{(0)}_X + \pi_1 \mathbb{P}^{(1)}_X$
- By Bayes rule:

$$\frac{\mathrm{d}\mathbb{P}^{(0)}_X}{\mathrm{d}\mathbb{P}^{(1)}_X}(\boldsymbol{x}) = \frac{\pi_1^*}{\pi_0^*} \frac{\mathrm{d}\mathbb{P}^{(0)}_{Z,X}}{\mathrm{d}\mathbb{P}^{(1)}_{Z,X}}(\boldsymbol{x}) = \frac{\pi_1^*}{\pi_0^*} \cdot \frac{\mathbb{P}(Z = 0|X = \boldsymbol{x})}{\mathbb{P}(Z = 1|X = \boldsymbol{x})}.$$

- With data $\{\boldsymbol{x}_i^{(k)}\}_{i=1}^{n_k}$, we can approximate π_k^* with $\hat{\pi}_k = n_k/(n_0 + n_1)$, $k = 0, 1$. It suffices to estimate the propensity score $\mathbb{P}(Z = 1|X = \boldsymbol{x})$.
- This idea generalizes the intuition we saw before from the connection between KMM and SVM: ``Classifying the target and source domains"



Discriminative learning

In practice, we can follow a two-step **discriminative learning** method (Bickel et al., 2007, 2009):

- Step 1: Learn the propensity score $\mathbb{P}(Z=k|X=\boldsymbol{x})$ by your favorite model $\Rightarrow \widehat{\mathbb{P}}(Z=k|X=\boldsymbol{x})$
- Step 2: Use the weight $\hat{w}(\boldsymbol{x}) = \frac{\hat{\pi}_1}{\hat{\pi}_0} \cdot \frac{\widehat{\mathbb{P}}(Z=0|X=\boldsymbol{x})}{\widehat{\mathbb{P}}(Z=1|X=\boldsymbol{x})}$ to reweight the source data then solve the ERM problem

Remarks:

- A similar connection between learning the weight and classifying target/source domains has been studied before in literature for sample selection bias correction (Zadrozny, 2004; Cortes et al., 2008)

[1] Bickel, S., Brückner, M., & Scheffer, T. (2007, June). Discriminative learning for differing training and test distributions. In Proceedings of the 24th international conference on Machine learning (pp. 81-88).

[2] Bickel, S., Brückner, M., & Scheffer, T. (2009). Discriminative learning under covariate shift. Journal of Machine Learning Research, 10(9).

[3] Zadrozny, B. (2004, July). Learning and evaluating classifiers under sample selection bias. In Proceedings of the twenty-first international conference on Machine learning (p. 114).

[4] Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008, October). Sample selection bias correction theory. In International conference on algorithmic learning theory (pp. 38-53). Berlin, Heidelberg: Springer Berlin Heidelberg.

Discriminative learning

Remarks: (continued)

- This connects to the inverse propensity score weighting (IPW) estimator for ATE in causal inference (Horvitz and Thompson, 1952; Ding, 2023):

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{y_i^{(1)}}{\widehat{\mathbb{P}}(Z=1|X=\boldsymbol{x}_i^{(1)})}-\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{y_i^{(0)}}{\widehat{\mathbb{P}}(Z=0|X=\boldsymbol{x}_i^{(0)})}.$$

  - IPW estimator corrects the selection bias for both treatment and control groups.
  - Discriminative learning corrects the bias in the source and the target serves as the reference

- The region of $\boldsymbol{x}$ where $\widehat{\mathbb{P}}(Z=1|X=\boldsymbol{x})$ is close to 1 is dangerous.
- Overfitting might be an issue due to the reason above. We can use a separate sample to estimate the propensity score, but then we need to ``interpolate"/``extrapolate" (originally we only need the propensity scores at those source observations)

---

[1] Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association, 47(260), 663-685.

[2] Ding, P. (2023). A first course in causal inference. arXiv preprint arXiv:2305.18793.

Discriminative learning

Integrated model: Bickel et al. (2007, 2009) also propose to estimate the $Y|X = \boldsymbol{x}$ model and propensity score in one shot (instead of the previous two-step procedure ``weight estimating - reweighted ERM")

- They consider the Bayes model

$$p(y, z, \boldsymbol{\theta}, \boldsymbol{v}|\boldsymbol{x}) = \underbrace{p(y|z, \boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{v})}_{\text{main model}} \times \underbrace{p(z|\boldsymbol{x}, \boldsymbol{v})}_{\text{propensity score model}} \times \underbrace{p(\boldsymbol{\theta})p(\boldsymbol{v})}_{\text{priors}},$$

  where $\boldsymbol{\theta}, \boldsymbol{v}$ are parameters and p represents densities. Then they find the maximum a posteriori (MAP) estimator for $\boldsymbol{\theta}, \boldsymbol{v}$.

- Two-step method is easier to use in practice.

---

[1] Bickel, S., Brückner, M., & Scheffer, T. (2007, June). Discriminative learning for differing training and test distributions. In Proceedings of the 24th international conference on Machine learning (pp. 81-88).

[2] Bickel, S., Brückner, M., & Scheffer, T. (2009). Discriminative learning under covariate shift. Journal of Machine Learning Research, 10(9).

Discriminative learning: theory

There are very few discussions on the generalization error of discriminative learning. But we can have the following intuition.

Recall our previous analysis: consider the reweighted ERM on the source data

$$\hat{h} \in \arg\min_{h\in\mathcal{H}} \left\{ n^{-1} \sum_{i=1}^{n} \hat{w}(\boldsymbol{x}_i^{(1)}) \ell(h(\boldsymbol{x}_i^{(1)}), y_i^{(1)}) \right\}.$$

Then if the loss function ℓ is bounded:

$$\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}} \ell(\hat{h}(X), Y) \leq \underbrace{\min_{h\in\mathcal{H}} \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(1)}} [w(X)\ell(h(X), Y)]}_{\text{oracle}} + \underbrace{o_{\mathbb{P}}(1)}_{\text{uniform convergence}} + \underbrace{2C\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}} |\hat{w}(X) - w(X)|}_{\text{cost of estimating the weight}}.$$

When the propensity score is well estimated, $\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}} |\hat{w}(X) - w(X)|$ can also be well bounded. But additional work might be necessary (e.g. how to translate the estimation error of propensity score to the estimation error of w)

§3.4: Density ratio estimation

- §3.4.1 A naive method: separate density estimation
- §3.4.2 Histogram-based method
- §3.4.3 Kernel mean matching
- §3.4.4 Discriminative learning
- §3.4.5 Kullback-Leibler method
- §3.4.6 Semi-parametric method
- §3.4.7 Least square method

Kullback-Leibler method

- It is known that MLE is equivalent to minimizing the KL divergence or cross-entropy between the empirical and underlying true distribution.
- Here we borrow this idea to estimate the weight w. Consider $\tilde{\mathbb{P}}^{(0)} = \hat{w}\mathbb{P}^{(1)}$.

$$\begin{aligned}\mathsf{KL}(\mathbb{P}^{(0)} \| \tilde{\mathbb{P}}^{(0)}) &= \mathbb{E}_{X\sim\mathbb{P}^{(0)}}\left[\log\left(\frac{\mathrm{d}\mathbb{P}^{(0)}}{\hat{w}\mathrm{d}\mathbb{P}^{(1)}}\right)\right] \\ &= \underbrace{\mathbb{E}_{X\sim\mathbb{P}^{(0)}}\left[\log\left(\frac{\mathrm{d}\mathbb{P}^{(0)}}{\mathrm{d}\mathbb{P}^{(1)}}(X)\right)\right]}_{\text{constant}} - \mathbb{E}_{X\sim\mathbb{P}^{(0)}}[\log\hat{w}(X)].\end{aligned}$$

Hence we can estimate w by finding $\hat{w}$ maximizing

$$\mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(0)}}[\log\hat{w}(X)] = n_0^{-1}\sum_{i=1}^{n_0}\log\hat{w}(\boldsymbol{x}_i^{(0)}).$$

- We can consider the estimators of form $\hat{w}(\boldsymbol{x}) = \sum_{r=1}^{b}\alpha_r\varphi_r(\boldsymbol{x})$, where $\{\varphi_r(\boldsymbol{x})\}_{r=1}^{b}$ are the basis functions with $\varphi_r(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathcal{X}$ and $r = 1 : b$.

Kullback-Leibler method

KL importance estimation procedure: (Sugiyama et al., 2007, 2008)

$$\max_{\{\alpha_r\}_{r=1}^b} \sum_{i=1}^{n_0} \log\left(\sum_{r=1}^{b} \alpha_r \varphi_r(\boldsymbol{x}_i^{(0)})\right)$$

$$\text{s.t.} \sum_{i=1}^{n_1}\sum_{r=1}^{b} \alpha_r \varphi_r(\boldsymbol{x}_i^{(1)}) = n_1,\ \alpha_r \geq 0,\ r = 1:b.$$

- Source data $\{\boldsymbol{x}_i^{(1)}\}_{i=1}^{n_1}$ is only used in the constraint.
- The optimization problem is convex.
- The basis functions $\{\varphi_r(\boldsymbol{x})\}_{r=1}^b$ and hyperparameter b can be chosen by cross-validation.
- Sugiyama et al. (2008) considers $b = n_0$ and $\varphi_i(\boldsymbol{x}) = \Phi(\boldsymbol{x}_i^{(0)})(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{x}_i^{(0)})$ with K and Φ being the Gaussian kernel and the associated feature map.

---

[1] Sugiyama, M., Nakajima, S., Kashima, H., Bünau, P., & Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. Advances in neural information processing systems, 20.

[1] Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., Von Bünau, P., & Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. Annals of the Institute of Statistical Mathematics, 60, 699-746.

Kullback-Leibler method: theory

Suppose $n_0 = n_1 = n$ and denote the entire function class as

$$\mathcal{G} = \{\sum_{r=1}^{b} \alpha_r \varphi_{\theta_r}, \varphi_{\theta_r} \in \mathcal{F}, \alpha_r \geq 0, b \geq 1\}.$$

Assumptions:

- $w = \frac{\mathrm{d}\mathbb{P}^{(0)}}{\mathrm{d}\mathbb{P}^{(1)}}$ is bounded away from 0 and $+\infty$
- $\mathbb{E}_{X \sim \mathbb{P}^{(1)}} \varphi_{\theta_r}(X)$ is away from 0 and φ_{θ_r} is bounded a.e., $\forall \varphi_{\theta_r} \in \mathcal{F}$
- The complexity of $\mathcal{G}$ is not too large. (measured by the metric entropy)

Theorem 3.4.1 (Sugiyama et al., 2008)

The generalized Hellinger distance between $\hat{w}$ and w can be bounded as

$$H_{\mathbb{P}^{(1)}}(\hat{w}, w) := [\mathbb{E}_{X \sim \mathbb{P}^{(1)}}(\sqrt{\hat{w}(X)} - \sqrt{w(X)})^2]^{1/2} \lesssim_{\mathbb{P}} n^{-\frac{1}{2+\gamma}} + \textit{approx. error},$$

where $\gamma \in (0, 2)$ is related to the complexity of $\mathcal{G}$.

There is an approximation error because we do not assume $w \in \mathcal{G}$.

Kullback-Leibler method: theory

Main result from last slide:

$$H_{\mathbb{P}^{(1)}}(\hat{w}, w) := [\mathbb{E}_{X\sim\mathbb{P}^{(1)}}(\sqrt{\hat{w}(X)} - \sqrt{w(X)})^2]^{1/2} \lesssim_{\mathbb{P}} n^{-\frac{1}{2+\gamma}} + \text{approx. error},$$

where $\gamma \in (0, 2)$ is related to the complexity of $\mathcal{G}$.

- Sugiyama et al. (2008) also presents a bound for a parametric function class $\mathcal{G}$.
- It is unclear how the bound of $H_{\mathbb{P}^{(1)}}(\hat{w}, w)$ can be translated to the generalization error on the target domain. E.g., as we showed before, an upper bound of the $L_1(\widehat{\mathbb{P}}^{(1)})$-estimation error of w, i.e. $\mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(1)}}|\hat{w}(X) - w(X)|$ can be useful. But

  $$H^2_{\mathbb{P}^{(1)}}(\hat{w}, w) \lesssim \mathbb{E}_{X\sim\mathbb{P}^{(1)}}|\hat{w}(X) - w(X)|,$$

  which seems to say the current result is not strong enough.

Kullback-Leibler method: a different perspective

Nguyen et al. (2010) developed a similar KL-based estimator by the variational form of f-divergence:

$$D_f(\mathbb{P},\mathbb{Q}) := \mathbb{E}_{X\sim\mathbb{Q}}\left[f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\right] = \sup_{g:\mathcal{X}\to\mathbb{R}}\left\{\mathbb{E}_{X\sim\mathbb{P}}[g(X)] - \mathbb{E}_{X\sim\mathbb{Q}}[f^*(g(X))]\right\},$$

where $f^*(x) = \sup_{y\in\mathbb{R}}\{xy - f(y)\}$ is the conjugate function of f, the supremum is taken at $g \in \partial f(\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q})$, and $\partial f^*(x)$ is the subdifferential of f^* at x.

Consequence for KL-divergence: set

$$f(x) = (x\log x)\cdot\mathbb{1}(x>0) + (+\infty)\cdot\mathbb{1}(x\le 0),$$

then

$$\begin{aligned} D_{\mathsf{KL}}(\mathbb{P}^{(0)}\|\mathbb{P}^{(1)}) &= \sup_{g:\mathcal{X}\to\mathbb{R}}\{\mathbb{E}_{X\sim\mathbb{P}^{(0)}}[g(X)] - \mathbb{E}_{X\sim\mathbb{P}^{(1)}}[e^{g(X)-1}]\} \\ &= \sup_{h:\mathcal{X}\to\mathbb{R}_+}\{\mathbb{E}_{X\sim\mathbb{P}^{(0)}}[\log h(X)] - \mathbb{E}_{X\sim\mathbb{P}^{(1)}}[h(X)] + 1\}, \end{aligned}$$

where the supremum is attained at $h = \mathrm{d}\mathbb{P}^{(0)}/\mathrm{d}\mathbb{P}^{(1)}$.

[1] Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory, 56(11), 5847-5861.

Kullback-Leibler method: a different perspective

Theorem 3.4.2 (Nguyen et al. (2010))

$$D_{\mathsf{KL}}(\mathbb{P}^{(0)}\|\mathbb{P}^{(1)}) = \sup_{h:\mathcal{X}\to\mathbb{R}_+} \{\mathbb{E}_{X\sim\mathbb{P}^{(0)}}[\log h(X)] - \mathbb{E}_{X\sim\mathbb{P}^{(1)}}[h(X)] + 1\}$$

where the supremum is attained at $h = \mathrm{d}\mathbb{P}^{(0)}/\mathrm{d}\mathbb{P}^{(1)}$.

This motivates the following KL-based estimator.

A second KL-based estimator of the density ratio: (Nguyen et al., 2010)

$$\hat{w} = \arg\max_{h:\mathcal{X}\to\mathbb{R}_+} \{\mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(0)}}[\log h(X)] - \mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(1)}}[h(X)]\}.$$

In practice, we use a specific function class $\mathcal{G}$ (e.g. RKHS with a universal kernel) to estimate $h : \mathcal{X} \to \mathbb{R}_+$.

[1] Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory, 56(11), 5847-5861.

Kullback-Leibler method: a different perspective

A second KL-based estimator of the density ratio: (Nguyen et al., 2010)

$$\hat{w} = \underset{h:\mathcal{X}\to\mathbb{R}_+}{\arg\max}\{\mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(0)}}[\log h(X)] - \mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(1)}}[h(X)]\}. \qquad (\star)$$

Recall the previous KL-based estimator (Sugiyama et al., 2007, 2008):

$$\max_{h:\mathcal{X}\to\mathbb{R}_+} \mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(0)}}[\log h(X)]$$
$$\text{s.t. } \mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(1)}} h(X) = 1.$$

The second one can be viewed as a Lagrangian form of the previous one by moving the constraint to the objective function.

---

[1] Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory, 56(11), 5847-5861.

[2] Sugiyama, M., Nakajima, S., Kashima, H., Bünau, P., & Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. Advances in neural information processing systems, 20.

[3] Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., Von Bünau, P., & Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. Annals of the Institute of Statistical Mathematics, 60, 699-746.

Kullback-Leibler method: a different perspective

A second KL-based estimator of the density ratio: (Nguyen et al., 2010)

$$\hat{w} = \arg\max_{h \in \mathcal{G}} \{\mathbb{E}_{X \sim \widehat{\mathbb{P}}^{(0)}}[\log h(X)] - \mathbb{E}_{X \sim \widehat{\mathbb{P}}^{(1)}}[h(X)]\}. \qquad (\star)$$

Theorem 3.4.3 (Nguyen et al., 2010)

With similar assumptions as before, and $w \in \mathcal{G}$, we have

$$H_{\mathbb{P}^{(1)}}(\hat{w}, w) := [\mathbb{E}_{X \sim \mathbb{P}^{(1)}}(\sqrt{\hat{w}(X)} - \sqrt{w(X)})^2]^{1/2} \lesssim_{\mathbb{P}} n^{-\frac{1}{2+\gamma}},$$

where $\gamma \in (0, 2)$ is related to the complexity of $\mathcal{G}$.

- They also propose to use the RKHS with Gaussian kernel as $\mathcal{G}$
- By plugging $\hat{w}$ in $(\star)$, we also get an estimator for $D_{\mathsf{KL}}(\mathbb{P}^{(0)} \| \mathbb{P}^{(1)})$:

$$\widehat{D}_{\mathsf{KL}} = \mathbb{E}_{X \sim \widehat{\mathbb{P}}^{(0)}}[\log \hat{w}(X)] - \mathbb{E}_{X \sim \widehat{\mathbb{P}}^{(1)}}[\hat{w}(X)]$$

Nguyen et al. (2010) show that $|\widehat{D}_{\mathsf{KL}} - D_{\mathsf{KL}}| \lesssim_{\mathbb{P}} n^{-1/2}$ under slightly stronger conditions.

[1] Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory, 56(11), 5847-5861.

§3.4: Density ratio estimation

- §3.4.1 A naive method: separate density estimation
- §3.4.2 Histogram-based method
- §3.4.3 Kernel mean matching
- §3.4.4 Discriminative learning
- §3.4.5 Kullback-Leibler method
- §3.4.6 Semi-parametric method
- §3.4.7 Least square method

Semi-parametric method

When deriving the first KL method (Sugiyama et al., 2008), we did not model $\mathbb{P}_X^{(0)}$ and $\mathbb{P}_X^{(1)}$ explicitly. Recall that we considered $\tilde{\mathbb{P}}^{(0)} = \hat{w}\mathbb{P}^{(1)}$, and

$$\begin{aligned}\mathsf{KL}(\mathbb{P}^{(0)} \| \tilde{\mathbb{P}}^{(0)}) &= \mathbb{E}_{X\sim\mathbb{P}^{(0)}}\left[\log\left(\frac{\mathrm{d}\mathbb{P}^{(0)}}{\hat{w}\mathrm{d}\mathbb{P}^{(1)}}\right)\right] \\ &= \underbrace{\mathbb{E}_{X\sim\mathbb{P}^{(0)}}\left[\log\left(\frac{\mathrm{d}\mathbb{P}^{(0)}}{\mathrm{d}\mathbb{P}^{(1)}}(X)\right)\right]}_{\text{constant}} - \mathbb{E}_{X\sim\mathbb{P}^{(0)}}[\log\hat{w}(X)].\end{aligned}$$

We treated the first term as constant and threw it away.

Motivation: However, by replacing the ordinary likelihood and KL divergence with the profile likelihood (Owen, 2001), we are able to model $\mathbb{P}_X^{(k)}$ at the same time and obtain a semi-pamametric estimator.

[1] Owen, A. B. (2001). Empirical likelihood. Chapman and Hall/CRC.

Semi-parametric method

Consider the semi-parametric setting:

$$\{\boldsymbol{x}_i^{(k)}\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} \mathbb{P}^{(k)}, \qquad \frac{\mathrm{d}\mathbb{P}^{(0)}}{\mathrm{d}\mathbb{P}^{(1)}}(\boldsymbol{x}) = g(\boldsymbol{x};\boldsymbol{\theta}),$$

with a **known** link function g.

- **Profile empirical log-likelihood ratio function** (Qin and Lawless, 1994)

$$l = \sup\left\{ \sum_{k=0}^{1}\sum_{i=1}^{n_k} \log p_i^{(k)} + \sum_{i=1}^{n_1} \log g(\boldsymbol{x}_i^{(1)};\boldsymbol{\theta}) : \sum_{k=0}^{1}\sum_{i=1}^{n_k} p_i^{(k)} = 1, p_i^{(k)} \geq 0, \right.$$
$$\left. \sum_{k=0}^{1}\sum_{i=1}^{n_k} p_i^{(k)} g(\boldsymbol{x}_i^{(1)};\boldsymbol{\theta}) = 1 \right\}$$

  The supremum is attained at $p_i^{(k)} = \frac{1}{n}\frac{1}{1+\lambda[g(\boldsymbol{x}_i^{(k)};\boldsymbol{\theta})+1]}$ with Lagrangian multiplier $\lambda \geq 0$. We can plug it into ℓ.

[1] Qin, J., & Lawless, J. (1994). Empirical likelihood and general estimating equations. the Annals of Statistics, 22(1), 300-325.

Semi-parametric method

- **Profile empirical log-likelihood ratio function** (Qin and Lawless, 1994): (continued)

$$l(\boldsymbol{\theta}, \lambda) = -\sum_{k=0}^{1}\sum_{i=1}^{n_k} \log[1 + \lambda(g(\boldsymbol{x}_i^{(k)}; \boldsymbol{\theta}) + 1)] + \sum_{i=1}^{n_1} \log g(\boldsymbol{x}_i^{(1)}; \boldsymbol{\theta}).$$

A semi-parametric estimator: (Qin, 1998)

Step 1: Let $\frac{\partial l}{\partial \boldsymbol{\theta}} = \frac{\partial l}{\partial \lambda} = 0 \quad \Rightarrow \quad (\hat{\boldsymbol{\theta}}, \hat{\lambda})$

Step 2: Estimate the density ratio as $\hat{w}(\boldsymbol{x}) = g(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$

Suppose the true density ratio $w(\boldsymbol{x}) = g(\boldsymbol{x}; \boldsymbol{\theta}^*)$ and $n_0 \asymp n_1$.

Theorem 3.4.1 (Qin, 1998)

Under some regularity conditions on g (e.g. curvature near $\boldsymbol{\theta}$, smoothness, etc.), within the region $\{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq n^{-1/3}\}$, l admits a local maximizer $(\hat{\boldsymbol{\theta}}, \hat{\lambda})$ with

$$\sqrt{n}(\hat{\boldsymbol{\theta}}, \hat{\lambda})^\top \xrightarrow{\mathsf{d}} N(\mathbf{0}, \boldsymbol{W}).$$

[1] Qin, J., & Lawless, J. (1994). Empirical likelihood and general estimating equations. the Annals of Statistics, 22(1), 300-325.
[2] Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. Biometrika, 85(3), 619-630.

Semi-parametric method

Main result from last slide: Within the region $\{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq n^{-1/3}\}$, l admits a local maximizer $(\hat{\boldsymbol{\theta}}, \hat{\lambda})$ with

$$\sqrt{n}(\hat{\boldsymbol{\theta}}, \hat{\lambda})^\top \xrightarrow{\mathsf{d}} N(\mathbf{0}, \boldsymbol{W}).$$

- The asymptotic normality only holds for solutions in the local area around $\boldsymbol{\theta}^*$ (which is reasonable because there is no global shape restrictions on g). A proper initializer might be needed in practice.
- Under certain conditions, we can get

$$\mathbb{E}_{X \sim \widehat{\mathbb{P}}^{(1)}} |g(X; \hat{\boldsymbol{\theta}}) - g(X; \boldsymbol{\theta}^*)| \lesssim \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \lesssim_{\mathbb{P}} n^{-1/2},$$

  which could be translated to a bound of the generalization error on the target domain.
- Qin (1998) also studies a special choice $g(\boldsymbol{x}; \boldsymbol{\theta}) = \exp\{\alpha + \phi(\boldsymbol{x}; \boldsymbol{\beta})\}$ (exponential tilting) and gives similar results.

§3.4: Density ratio estimation

- §3.4.1 A naive method: separate density estimation
- §3.4.2 Histogram-based method
- §3.4.3 Kernel mean matching
- §3.4.4 Discriminative learning
- §3.4.5 Kullback-Leibler method
- §3.4.6 Semi-parametric method
- §3.4.7 Least square method

Least square method

Recall our previous analysis for bounded loss ℓ:

$$\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(0)}}\ell(\hat{h}(X),Y) \le \underbrace{\min_{h\in\mathcal{H}}\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(1)}}[w(X)\ell(h(X),Y)]}_{\text{oracle}} + \underbrace{\mathcal{O}_{\mathbb{P}}(1)}_{\text{uniform convergence}} + \underbrace{2C\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}|\hat{w}(X)-w(X)|}_{\text{cost of estimating the weight}}.$$

Consider $\left(\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}|\hat{w}(X)-w(X)|\right)^2 \le \mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}|\hat{w}(X)-w(X)|^2$. And

$$\begin{aligned}\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}|\hat{w}(X)-w(X)|^2 &= \frac{1}{2}\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}[\hat{w}^2(X)] - \mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}[\hat{w}(X)w(X)] \\ &\quad + \frac{1}{2}\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}[w^2(X)] \\ &\approx \frac{1}{2}\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}[\hat{w}^2(X)] - \mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(0)}}[\hat{w}(X)] \\ &\quad + \underbrace{\frac{1}{2}\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}[w^2(X)]}_{\text{constant}}\end{aligned}$$

Least square method

- Hence we can search over some function class to find $\hat{w}$ that minimizes $\frac{1}{2}\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}[\hat{w}^2(X)] - \mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(0)}}[\hat{w}(X)]$.
- Note that

$$\frac{1}{2}\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}[\hat{w}^2(X)] - \mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(0)}}[\hat{w}(X)] \approx \frac{1}{2}\mathbb{E}_{X\sim\mathbb{P}_X^{(1)}}[\hat{w}^2(X)] - \mathbb{E}_{X\sim\mathbb{P}_X^{(0)}}[\hat{w}(X)].$$

We can connect this estimator with the f-divergence estimator (Nguyen et al., 2010) we discussed before, where

$$\begin{aligned} D_f(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) &:= \mathbb{E}_{X\sim\mathbb{P}^{(1)}}\left[f\left(\frac{\mathrm{d}\mathbb{P}^{(0)}}{\mathrm{d}\mathbb{P}^{(1)}}\right)\right] \\ &= \sup_{g:\mathcal{X}\to\mathbb{R}_+}\left\{-\mathbb{E}_{X\sim\mathbb{P}^{(1)}}[f^*(g(X))] + \mathbb{E}_{X\sim\mathbb{P}^{(0)}}[g(X)]\right\}, \end{aligned}$$

and the solution of the supremum is $g = \mathrm{d}\mathbb{P}^{(0)}/\mathrm{d}\mathbb{P}^{(1)}$.

Take $f^*(x) = \frac{1}{2}x^2$, then two estimators are equivalent. [4]

[4] This choice of f and f^* does not satisfy the f-divergence definition though.

[1] Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory, 56(11), 5847-5861.

Least square method

- Consider the estimators of form $\hat{w}(\boldsymbol{x}) = \sum_{r=1}^{b} \alpha_r \varphi_r(\boldsymbol{x})$, where $\{\varphi_r(\boldsymbol{x})\}_{r=1}^{b}$ are the basis functions with $\varphi_r(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathcal{X}$ and $r = 1 : b$.

Least square method (the constrained version): (Kanamori et al., 2009)

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^b} \frac{1}{2} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \hat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda \mathbf{1}_b^\top \boldsymbol{\alpha}$$
$$\text{s.t. } \boldsymbol{\alpha} \geq \mathbf{0}_b,$$

where $\widehat{H}_{rr'} = \mathbb{E}_{X \sim \widehat{\mathbb{P}}(1)} [\varphi_r(X) \varphi_{r'}(X)]$, $\hat{h}_r = \mathbb{E}_{X \sim \widehat{\mathbb{P}}(0)} \varphi_r(X)$, $\lambda \geq 0$.

- In many cases, Hessian $\widehat{\boldsymbol{H}}$ is strictly postitive-definite, which makes the problem a convex one with reasonable solutions.
- The penalty term $\lambda \mathbf{1}_b^\top \boldsymbol{\alpha}$ might lead to sparse solutions (similar to Lasso penalty)

[1] Kanamori, T., Hido, S., & Sugiyama, M. (2009). A least-squares approach to direct importance estimation. The Journal of Machine Learning Research, 10, 1391-1445.

Least square method

Least square method (the constrained version): (Kanamori et al., 2009)

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^b} \frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \hat{\boldsymbol{h}}^\top\boldsymbol{\alpha} + \lambda \mathbf{1}_b^\top \boldsymbol{\alpha}$$
$$\text{s.t. } \boldsymbol{\alpha} \geq \mathbf{0}_b,$$

where $\widehat{H}_{rr'} = \mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(1)}}[\varphi_r(X)\varphi_{r'}(X)]$, $\hat{h}_r = \mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(0)}}\varphi_r(X)$, $\lambda \geq 0$.

Theorem 3.4.1 (Kanamori et al., 2009)

Under certain conditions and $n_0 \gtrsim n_1^2$

$$J(\hat{\boldsymbol{\alpha}}) \leq J(\boldsymbol{\alpha}^*) + 1/n_1$$

where $J(\boldsymbol{\alpha}) := \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{H}\boldsymbol{\alpha} - \boldsymbol{h}^\top\boldsymbol{\alpha} + \lambda\mathbf{1}_b^\top\boldsymbol{\alpha}$, $\boldsymbol{H}$ *and* $\boldsymbol{h}$ *are population-level counterpart of* $\widehat{\boldsymbol{H}}$ *and* $\hat{\boldsymbol{h}}$, $\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}\geq\mathbf{0}_b} J(\boldsymbol{\alpha})$.

- This result does not immediately lead to estimation error of w.

[1] Kanamori, T., Hido, S., & Sugiyama, M. (2009). A least-squares approach to direct importance estimation. The Journal of Machine Learning Research, 10, 1391-1445.

Least square method

Least square method (the unconstrained version): (Kanamori et al., 2009)

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^b} \frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \hat{\boldsymbol{h}}^\top\boldsymbol{\alpha} + \lambda\|\boldsymbol{\alpha}\|_2^2.$$

where $\widehat{H}_{rr'} = \mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(1)}}[\varphi_r(X)\varphi_{r'}(X)]$, $\hat{h}_r = \mathbb{E}_{X\sim\widehat{\mathbb{P}}^{(0)}}\varphi_r(X)$, $\lambda \geq 0$.

- Explicit solution $\hat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_{b\times b})^{-1}\hat{\boldsymbol{h}}$
- We can truncate $\hat{\boldsymbol{\alpha}}$ coordinate-wisely at 0 to guarantee positivity of the estimated density ratio $\hat{w}(\boldsymbol{x}) = \sum_{r=1}^b \hat{\alpha}_r\varphi_b(\boldsymbol{x})$
- Some theory are derived in Kanamori et al. (2009). See Section 3 in their paper.
- Kanamori et al. (2009) observes that the unconstrained method performed better than the constrained version in practice.

[1] Kanamori, T., Hido, S., & Sugiyama, M. (2009). A least-squares approach to direct importance estimation. The Journal of Machine Learning Research, 10, 1391-1445.

Least square method: a kernelized version

Consider estimating w by functions in an RKHS $\mathcal{H}$ with some kernel K.
Recall that in least square method, we want to find $\hat{w}$ that minimizes

$$\frac{1}{2}\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}[\hat{w}^2(X)] - \mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(0)}}[\hat{w}(X)].$$

We can add a ridge penalty $\lambda\|\hat{w}\|_{\mathcal{H}}^2$. By the representation property of RKHS, it suffices to consider

$$\hat{w}(\boldsymbol{x}) = \sum_{k=0}^{1}\sum_{i=1}^{n_k}\alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i^{(k)}).$$

Least square method (the kernelized version): (Kanamori et al., 2012)

$$\frac{1}{2}\mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(1)}}[\hat{w}^2(X)] - \mathbb{E}_{X\sim\widehat{\mathbb{P}}_X^{(0)}}[\hat{w}(X)] + \lambda\|\hat{w}\|_{\mathcal{H}}^2,$$

where $\lambda \geq 0$.

- An explicit solution exists. See Theorem 1 of Kanamori et al. (2012).
- Upper bound of $\mathbb{E}_{X\sim\mathbb{P}^{(1)}}\|\hat{w} - w\|^2$ can be proved. See Section 3.3 of Kanamori et al. (2012).

[1] Kanamori, T., Suzuki, T., & Sugiyama, M. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. Machine Learning, 86, 335-367.

References I

Bickel, S., Brückner, M., and Scheffer, T. (2007). Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81--88.

Bickel, S., Brückner, M., and Scheffer, T. (2009). Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9).

Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. (2008). Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38--53. Springer.

Ding, P. (2023). A first course in causal inference. *arXiv preprint arXiv:2305.18793*.

Ge, J., Tang, S., Fan, J., Ma, C., and Jin, C. (2024). Maximum likelihood estimation is all you need for well-specified covariate shift. In *The Twelfth International Conference on Learning Representations*.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2008). Covariate shift by kernel mean matching.

References II

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663--685.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. (2006). Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19.

Kanamori, T., Hido, S., and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391--1445.

Kanamori, T., Suzuki, T., and Sugiyama, M. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86:335--367.

Kpotufe, S. (2017). Lipschitz density-ratios, structured data, and data-driven tuning. In *Artificial Intelligence and Statistics*, pages 1320--1328. PMLR.

Lepski, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512--2546.

References III

Li, F., Lam, H., and Prusty, S. (2020). Robust importance weighting for covariate shift. In *International conference on artificial intelligence and statistics*, pages 352--362. PMLR.

Ma, C., Pathak, R., and Wainwright, M. J. (2023). Optimally tackling covariate shift in rkhs-based nonparametric regression. *The Annals of Statistics*, 51(2):738--761.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847--5861.

Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619--630.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *the Annals of Statistics*, 22(1):300--325.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227--244.

References IV

Siegel, A. F. (2017). *Counterexamples in probability and statistics*. Routledge.

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153--172.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2011). Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7).

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67--93.

Stoyanov, J. M. (2013). *Counterexamples in probability*. Courier Corporation.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., and Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

References V

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*, 60:699--746.

Wang, K. (2023). Pseudo-labeling for kernel ridge regression under covariate shift. *arXiv preprint arXiv:2302.10160*.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Yu, Y. and Szepesvári, C. (2012). Analysis of kernel mean matching under covariate shift. *arXiv preprint arXiv:1206.4650*.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114.

Special thanks to

Haotian Lin (Penn State University, Statistics)
Binghe Zhu (Columbia University, Statistics)
Gan Yuan (Columbia University, Statistics)
Haiyan Zheng (University of Bath, Math)

for helpful discussions and providing references