

## Lecture 3: Covariate Shift

Yang Feng<sup>1</sup>, Ye Tian<sup>2</sup>

<sup>1</sup>Department of Biostatistics, School of Global Public Health, New York University

<sup>2</sup>Department of Statistics, Columbia University

May 20, 2024

# 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 §3.5: Other methods
  - §3.5.1 Common issues of existing reweighting methods
  - §3.5.2 Domain-invariant methods
  - §3.5.3 Marginal transfer learning methods
  - §3.5.4 Transformation-based methods
- 6 §3.6: Extended readings
- 7 References

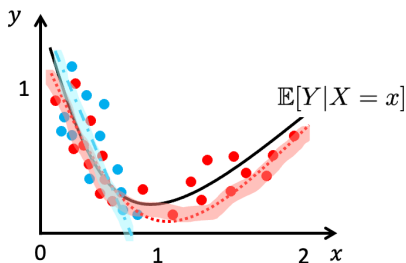
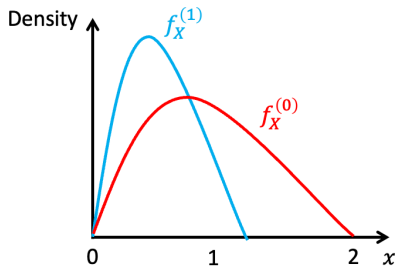
## §3.1: Covariate shift

---

# Covariate shift

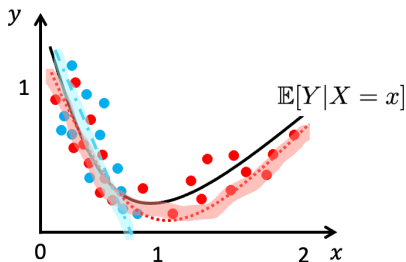
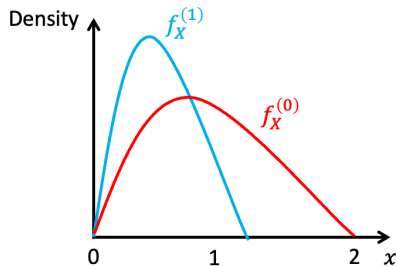
For simplicity, we consider the target and one single source in this section.

- 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



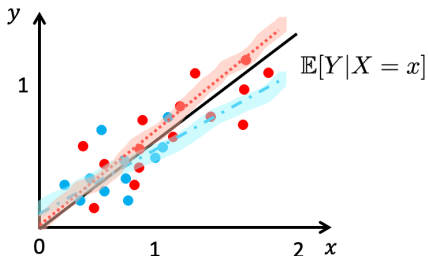
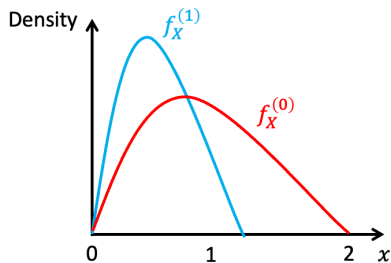
## §3.2: Adaptivity to covariate shift for free?

# 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  $Y = X^\top \theta^* + N(0, \sigma^2)$  for both domains)
  - OLS on the full source data of size  $n \Rightarrow \hat{\theta}$  with  $\|\hat{\theta} - \theta^*\|_2 \lesssim_{\mathbb{P}} n^{-1/2}$ ;
  - Prediction error on the target domain:

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

## Understand this phenomenon from MLE

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

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

and the population-level risk

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

It is easy to see that

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

for any  $\mathbf{x}$ . Hence

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

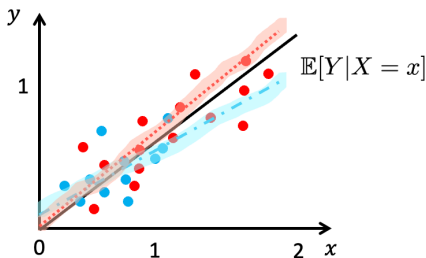
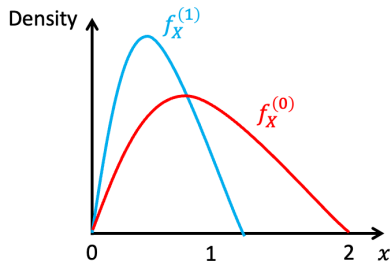
The initial analysis of this phenomenon was conducted in [Shimodaira \(2000\)](#).

---

[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  $Y = X^\top \theta^* + N(0, \sigma^2)$  for both domains)

- (1) OLS on the full source data of size  $n \Rightarrow \hat{\theta}$  with  $\|\hat{\theta} - \theta^*\|_2 \lesssim_{\mathbb{P}} n^{-1/2}$ ;
- (2) Prediction error on target domain:  $\mathbb{E}_{(X,Y) \sim \mathbb{P}^{(0)}} [(X^\top \hat{\theta} - Y)^2] \lesssim_{\mathbb{P}} \sigma^2 + n^{-1}$ .

## Necessary ingredients:

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

# Adaptivity to covariate shift for free

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

## Theorem 3.2.1 (Ge et al. (2024))

Under certain conditions,  $R^{(0)}(\hat{\boldsymbol{\theta}}) - R^{(0)}(\boldsymbol{\theta}^*) \lesssim_{\mathbb{P}} \frac{\text{Tr}(\mathcal{I}^{(0)}(\mathcal{I}^{(1)})^{-1})}{n_1}$ , 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  $Y|X = \mathbf{x}$ : ✓
- Good “curvature” for source risk around  $\boldsymbol{\theta}^*$  (to guarantee recovery of  $\boldsymbol{\theta}^*$ ):

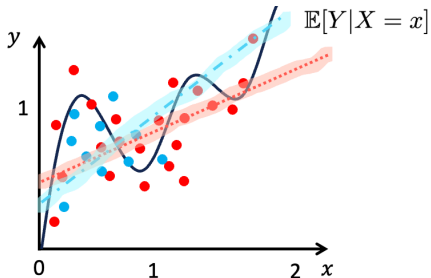
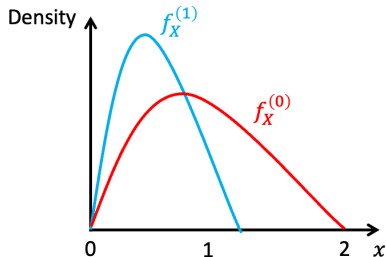
✓, see  $(\mathcal{I}^{(1)})^{-1}$  term

“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”  $\theta^*$ : guarantees a good estimation error of  $\theta^*$ , e.g.  $\|\hat{\theta} - \theta^*\|_2 \lesssim_{\mathbb{P}} \dots$
- The target excess risk can be bounded by the estimation error of  $\theta^*$ , e.g.  $\|\hat{\theta} - \theta^*\|_2$

## Beyond parametric models

There are some recent works on kernel ridge regression (KRR) where similar phenomena 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  $\frac{d\mathbb{P}_X^{(0)}}{d\mathbb{P}_X^{(1)}}$  is bounded,  $\mathbb{E}_{X \sim \mathbb{P}^{(0)}} |\hat{f}^{\text{KRR}}(X) - f^*(X)|^2$  can be well controlled
- [Wang \(2023\)](#) showed that if the regression function  $f^* \in \mathcal{H}$ , even when the density ratio  $\frac{d\mathbb{P}_X^{(0)}}{d\mathbb{P}_X^{(1)}}$  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  $\infty$ -dimensional space
- **No:** With Mercer kernels, RKHS  $\mathcal{H}$  can be embedded into  $\ell_2(\mathbb{N})$ , hence our previous intuitions (estimation error  $\rightarrow$  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.  $\mathcal{X}$
- 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  $R^{(0)}(h)$

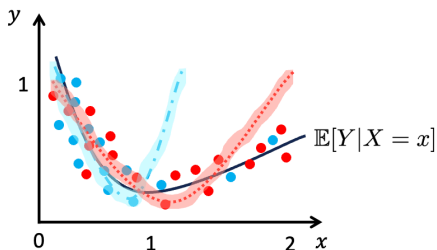
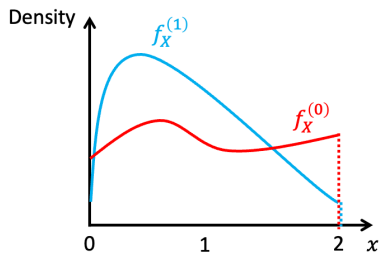
With labeled source data  $\{(\mathbf{x}_i^{(1)}, y_i^{(1)})\}_{i=1}^{n_1}$ , we are able to conduct ERM:

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

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

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

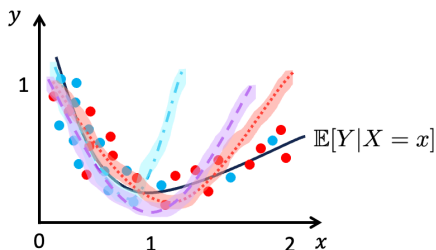
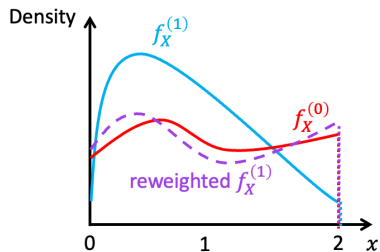
# Motivation



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

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

## Solution: reweighting the loss by density ratio



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

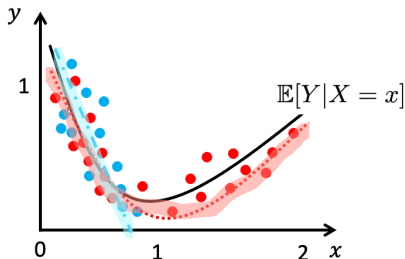
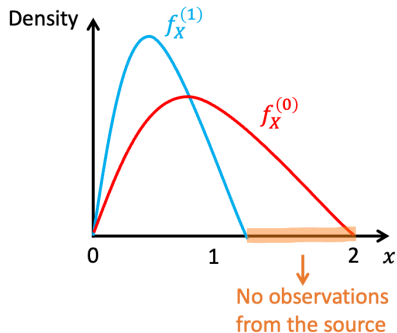
**Rewighted ERM:** ERM on source data with reweighted loss function

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

# 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  $\sigma$  (e.g., Lebesgue) s.t.  $\mathbb{P}^{(0)}, \mathbb{P}^{(1)} \ll \sigma$  and the density ratio  $w(X) = \frac{d\mathbb{P}^{(0)}/d\sigma}{d\mathbb{P}^{(1)}/d\sigma}$  is **bounded**

# Why reweighting works: a simple theoretical justification

Consider the reweighted ERM on the source data

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

Then if the loss function  $\ell$  is bounded:

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

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  $w(\mathbf{x}) = \frac{d\mathbb{P}^{(0)}/d\sigma}{d\mathbb{P}^{(1)}/d\sigma}(\mathbf{x}) := \frac{f^{(0)}}{f^{(1)}}(\mathbf{x})$ .

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

### Common density estimation methods:

- Histogram

- ▷ Divide the space into bins  $B_1, \dots, B_m$
- ▷ Estimate the density as  $\hat{f}(\mathbf{x}) = \sum_{j=1}^m \frac{\#\{i: \mathbf{x}_i \in B_j\}}{n \cdot \text{Vol}(B_j)} \mathbb{1}(\mathbf{x} \in B_j)$

- Kernel density estimation (KDE)

- ▷ Choose a kernel function  $K: \mathbb{R}^p \rightarrow \mathbb{R}^+$
- ▷ Estimate the density as  $\hat{f}(\mathbf{x}) = \sum_{i=1}^n \frac{1}{nh} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$
- ▷ Examples of kernels ( $p = 1$ ):
  - Box kernel:  $K(x) = \mathbb{1}(|x| \leq 1)$
  - Gaussian kernel:  $K(x) = (2\pi)^{-1} \exp\{-x^2/2\}$
  - Epanechnikov kernel:  $K(x) = \frac{3}{4}(1 - x^2)\mathbb{1}(|x| \leq 1)$

See [Wasserman \(2006\)](#) for more details.

---

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

## A naive method: separate density estimation

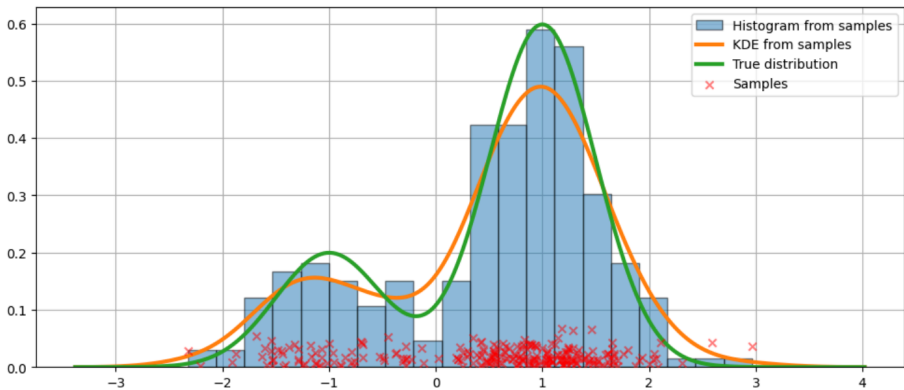


Image URL: [https://www.statsmodels.org/stable/examples/notebooks/generated/kernel\\_density.html](https://www.statsmodels.org/stable/examples/notebooks/generated/kernel_density.html)

# A naive method: separate density estimation

## Common density estimation methods: (continued)

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

## Other methods:

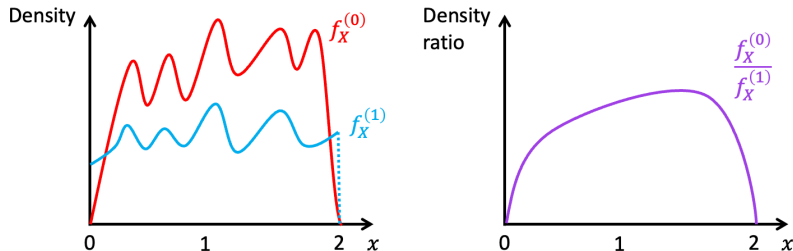
- Nearest neighbors
- Bayesian method
- ...

See [Sugiyama et al. \(2012\)](#) for more details.

# A naive method: separate density estimation

## Issues:

- Dividing by a density could inflate the estimation error on the numerator
- 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  $\infty$ , and  $f_X^{(0)}$  is bounded away from  $\infty$ , 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:  $f^{(k)}(\mathbf{x}) = \prod_{j=1}^d f_j^{(k)}(x_j)$ , with  $f_j^{(1)} - f_j^{(0)} \equiv 0$  for  $j = 2 : d$ . Then  $f^{(0)}/f^{(1)}(\mathbf{x}) = f_1^{(0)}/f_1^{(1)}(x_1)$ , 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  $\hat{f}(\mathbf{x}) = \sum_{i=1}^n \frac{1}{nh} \mathbb{1}(\|\mathbf{x} - \mathbf{x}_i\|_2 \leq h)$  with kernel  $K(\mathbf{x}) = \mathbb{1}(\|\mathbf{x}\|_2 \leq 1)$ .

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

**Estimator** (Kpotufe, 2017):  $\tilde{w}(\mathbf{x}) = \frac{\hat{\mathbb{P}}^{(0)}(B(\mathbf{x}, r))}{\hat{\mathbb{P}}^{(1)}(B(\mathbf{x}, r))} = \frac{n_0^{-1} \sum_{i=1}^{n_0} \mathbb{1}(\|\mathbf{x} - \mathbf{x}_i^{(0)}\|_2 \leq r)}{n_1^{-1} \sum_{i=1}^{n_1} \mathbb{1}(\|\mathbf{x} - \mathbf{x}_i^{(1)}\|_2 \leq r)}$ , and  $\hat{w}(\mathbf{x}) = \tilde{w}(\mathbf{x}) \mathbb{1}(\hat{\mathbb{P}}^{(1)}(B(\mathbf{x}, r)) \geq \alpha)$

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  $\mathbb{R}^d$ <sup>1</sup>
- Bounded density ratio:  $\|f\|_\infty \leq B < \infty$
- $\beta$ -Hölder class:  $|w(\mathbf{x}) - w(\mathbf{x}')| \leq L \|\mathbf{x} - \mathbf{x}'\|_2^\beta$  with  $\beta \in (0, 1]$ .

<sup>1</sup>  $d$  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  $\beta$  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  $K(\mathbf{x}) = \mathbb{1}(\|\mathbf{x}\|_2 \leq 1)$  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  $x_i \stackrel{\text{i.i.d.}}{\sim} \mu$
- Method: Let  $\mathbb{E}_{X \sim \mu} X^r = \mathbb{E}_{X \sim \hat{\mu}} X^r = n^{-1} \sum_{i=1}^n x_i^r$ ,  $r = 1, 2, \dots$
- Requirement:  $\mu \in \mathcal{P}$  can be uniquely identified by the moments

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

- $\mathbb{E}_{X \sim \mathbb{P}^{(0)}} [\Phi(X)] = \mathbb{E}_{X \sim \mathbb{P}^{(1)}} [\Phi(X)w(X)]$  for any appropriate  $\Phi$
- Maybe we can let  $\mathbb{E}_{X \sim \hat{\mathbb{P}}^{(0)}} X^r = \mathbb{E}_{X \sim \hat{\mathbb{P}}^{(1)}} [w(X)X^r]$  then solve  $w$ ?
- Issues:
  - ▷ Consider finite  $r = 1 : R$ : Computable but not able to uniquely identify  $w$
  - ▷ Consider  $r = 1, 2, \dots$ : 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 \(2013\)](#))

---

<sup>2</sup>  $\mu \in \mathcal{P}$  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  $\Phi$  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  $\Phi$  has to be  $\infty$ -dimensional
- This can be used to estimate  $w$

**Consequence of the injection:** If  $\tilde{w} : \mathcal{X} \rightarrow \mathbb{R}_+$  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  $\hat{\mathbb{P}}^{(k)}$  for  $k = 0, 1$  and estimate  $w$ .

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

**Question:** What “anchor” functional  $\Phi$  can we use?

# Kernel mean matching

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

## Definition 3.4.1

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

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

then we call  $\mathcal{H}$  as an **RKHS** induced by  $K$ .

- $\Phi(\mathbf{x}) = K(\cdot, \mathbf{x})$  is often called a **feature map** ( $\mathcal{X} \rightarrow \mathbb{R}$ )
- Such an RKHS  $\mathcal{H}$  is unique and  $\mathcal{H} = \overline{\text{span}(\{\Phi(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\})}$

<sup>3</sup>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  $f : \mathcal{X} \rightarrow \mathbb{R}$  continuous and  $\epsilon > 0$ ,  $\exists h \in \mathcal{H}$  s.t.

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

It turns out that a universal kernel leads to a valid “anchor” functional  $\Phi$ .

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

When  $\Phi(\mathbf{x}) = K(\cdot, \mathbf{x})$  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  $\Phi$  associated with universal kernels is a good “anchor” functional.

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

- Gaussian kernel:  $K(\mathbf{x}, \mathbf{x}') = \exp \left\{ -\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\sigma^2} \right\}$
- Exponential (Laplace) kernel:  $K(\mathbf{x}, \mathbf{x}') = \exp \left\{ -\frac{\|\mathbf{x} - \mathbf{x}'\|_2}{\sigma} \right\}$
- $K(\mathbf{x}, \mathbf{x}') = \exp\{\langle \mathbf{x}, \mathbf{x}' \rangle\}$
- ..... (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  $\{\mathbf{x}_i^{(k)}\}_{i=1}^{n_k} \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}^{(k)}$ ,  $\epsilon, B > 0$  as input. Solve

$$\begin{aligned} \min_{\mathbf{w}} \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} w_i \Phi(\mathbf{x}_i^{(1)}) - \frac{1}{n_0} \sum_{i=1}^{n_0} \Phi(\mathbf{x}_i^{(0)}) \right\|_{\mathcal{H}}^2 \\ \text{s.t. } w_i \in [0, B], \left| \frac{1}{n_1} \sum_{i=1}^{n_1} w_i - 1 \right| \leq \epsilon. \end{aligned}$$

- By reproducing property,  $\left\| \frac{1}{n_1} \sum_{i=1}^{n_1} w_i \Phi(\mathbf{x}_i^{(1)}) - \frac{1}{n_0} \sum_{i=1}^{n_0} \Phi(\mathbf{x}_i^{(0)}) \right\|_{\mathcal{H}}^2 = \frac{1}{n_1^2} \mathbf{w}^\top \mathbf{K} \mathbf{w} - \frac{2}{n_1} \boldsymbol{\kappa}^\top \mathbf{w} + \text{constant}$ , where  $\kappa_i = \frac{1}{n_0} \sum_{j=1}^{n_0} K(\mathbf{x}_j^{(0)}, \mathbf{x}_i^{(1)})$ . 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

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

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

$\left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{w}_i \Phi(\mathbf{x}_i^{(1)}) - \frac{1}{n_0} \sum_{i=1}^{n_0} \Phi(\mathbf{x}_i^{(0)}) \right\|_{\mathcal{H}}$  (Gretton et al., 2008) or bound  $\left| \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{w}_i y_i^{(1)} - \mathbb{E}Y^{(0)} \right|$  (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  $\ell$ . However, we can follow their analysis to obtain the following result for KMM.

**Goal:** Do ERM on source data

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

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

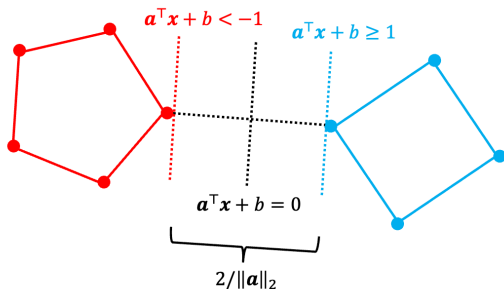
# Kernel mean matching: connection to SVM

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

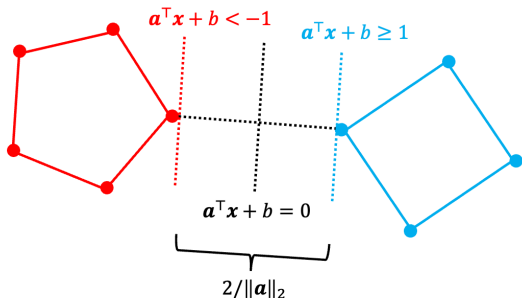
## Support vector machine (SVM)

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

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



## Kernel mean matching: connection to SVM



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

$$\min_{\mathbf{a}, b, \mathbf{u}, \mathbf{v}} \frac{1}{2} \|\mathbf{a}\|_2^2 + C(\mathbf{1}^\top \mathbf{u} + \mathbf{1}^\top \mathbf{v})$$

$$\text{s.t. } \mathbf{a}^\top \mathbf{x}_i^{(1)} + b \geq 1 - u_i, \quad i = 1 : n_1$$

$$\mathbf{a}^\top \mathbf{x}_i^{(0)} + b \leq -1 + v_i, \quad i = 1 : n_0$$

$$\mathbf{u} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0}, \mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$$

## Kernel mean matching: connection to SVM

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

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

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

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

Let us connect the dual form to the KMM.

# Kernel mean matching: connection to SVM

**KMM:**

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

**SVM:**

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

- Let  $\boldsymbol{\lambda} = \frac{1}{n_0} \mathbf{1}_{n_0}$ ,  $\boldsymbol{\mu} = \frac{1}{n_1} \mathbf{w}$ ,  $C \rightarrow +\infty$ : 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 “**discriminative 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

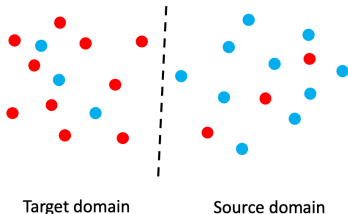
## Discriminative learning: a Bayes view

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

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

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

- With data  $\{\mathbf{x}_i^{(k)}\}_{i=1}^{n_k}$ , we can approximate  $\pi_k^*$  with  $\hat{\pi}_k = n_k / (n_0 + n_1)$ ,  $k = 0, 1$ . It suffices to estimate the **propensity score**  $\mathbb{P}(Z = 1|X = \mathbf{x})$ .
- 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  $\mathbb{P}(Z = k|X = \mathbf{x})$  by your favorite model  
 $\Rightarrow \widehat{\mathbb{P}}(Z = k|X = \mathbf{x})$
- Step 2: Use the weight  $\hat{w}(\mathbf{x}) = \frac{\hat{\pi}_1}{\hat{\pi}_0} \cdot \frac{\widehat{\mathbb{P}}(Z=0|X=\mathbf{x})}{\widehat{\mathbb{P}}(Z=1|X=\mathbf{x})}$  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):

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

- ▷ 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  $\mathbf{x}$  where  $\widehat{\mathbb{P}}(Z = 1|X = \mathbf{x})$  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  $Y|X = \mathbf{x}$  model and propensity score in one shot (instead of the previous two-step procedure “weight estimating - reweighted ERM”)

- They consider the Bayes model

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

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

- 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

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

Then if the loss function  $\ell$  is bounded:

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

When the propensity score is well estimated,  $\mathbb{E}_{X \sim \hat{\mathbb{P}}_X^{(1)}} |\hat{w}(X) - w(X)|$  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  $\tilde{\mathbb{P}}^{(0)} = \hat{w}\mathbb{P}^{(1)}$ .

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

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

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

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

# Kullback-Leibler method

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

$$\begin{aligned} \max_{\{\alpha_r\}_{r=1}^b} \quad & \sum_{i=1}^{n_0} \log \left( \sum_{r=1}^b \alpha_r \varphi_r(\mathbf{x}_i^{(0)}) \right) \\ \text{s.t.} \quad & \sum_{i=1}^{n_1} \sum_{r=1}^b \alpha_r \varphi_r(\mathbf{x}_i^{(1)}) = n_1, \quad \alpha_r \geq 0, \quad r = 1 : b. \end{aligned}$$

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

---

[1] Sugiyama, M., Nakajima, S., Kashima, H., Büna, 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 Büna, 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

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

### Assumptions:

- $w = \frac{d\mathbb{P}^{(0)}}{d\mathbb{P}^{(1)}}$  is bounded away from 0 and  $+\infty$
- $\mathbb{E}_{X \sim \mathbb{P}^{(1)}} \varphi_{\theta_r}(X)$  is away from 0 and  $\varphi_{\theta_r}$  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

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

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  $L_1(\hat{\mathbb{P}}^{(1)})$ -estimation error of  $w$ , i.e.  $\mathbb{E}_{X \sim \hat{\mathbb{P}}^{(1)}} |\hat{w}(X) - w(X)|$  can be useful. But

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

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:

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

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

**Consequence for KL-divergence:** set

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

then

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

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

---

[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))

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

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

This motivates the following KL-based estimator.

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

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

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

---

[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)

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

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

$$\begin{aligned} & \max_{h: \mathcal{X} \rightarrow \mathbb{R}_+} \mathbb{E}_{X \sim \hat{\mathbb{P}}^{(0)}} [\log h(X)] \\ & \text{s.t. } \mathbb{E}_{X \sim \hat{\mathbb{P}}^{(1)}} h(X) = 1. \end{aligned}$$

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., Büna, 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 Büna, 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)

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

### 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  $D_{\text{KL}}(\mathbb{P}^{(0)} \parallel \mathbb{P}^{(1)})$ :

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

Nguyen et al. (2010) show that  $|\hat{D}_{\text{KL}} - D_{\text{KL}}| \lesssim_{\mathbb{P}} n^{-1/2}$  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

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

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-parametric estimator.

---

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

# Semi-parametric method

Consider the semi-parametric setting:

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

with a **known** link function  $g$ .

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

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

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

[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)

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

## A semi-parametric estimator: (Qin, 1998)

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

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

Suppose the true density ratio  $w(\mathbf{x}) = g(\mathbf{x}; \boldsymbol{\theta}^*)$  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*

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

[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

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

- 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

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

which could be translated to a bound of the generalization error on the target domain.

- [Qin \(1998\)](#) also studies a special choice  $g(\mathbf{x}; \boldsymbol{\theta}) = \exp\{\alpha + \phi(\mathbf{x}; \boldsymbol{\beta})\}$  (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  $\ell$ :

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

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

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

## Least square method

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

- Note that

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

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

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

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

Take  $f^*(x) = \frac{1}{2}x^2$ , then two estimators are equivalent. <sup>4</sup>

<sup>4</sup>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  $\hat{w}(\mathbf{x}) = \sum_{r=1}^b \alpha_r \varphi_r(\mathbf{x})$ , where  $\{\varphi_r(\mathbf{x})\}_{r=1}^b$  are the basis functions with  $\varphi_r(\mathbf{x}) \geq 0$  for all  $\mathbf{x} \in \mathcal{X}$  and  $r = 1 : b$ .

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

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

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  $\widehat{\mathbf{H}}$  is strictly positive-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)

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

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

## Theorem 3.4.1 (Kanamori et al., 2009)

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

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

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

◦ 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)

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

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

- Explicit solution  $\widehat{\alpha} = (\widehat{\mathbf{H}} + \lambda \mathbf{I}_{b \times b})^{-1} \widehat{\mathbf{h}}$
- We can truncate  $\widehat{\alpha}$  coordinate-wisely at 0 to guarantee positivity of the estimated density ratio  $\widehat{w}(\mathbf{x}) = \sum_{r=1}^b \widehat{\alpha}_r \varphi_b(\mathbf{x})$
- 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

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

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

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

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

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

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.

## §3.5: Other methods

---

- §3.5.1 Common issues of existing reweighting methods
- §3.5.2 Domain-invariant methods
- §3.5.3 Marginal transfer learning methods
- §3.5.4 Transformation-based methods

## §3.5: Other methods

---

- §3.5.1 Common issues of existing reweighting methods
- §3.5.2 Domain-invariant methods
- §3.5.3 Marginal transfer learning methods
- §3.5.4 Transformation-based methods

## Common issues of existing reweighting methods

- We usually first fit the density ratio then reweight the loss function on the source data to do ERM. The density ratio estimation is often independent of the ERM procedure, which may not be optimal.
- Defining  $\frac{d\mathbb{P}^{(0)}}{d\mathbb{P}^{(1)}}$  requires  $\mathbb{P}^{(0)} \ll \mathbb{P}^{(1)}$ , which implies  $\text{supp}(\mathbb{P}^{(0)}) \subseteq \text{supp}(\mathbb{P}^{(1)})$ . Furthermore, almost all reweighting methods require  $\frac{d\mathbb{P}^{(0)}}{d\mathbb{P}^{(1)}}$  to be **bounded**.

**Question:** Can we go beyond the density ratio estimation and the reweighting method in covariate shift or even more general distribution shift problems?

The answer is **Yes**. We will briefly talk about a few other solutions.

- Domain-invariant methods
- Marginal transfer learning methods
- Transformation-based methods

## §3.5: Other methods

---

- §3.5.1 Common issues of existing reweighting methods
- **§3.5.2 Domain-invariant methods**
- §3.5.3 Marginal transfer learning methods
- §3.5.4 Transformation-based methods

## Domain-invariant methods

- Previously, we considered how to generalize the source domain to the target domain from where we observe some unlabeled data. How can we generalize to an **unseen** target domain?
- Robey et al. (2021) provides a way through **domain-invariant** learners.

Robey et al. (2021) suggested solving the following problem:

Let  $\mathcal{E}_{\text{all}}$  be a set of all **domains** (or **environments**), and  $\mathcal{E}_{\text{train}} \subseteq \mathcal{E}_{\text{all}}$  is a finite subset. From each  $k \in \mathcal{E}_{\text{train}}$ , we observe data  $D^{(k)} = \{\mathbf{x}_i^{(k)}, y_i^{(k)}\}_{i=1}^{n_k}$ . Solve:

$$\hat{h} \in \min_{h \in \mathcal{H}} \max_{k \in \mathcal{E}_{\text{all}}} \mathbb{E}_{(X, Y) \sim \mathbb{P}^{(k)}} [\ell(h(X), Y)].$$

- The maximum over  $k \in \mathcal{E}_{\text{all}}$  guarantees the robustness of the learner  $\hat{h}$ .
- This can be seen a special form of **distributionally robust optimization (DRO)** (Rahimian and Mehrotra, 2019).
- Without further problem structure, finding  $\hat{h}$  is **impossible!**

[1] Robey, A., Pappas, G. J., & Hassani, H. (2021). Model-based domain generalization. Advances in Neural Information Processing Systems, 34, 20210-20229.

[2] Rahimian, H., & Mehrotra, S. (2019). Distributionally robust optimization: A review. arXiv preprint arXiv:1908.05659.

# Domain-invariant methods

## Definition 3.5.1 (Robey et al., 2021)

$G$  is called a **domain transformation** if  $\exists X \sim \mathbb{P}$  such that  $X^{(k)} \sim \mathbb{P}^{(k)}$  satisfies

$$X^{(k)} = G(X, k), \quad \forall k \in \mathcal{E}_{\text{all}}.$$

Given that  $\mathbb{P}_{Y|X}^{(k)} \equiv \mathbb{P}_{Y|X}$ , we have the following equivalence.

## Theorem 3.5.2 (Robey et al., 2021)

*Solving*

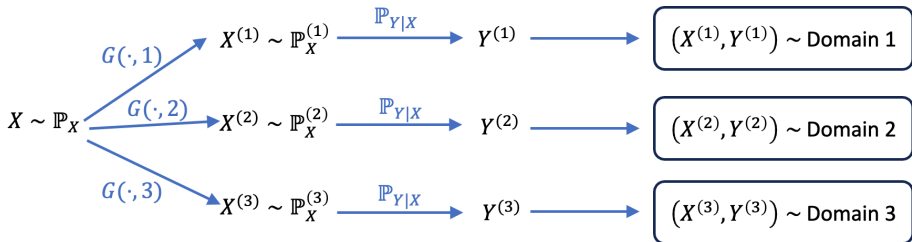
$$\min_{h \in \mathcal{H}} \max_{k \in \mathcal{E}_{\text{all}}} \mathbb{E}_{(X, Y) \sim \mathbb{P}^{(k)}} [\ell(h(X), Y)]$$

*is equivalent to solving*

$$\begin{aligned} & \min_{h \in \mathcal{H}} \mathbb{E}_{(X, Y) \sim \mathbb{P}} [\ell(h(X), Y)] \\ & \text{s.t. } h(\mathbf{x}) = h(G(\mathbf{x}, k)), \text{ a.s. } \mathbf{x} \sim \mathbb{P}, \forall k \in \mathcal{E}_{\text{all}} \end{aligned}$$

[1] Robey, A., Pappas, G. J., & Hassani, H. (2021). Model-based domain generalization. *Advances in Neural Information Processing Systems*, 34, 20210-20229.

# Domain-invariant methods



## Domain-invariant generalization: (Robey et al., 2021)

$$\min_{h \in \mathcal{H}} \mathbb{E}_{(X, Y) \sim \mathbb{P}} [\ell(h(X), Y)]$$

$$\text{s.t. } h(\mathbf{x}) = h(G(\mathbf{x}, k)), \text{ a.s. } \mathbf{x} \sim \mathbb{P}, \forall k \in \mathcal{E}_{\text{all}}$$

- It is not easy to solve...
  - ▷  $G$  and  $\mathbb{P}$  are unknown
  - ▷ We only observe  $\mathcal{E}_{\text{train}} \subseteq \mathcal{E}_{\text{all}}$
  - ▷ The constraint is stringent

See [Robey et al. \(2021\)](#) for details.

[1] Robey, A., Pappas, G. J., & Hassani, H. (2021). Model-based domain generalization. *Advances in Neural Information Processing Systems*, 34, 20210-20229.

## §3.5: Other methods

---

- §3.5.1 Common issues of existing reweighting methods
- §3.5.2 Domain-invariant methods
- **§3.5.3 Marginal transfer learning methods**
- §3.5.4 Transformation-based methods

# Marginal transfer learning methods

- Previously, we know which environment/domain we observe the data from. However, in practice, even this knowledge may not be available.
- [Blanchard et al. \(2011\)](#) considers the following setting:
  - ▷ Each source distribution  $\mathbb{P}^{(k)} \stackrel{\text{i.i.d.}}{\sim} \mu, k = 1 : K$
  - ▷ Each dataset  $\{(\mathbf{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k} \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}^{(k)}$
  - ▷ Target distribution  $\mathbb{P}^{(0)} \sim \mu$ , we observe  $\{\mathbf{x}_i^{(0)}\}_{i=1}^{n_0} \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_X^{(0)}$ , and want to predict  $\{y_i^{(0)}\}_{i=1}^{n_0}$
  - ▷ Goal: Find learner  $h \in \mathcal{H}$  to minimize

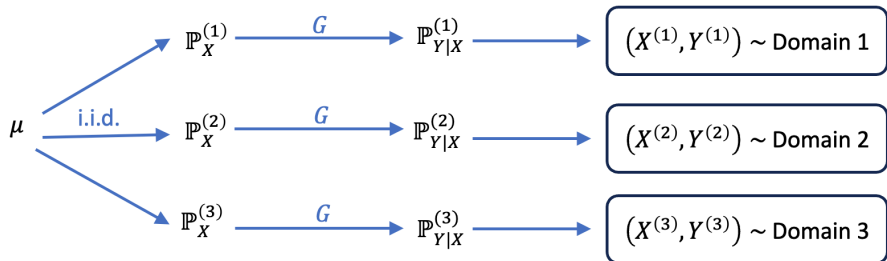
$$\mathbb{E}_{\mathbb{P}^{(0)} \sim \mu} \mathbb{E}_{\{\mathbf{x}_i^{(0)}\}_{i=1}^{n_0} \sim (\mathbb{P}_X^{(0)})^{\otimes n_0}} \left[ \frac{1}{n_0} \sum_{i=1}^{n_0} \ell(h(\mathbf{x}_i^{(0)}), y_i^{(0)}) \right].$$

- Without further conditions, this is **impossible** because  $\mathbb{P}_{Y|X}^{(0)}$  can be arbitrary
- It is possible if there is a fixed “route” for each  $\mathbb{P}_{X,Y} \sim \mu: \mathbb{P}_X \mapsto \mathbb{P}_{Y|X}$

---

[1] Blanchard, G., Lee, G., & Scott, C. (2011). Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24.

# Marginal transfer learning methods



- We can assume there exists a mapping  $G : \mathbb{P}_X \rightarrow \mathbb{P}_{Y|X}$  for any  $\mathbb{P}_{X,Y} \sim \mu$
- This includes the covariate shift setting, where  $G$  is a constant mapping
- The marginal  $\mathbb{P}_X$  is helpful in learning  $h$  (a predictor for  $Y$  given  $X$ )
- We can consider learner  $h$  defined on an "augmented feature space" of  $(\mathbb{P}_X, X) \in \mathcal{B}_X \times \mathcal{X}$ , where  $\mathcal{B}_X$  is the set of all prob. distributions on  $\mathcal{X}$

# Marginal transfer learning methods

If there exists a mapping  $G : \mathbb{P}_X \rightarrow \mathbb{P}_{Y|X}$  for any  $\mathbb{P}_{X,Y} \sim \mu$ , then we it is possible to generalize from sources to **any** target domain with unlabeled target data.

## Theorem 3.5.1 (Blanchard et al., 2011, 2021)

The **domain-specific** Bayes risk is theoretically achivable in the sense that

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}^{(0)} \sim \mu} \underbrace{\min_{h'}}_{\text{domain-specific}} \mathbb{E}_{\{\mathbf{x}_i^{(0)}\}_{i=1}^{n_0} \sim (\mathbb{P}_X^{(0)})^{\otimes n_0}} \left[ \frac{1}{n_0} \sum_{i=1}^{n_0} \ell(h'(\mathbf{x}_i^{(0)}), y_i^{(0)}) \right] \\ &= \underbrace{\min_h}_{\text{domain-invariant}} \mathbb{E}_{\mathbb{P}^{(0)} \sim \mu} \mathbb{E}_{\{\mathbf{x}_i^{(0)}\}_{i=1}^{n_0} \sim (\mathbb{P}_X^{(0)})^{\otimes n_0}} \left[ \frac{1}{n_0} \sum_{i=1}^{n_0} \ell(h(\mathbb{P}_X^{(0)}, \mathbf{x}_i^{(0)}), y_i^{(0)}) \right]. \end{aligned}$$

- In general, the LHS  $\leq$  the RHS
- The existence of domain-invariant transform  $G$  is crucial

[1] Blanchard, G., Lee, G., & Scott, C. (2011). Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24.

[2] Blanchard, G., Deshmukh, A. A., Dogan, U., Lee, G., & Scott, C. (2021). Domain generalization by marginal transfer learning. *Journal of machine learning research*, 22(2), 1-55.

# Marginal transfer learning methods

Therefore, it suffices to solve

$$\underbrace{\min_h}_{\text{domain-invariant}} \mathbb{E}_{\mathbb{P}^{(0)} \sim \mu} \mathbb{E}_{\{\mathbf{x}_i^{(0)}\}_{i=1}^{n_0} \sim (\mathbb{P}_X^{(0)})^{\otimes n_0}} \left[ \frac{1}{n_0} \sum_{i=1}^{n_0} \ell(h(\mathbb{P}_X^{(0)}, \mathbf{x}_i^{(0)}), y_i^{(0)}) \right].$$

- We can do an ERM on source data by finding  $h$  to minimize

$$\frac{1}{K} \sum_{k=1}^K \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(h(\hat{\mathbb{P}}_X^{(k)}, \mathbf{x}_i^{(k)}), y_i^{(k)}).$$

- To learn  $h$ , [Blanchard et al. \(2011, 2021\)](#) consider RKHS with universal kernels and a kernel mean embedding  $\mathbb{P}_X \rightarrow \mathbb{E}_{X \sim \mathbb{P}_X} \Phi(X)$ .
- The embedding  $\mathbb{E}_{X \sim \mathbb{P}_X} \Phi(X)$  serves as the input of  $h$ , which makes sense because we have proved that  $\mathbb{P}_X \rightarrow \mathbb{E}_{X \sim \mathbb{P}_X} \Phi(X)$  is **injective** for universal kernels
- [Blanchard et al. \(2011, 2021\)](#) call this method as **marginal transfer learning**

---

[1] Blanchard, G., Lee, G., & Scott, C. (2011). Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24.

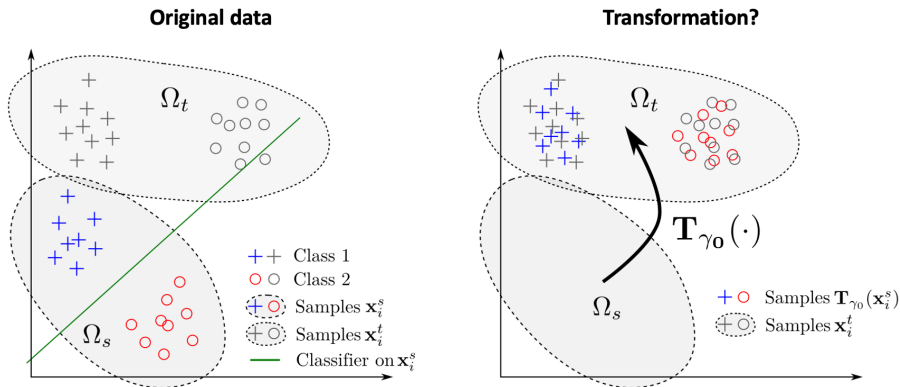
[2] Blanchard, G., Deshmukh, A. A., Dogan, U., Lee, G., & Scott, C. (2021). Domain generalization by marginal transfer learning. *Journal of machine learning research*, 22(2), 1-55.

## §3.5: Other methods

---

- §3.5.1 Common issues of existing reweighting methods
- §3.5.2 Domain-invariant methods
- §3.5.3 Marginal transfer learning methods
- §3.5.4 Transformation-based methods

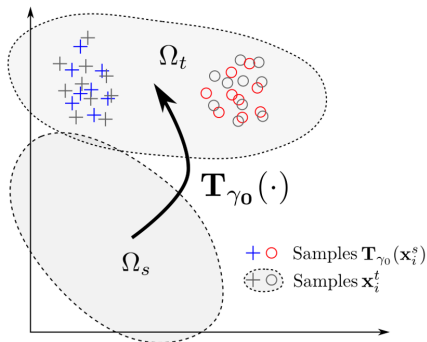
# Intuition



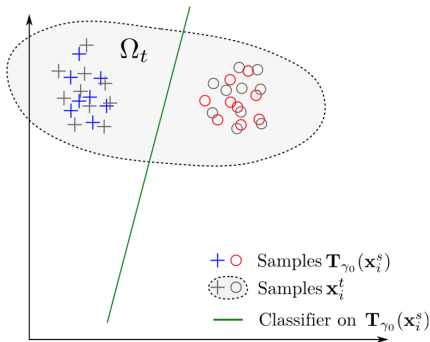
Picture source: Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. IEEE transactions on pattern analysis and machine intelligence, 39(9), 1853-1865.

# Intuition

## Transformation



## Classification after transformation



Picture source: Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. IEEE transactions on pattern analysis and machine intelligence, 39(9), 1853-1865.

# Transformation-based methods

Denote  $\mathcal{X}^{(0)}$ ,  $\mathcal{X}^{(1)}$  as the target and source feature domains, respectively.

**Assumption:**  $\exists$  a transform  $T : \mathcal{X}^{(1)} \rightarrow \mathcal{X}^{(0)}$  s.t.

$$\mathbb{P}_{Y|X}^{(1)}(\cdot|X = \mathbf{x}) = \mathbb{P}_{Y|X}^{(0)}(\cdot|X = T(\mathbf{x})), \quad \mathbb{P}^{(1)\text{-a.e.}} \mathbf{x}.$$

## Domain adaptation based on transformation:

- (i) First learn a transform  $T$  from source feature domain to target feature domain
- (ii) Then apply  $T$  on source training data to get transformed features
- (iii) Use transformed source data (new feature + the original response) to train the model

**Question:** How to learn the transform  $T$ ?

# Transformation-based methods

Courty et al. (2016) proposes to learn it by solving the optimal transport (OT) problem.

**Monge's OT problem:** Given some cost function  $c : \mathcal{X}^{(1)} \times \mathcal{X}^{(0)} \rightarrow \mathbb{R}^+$ , solve

$$\begin{aligned} \min_{T: \mathcal{X}^{(1)} \rightarrow \mathcal{X}^{(0)}} & \int c(\mathbf{x}, T(\mathbf{x})) d\mathbb{P}_X^{(1)}, \\ \text{s.t.} & T(X^{(1)}) \sim \mathbb{P}_X^{(0)}. \end{aligned}$$

- This problem is very difficult to solve.
- The solution may not exist, if  $\mathbb{P}_X^{(0)}$  and  $\mathbb{P}_X^{(1)}$  are discrete.
- Unfortunately, for our use case, we need to consider  $\hat{\mathbb{P}}^{(0)}$  and  $\hat{\mathbb{P}}^{(1)}$  which are indeed discrete.

---

[1] Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. IEEE transactions on pattern analysis and machine intelligence, 39(9), 1853-1865.

# Transformation-based methods

Instead, we can consider a relaxed formulation.

**Kantorovitch's OT problem:** Given some cost function  $c : \mathcal{X}^{(1)} \times \mathcal{X}^{(0)} \rightarrow \mathbb{R}^+$ , consider all joint distributions  $\gamma$  defined on  $\mathcal{X}^{(1)} \times \mathcal{X}^{(0)}$ , solve

$$\min_{\gamma} \int c(\mathbf{x}^{(1)}, \mathbf{x}^{(0)}) d\gamma(\mathbf{x}^{(1)}, \mathbf{x}^{(0)}),$$

$$\text{s.t. } \gamma \text{ has marginals } \mathbb{P}_X^{(1)}, \mathbb{P}_X^{(0)}.$$

- $\gamma$  is called a **transportation plan**
- In practice, we can replace  $\mathbb{P}_X^{(1)}$  and  $\mathbb{P}_X^{(0)}$  with empirical distributions  $\hat{\mathbb{P}}_X^{(1)}$  and  $\hat{\mathbb{P}}_X^{(0)}$ , then the problem becomes a **linear programming** problem:

$$\min_{\Gamma \in (\mathbb{R}^+)^{n_1 \times n_0}} \langle \Gamma, \mathbf{C} \rangle_F,$$

$$\text{s.t. } \Gamma \mathbf{1}_{n_0} = \hat{\mathbb{P}}_X^{(0)}, \Gamma^\top \mathbf{1}_{n_1} = \hat{\mathbb{P}}_X^{(1)}.$$

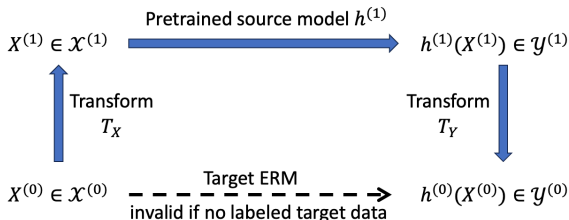
## Transformation-based methods

- After learned the transportation plan  $\Gamma = (\gamma_{i,j})_{n_1 \times n_0}$ , we can use the following barycentric mapping to transform the source data

$$\hat{\mathbf{x}}_i^{(1)} = \arg \min_{\mathbf{x} \in \mathcal{X}^{(1)}} \sum_{j=1}^{n_0} \gamma_{i,j} c(\mathbf{x}, \mathbf{x}_j^{(0)}).$$

For more details, see [Courty et al. \(2016\)](#)

- In fact, the idea of transformation can be extended to general concept drift problem. Instead of transforming from source to target, [Cao et al. \(2023\)](#) proposes a three-step learning diagram.



[1] Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9), 1853-1865.

[2] Cao, H., Gu, H., Guo, X., & Rosenbaum, M. (2023). Feasibility and transferability of transfer learning: A mathematical framework. *arXiv preprint arXiv:2301.11542*.

## §3.6: Extended readings

---

## Beyond bounded density ratio

As we mentioned before, many reweighted ERM methods (more specifically, the weight estimation methods they use) require the density ratio to be **bounded**. However, in many cases, this condition might be unnecessary.

- [Cortes et al. \(2010\)](#) shows that given the true weight, the excess target risk of reweighted ERM can be bounded using R enyi divergence between  $\mathbb{P}_X^{(0)}$  and  $\mathbb{P}_X^{(1)}$ . They show that the excess target risk can be  $\mathcal{O}_{\mathbb{P}}(1)$  even when the weight is unbounded.
- [Kpotufe and Martinet \(2021\)](#) proposes a notion called **transfer exponent**  $\gamma$  in the following sense

$$\mathbb{P}_X^{(1)}(\mathcal{B}(\mathbf{x}, r)) \geq \mathbb{P}_X^{(0)}(\mathcal{B}(\mathbf{x}, r)) \cdot C_\gamma \left( \frac{r}{\Delta} \right)^\gamma, \quad \forall \mathbf{x} \in \text{supp}(\mathbb{P}^{(0)}), \quad \forall r \in (0, \Delta].$$

This notion allows unbounded density ratio and they show that transfer learning is still possible under a non-parametric classification set-up.

---

[1] Cortes, C., Mansour, Y., & Mohri, M. (2010). Learning bounds for importance weighting. *Advances in neural information processing systems*, 23.

[2] Kpotufe, S., & Martinet, G. (2021). Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6), 3299-3323.

# 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).
- Blanchard, G., Deshmukh, A. A., Dogan, U., Lee, G., and Scott, C. (2021). Domain generalization by marginal transfer learning. *Journal of machine learning research*, 22(2):1–55.
- Blanchard, G., Lee, G., and Scott, C. (2011). Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24.
- Cao, H., Gu, H., Guo, X., and Rosenbaum, M. (2023). Feasibility and transferability of transfer learning: A mathematical framework. *arXiv preprint arXiv:2301.11542*.
- Cortes, C., Mansour, Y., and Mohri, M. (2010). Learning bounds for importance weighting. *Advances in neural information processing systems*, 23.

## References II

- 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.
- Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865.
- 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.
- 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.

## References III

- 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.
- Kpotufe, S. and Martinet, G. (2021). Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323.
- Lepski, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546.

## References IV

- 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.
- Rahimian, H. and Mehrotra, S. (2019). Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*.

## References V

- Robey, A., Pappas, G. J., and Hassani, H. (2021). Model-based domain generalization. *Advances in Neural Information Processing Systems*, 34:20210–20229.
- 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.
- 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.

## References VI

- 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.
- Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., Von Büna, 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**