Pseudo-Labeling for Covariate Shift Adaptation

Kaizheng Wang

IEOR & DSI
Columbia University

November 12th 2024

Outline

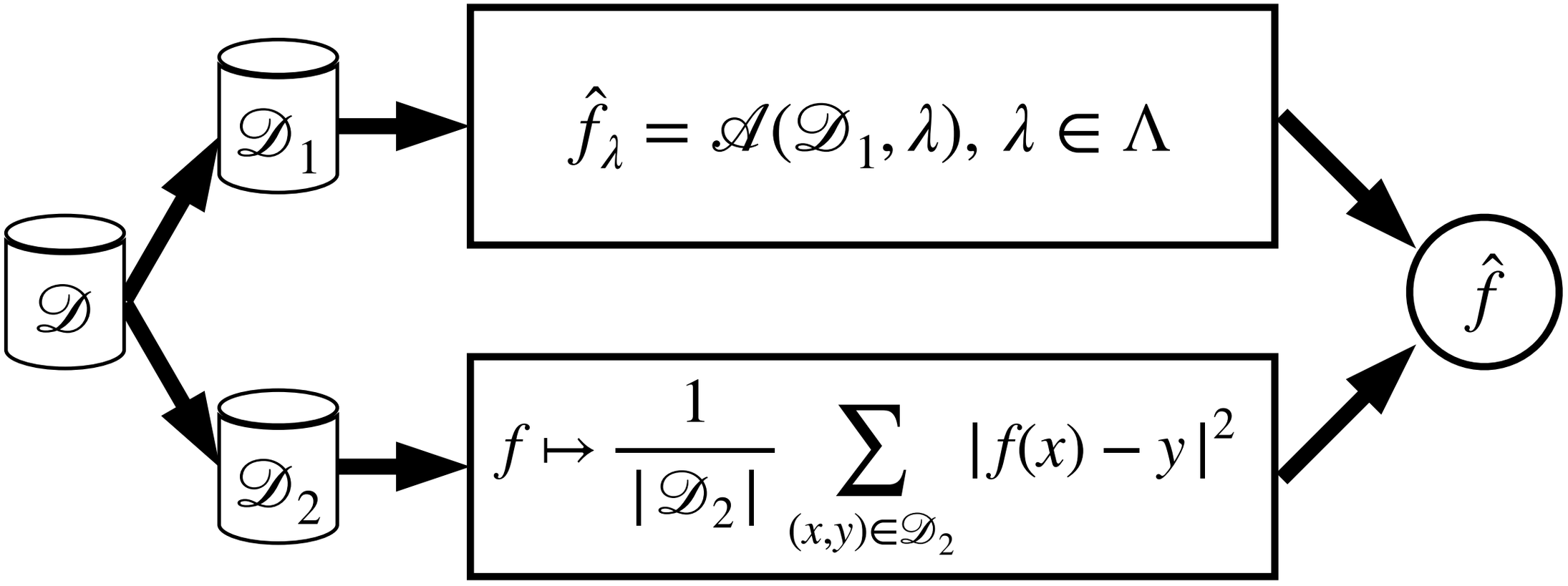
- **Introduction: covariate shift adaptation**
- Methodology: validation with pseudo-labels
- Model selection theory: general results
- Adaptivity guarantees: kernel ridge regression

Regression

- **Distribution**: $(x, y) \sim \mathcal{Q}$ over $\mathcal{X} \times \mathbb{R}$.
- **Risk** (MSE): $R(f) = \mathbb{E}_{(x,y)\sim\mathcal{Q}} |f(x) - y|^2$.
- **Data**: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ i.i.d. from $\mathcal{Q}$.
- **Goal**: learn a model $\hat{f}$ with small risk $R(\hat{f})$.
- **Estimator**: $\hat{f}_\lambda = \mathcal{A}(\mathcal{D}, \lambda)$ with hyperparameter $\lambda \in \Lambda$.

Hold-Out Validation

Training **Candidates**



Validation **Risk estimator**

Needs sufficient labeled data from $\mathscr{Q}$.

Covariate Shift

- **Goal**: learn a predictive model $\hat{f}$ that works well over $\mathcal{Q}$.
- **Issue**: Too costly to collect labeled data from $\mathcal{Q}$.
- **Data**: $\mathcal{D} = \{(x_{P,i}, y_{P,i})\}_{i=1}^{n_P} \sim \mathcal{P}$ and $\{x_{Q,i}\}_{i=1}^{n_Q} \sim \mathcal{Q}_X$.

Covariate Shift

- **Goal**: learn a predictive model $\hat{f}$ that works well over $\mathcal{Q}$.
- **Issue**: Too costly to collect labeled data from $\mathcal{Q}$.
- **Data**: $\mathcal{D} = \{(x_{P,i}, y_{P,i})\}_{i=1}^{n_P} \sim \mathcal{P}$ and $\{x_{Q,i}\}_{i=1}^{n_Q} \sim \mathcal{Q}_X$.

Examples

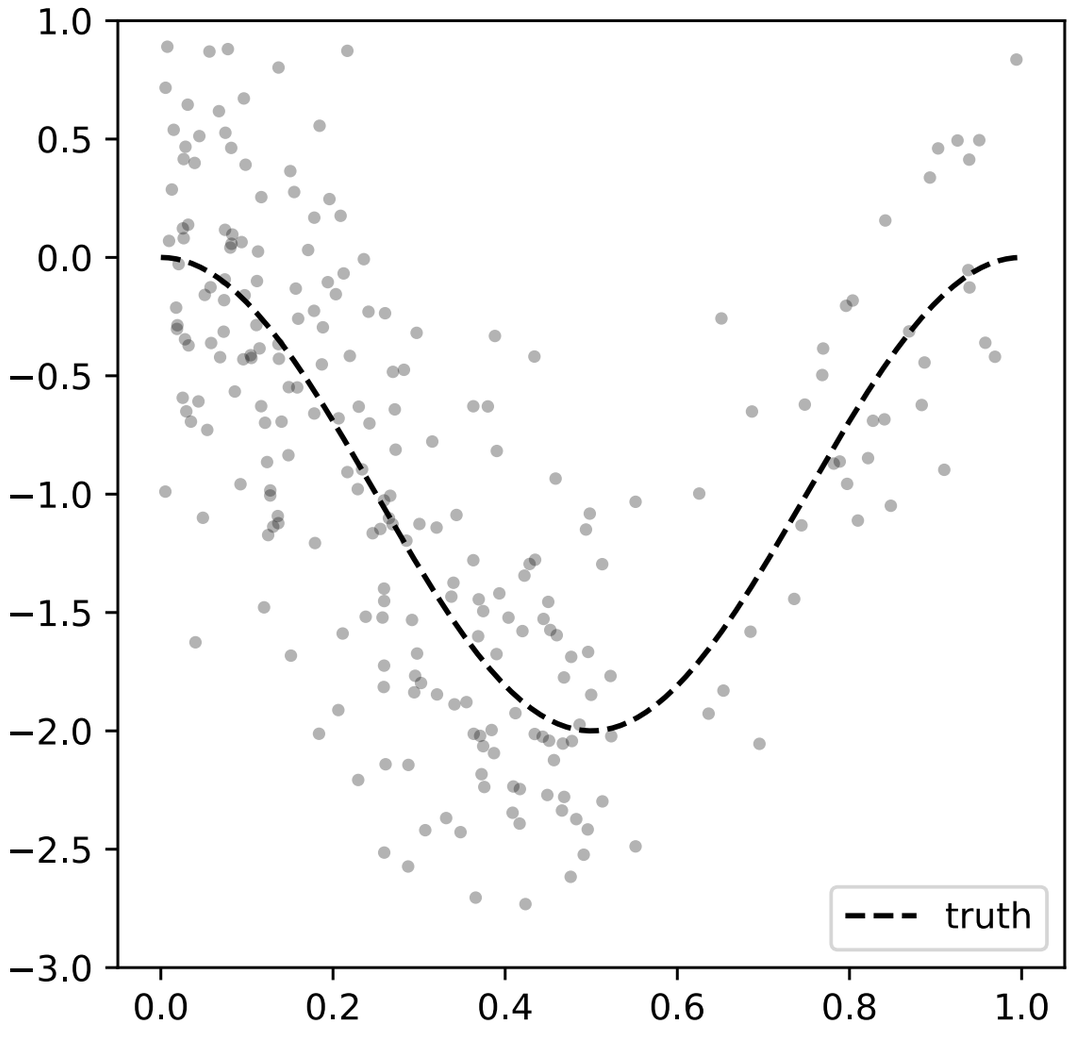
- **Chatbot**: $\mathcal{Q}$ = Business negotiations, $\mathcal{P}$ = News.
- **Self-driving**: $\mathcal{Q}$ = NYC, $\mathcal{P}$ = Berkeley.

Covariate Shift

- **Goal**: learn a predictive model $\hat{f}$ that works well over $\mathcal{Q}$.
- **Issue**: Too costly to collect labeled data from $\mathcal{Q}$.
- **Data**: $\mathcal{D} = \{(x_{P,i}, y_{P,i})\}_{i=1}^{n_P} \sim \mathcal{P}$ and $\{x_{Q,i}\}_{i=1}^{n_Q} \sim \mathcal{Q}_X$.
- $\mathcal{P}_X \neq \mathcal{Q}_X$ but $\mathcal{P}_{Y|X} = \mathcal{Q}_{Y|X}$.
- $\mathcal{P}$, $\mathcal{Q}$ are called **source** and **target**.

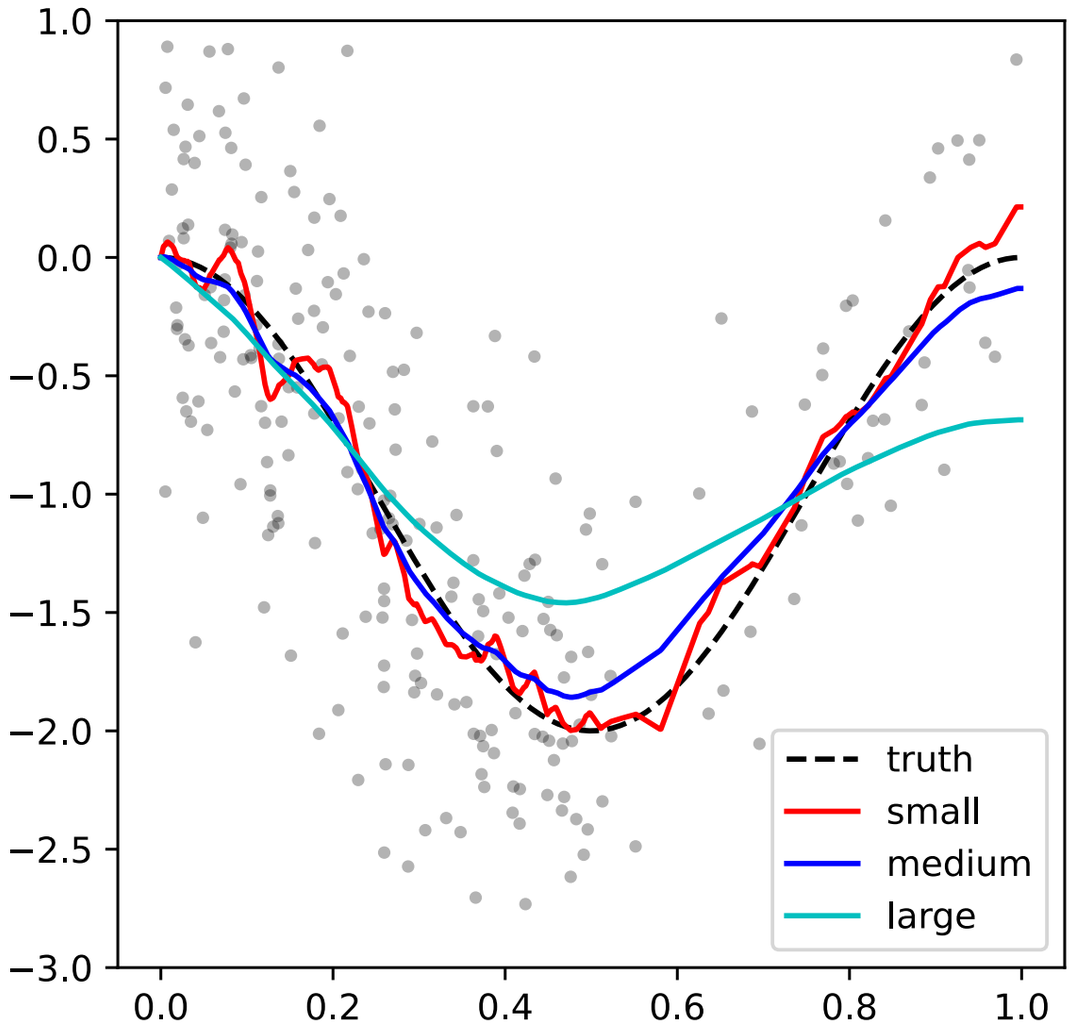
Pan and Yang (2010), Sugiyama and Kawanabe (2012)

Example: Kernel Ridge Regression



$$y = f^{\star}(x) + \varepsilon$$

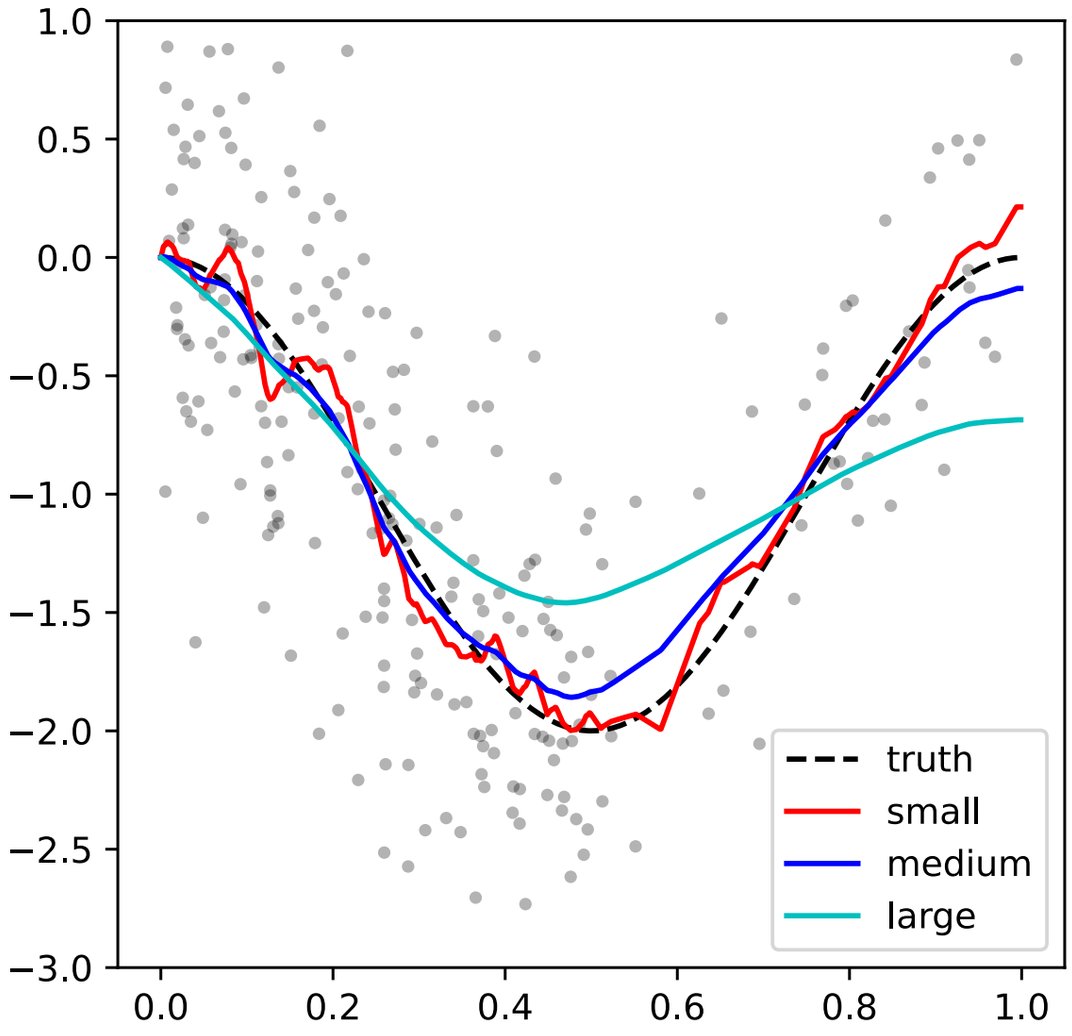
Example: Kernel Ridge Regression



KRR with S, M, L penalties.

Which one is the best?

Example: Kernel Ridge Regression



- If $\mathcal{Q}_X$ concentrates on $[0,0.5]$, **blue** is the best;
- If $\mathcal{Q}_X$ concentrates on $[0.5,1]$, **red** is the best.

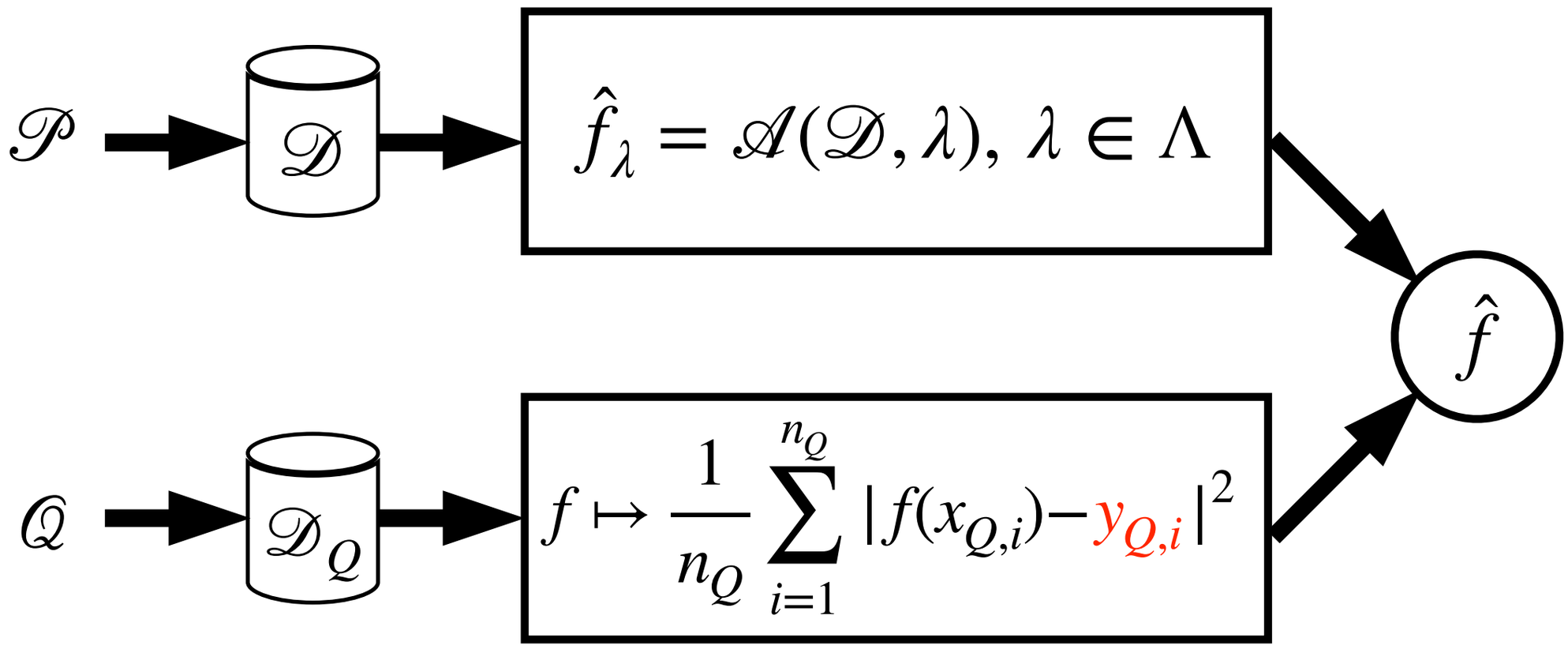
The "best" model depends on $\mathcal{Q}$.

Outline

- Introduction: covariate shift adaptation
- **Methodology: validation with pseudo-labels**
- Model selection theory: general results
- Adaptivity guarantees: kernel ridge regression

If Target Data were Labeled…

Source **Candidates**



Target **Risk estimator**

Issue: $\{y_{Q,i}\}_{i=1}^{n_Q}$ are missing.

Regression Imputation

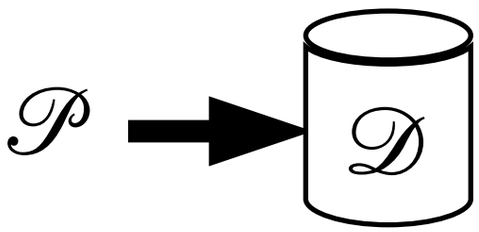
Idea: fill the missing labels $\{y_{Q,i}\}_{i=1}^{n_Q}$.

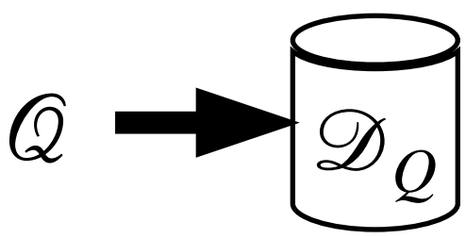
Train an **imputation model** $\tilde{f}$ to label the data: $\widetilde{y}_{Q,i} = \tilde{f}(x_{Q,i})$.

- **Pseudo-labeling and self-training** (Lee, 2013, Kumar et al., 2020, Cai et al., 2021, Liu et al., 2021)
- **Regression imputation** (Little and Rubin, 2019, Hirshberg et al., 2019, Kallus, 2020; Hirshberg and Wager, 2021; Mou et al., 2022)
- Different approaches: **propensity weighting** and **matching** (Shimodaira, 2000; Huang et al., 2006; Cortes et al., 2010; Ma et al., 2022)

Methodology: Start

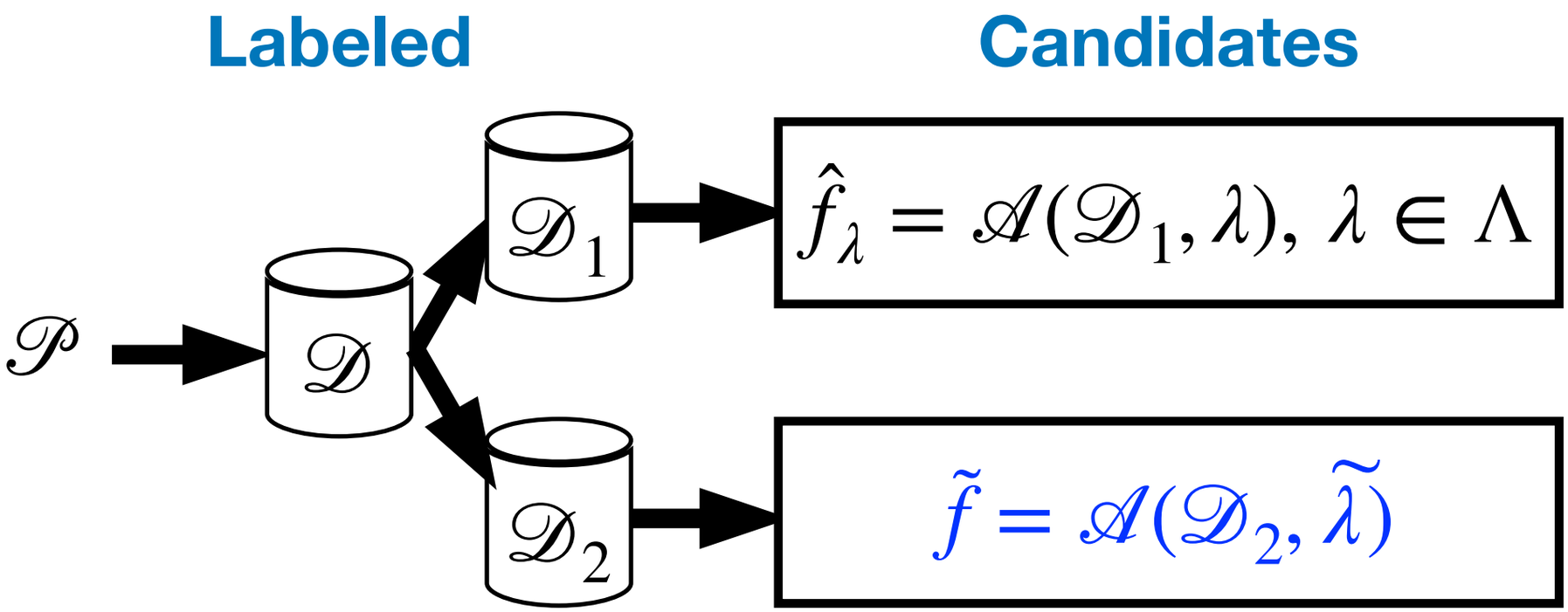
Labeled

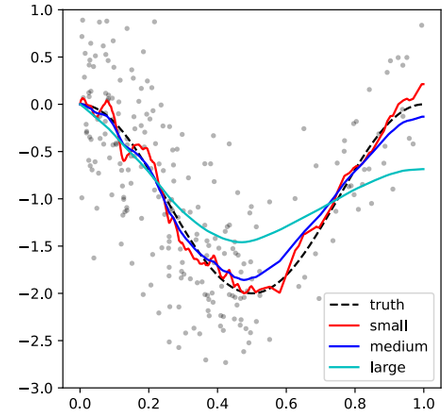




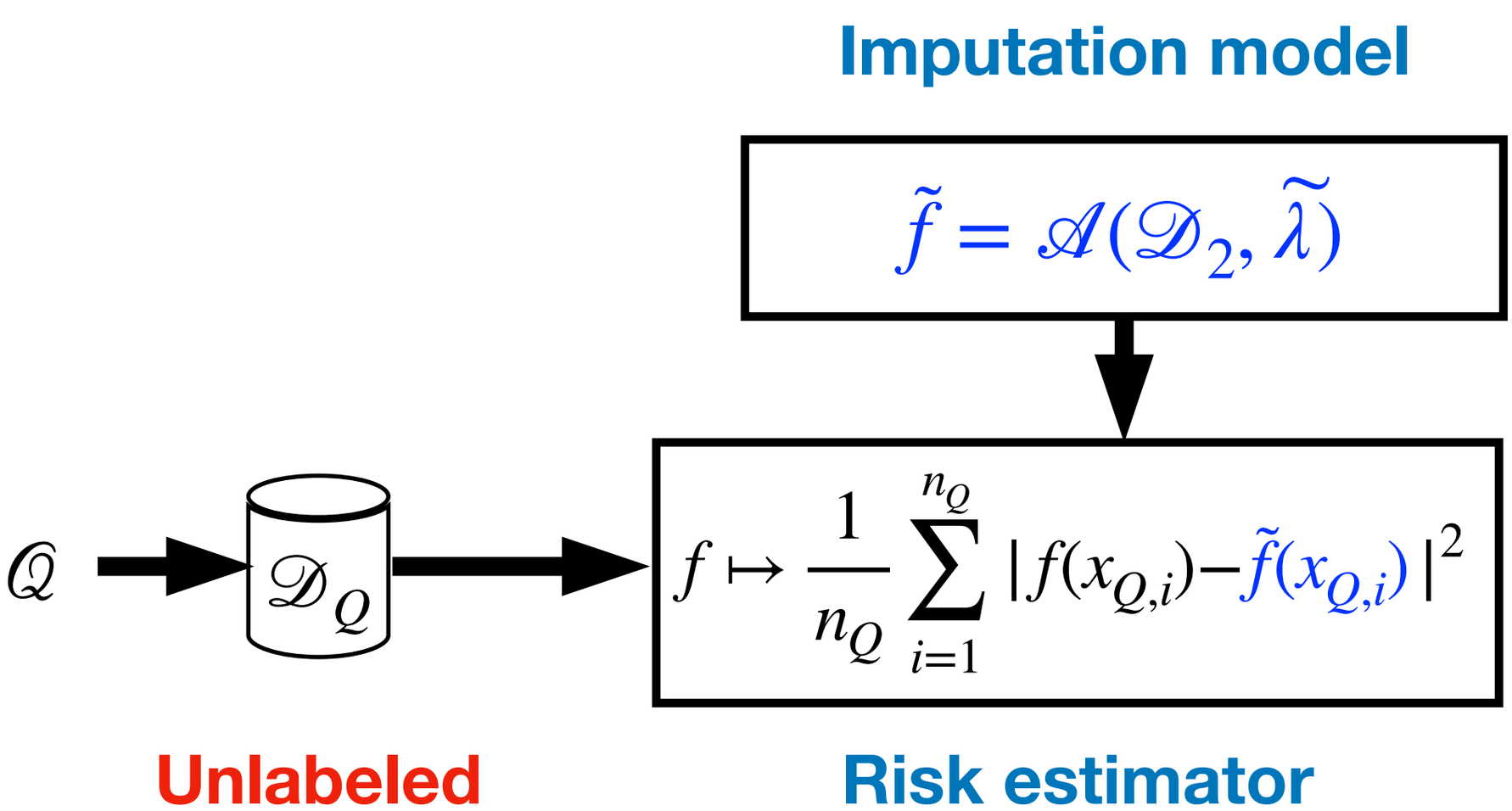
Unlabeled

Training

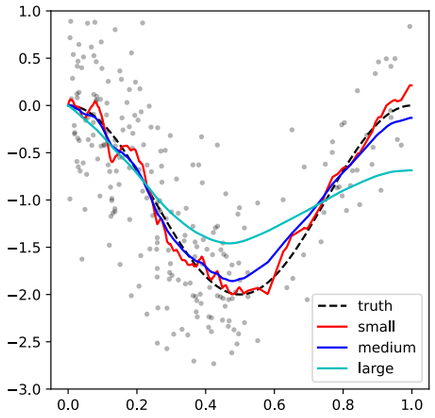

Labeled
Candidates
$\mathscr{P}$
$\mathscr{D}$
$\mathscr{D}_1$
$\mathscr{D}_2$
$\hat{f}_\lambda = \mathscr{A}(\mathscr{D}_1, \lambda),\ \lambda \in \Lambda$
$\tilde{f} = \mathscr{A}(\mathscr{D}_2, \tilde{\lambda})$
Imputation model


truth
small
medium
large

Pseudo-Labeling (Imputation)

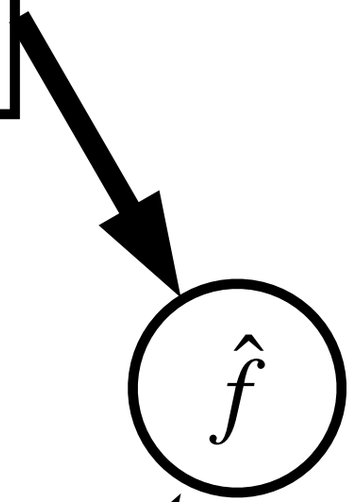


Model Selection



Candidates

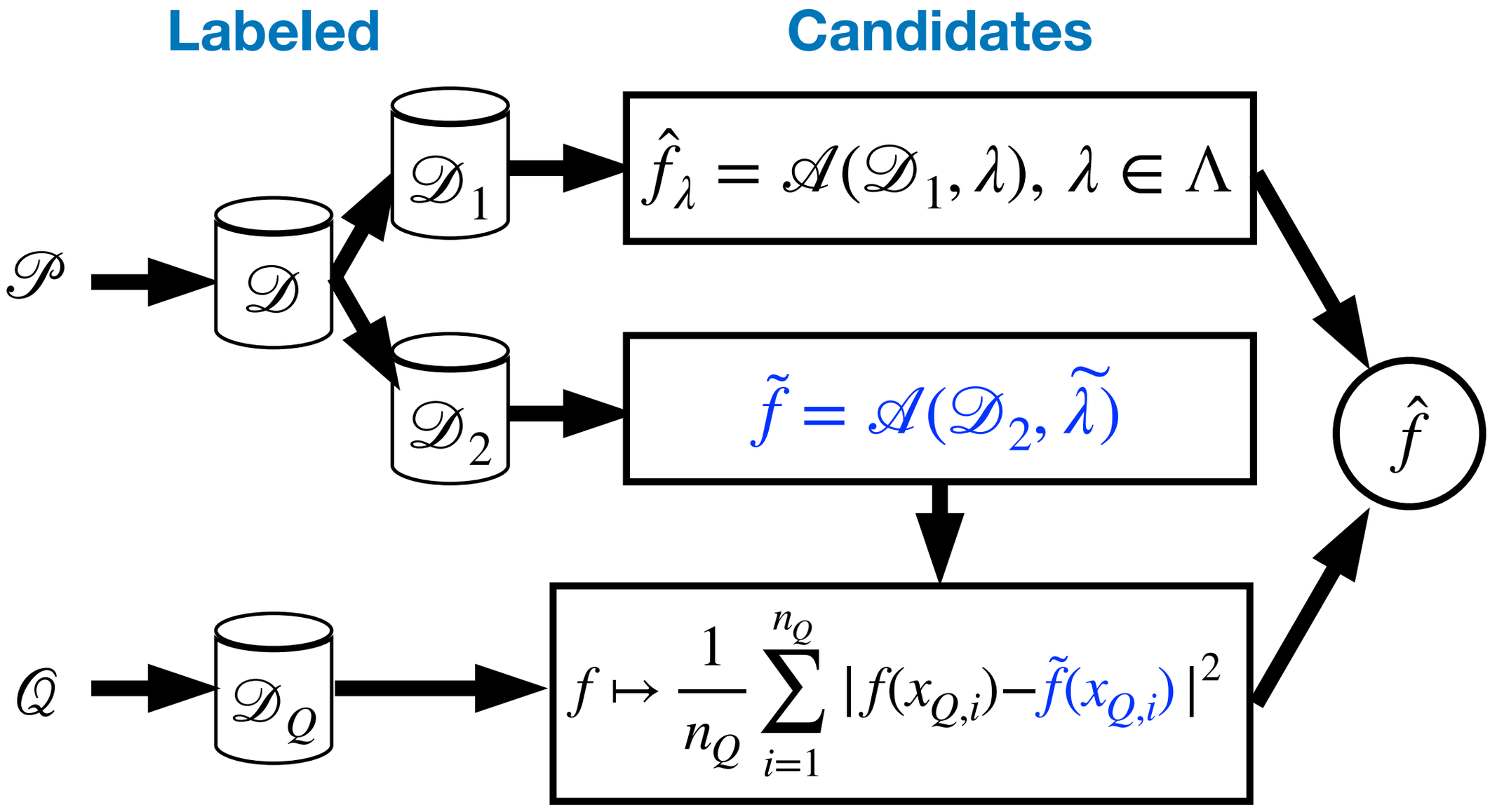
$$\hat{f}_\lambda = \mathscr{A}(\mathscr{D}_1, \lambda),\ \lambda \in \Lambda$$



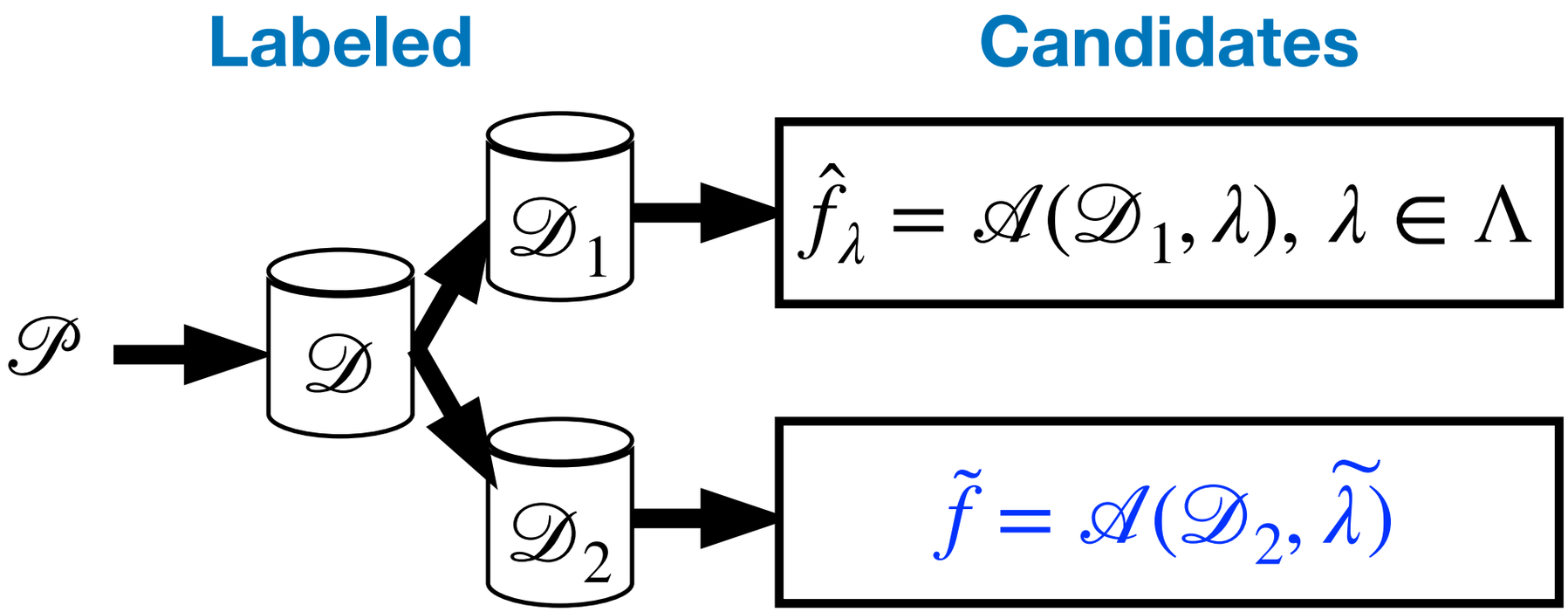
$$f \mapsto \frac{1}{n_Q} \sum_{i=1}^{n_Q} |f(x_{Q,i}) - \tilde{f}(x_{Q,i})|^2$$

Risk estimator

Methodology



A Chicken-and-Egg Dilemma?



- To select a good λ, need a good $\tilde{f}$, which needs a good $\tilde{\lambda}$.

Outline

- Introduction: covariate shift adaptation
- Methodology: validation with pseudo-labels
- **Model selection theory: general results**
- Adaptivity guarantees: kernel ridge regression

Validation with Pseudo-Labels

$\{x_i\}_{i=1}^n$ deterministic, $(\widetilde{y}_1, \cdots, \widetilde{y}_n)^\top \sim N(\overline{\mathbf{y}}, \mathbf{S})$.

Given $\{f_j\}_{j=1}^m$, select $\hat{j} \in \arg\min_{j \in [m]} \left\{ \frac{1}{n} \sum_{i=1}^n |f_j(x_i) - \widetilde{y}_i|^2 \right\}$.

Validation with Pseudo-Labels

$\{x_i\}_{i=1}^n$ deterministic, $(\widetilde{y}_1, \cdots, \widetilde{y}_n)^\top \sim N(\overline{\mathbf{y}}, \mathbf{S})$.

Given $\{f_j\}_{j=1}^m$, select $\hat{j} \in \arg\min_{j \in [m]} \left\{ \frac{1}{n} \sum_{i=1}^n |f_j(x_i) - \widetilde{y}_i|^2 \right\}$.

Define $\mathscr{E}(f) = \frac{1}{n} \sum_{i=1}^n |f(x_i) - f^\star(x_i)|^2$.

How does $\hat{f} = f_{\hat{j}}$ compare to the best in class?

Validation with Pseudo-Labels

$\{x_i\}_{i=1}^n$ deterministic, $(\widetilde{y}_1, \cdots, \widetilde{y}_n)^\top \sim N(\overline{\mathbf{y}}, \mathbf{S})$.

Given $\{f_j\}_{j=1}^m$, select $\hat{j} \in \arg\min_{j\in[m]} \left\{ \frac{1}{n} \sum_{i=1}^n |f_j(x_i) - \widetilde{y}_i|^2 \right\}$.

Define $\mathscr{E}(f) = \frac{1}{n} \sum_{i=1}^n |f(x_i) - f^\star(x_i)|^2$.

Theorem

$$\mathscr{E}(\hat{f}) \lesssim \min_{j\in[m]} \mathscr{E}(f_j) + \frac{1}{n} \sum_{i=1}^n |\overline{y}_i - f^\star(x_i)|^2 + \frac{\|\mathbf{S}\| \log m}{n}.$$

Bias-Variance Decomposition

Impact of $(\widetilde{y}_1, \cdots, \widetilde{y}_n)^\top \sim N(\overline{\mathbf{y}}, \mathbf{S})$ on model selection:

$$\mathscr{E}(\hat{f}) \lesssim \min_{j \in [m]} \mathscr{E}(f_j) + \frac{1}{n} \sum_{i=1}^{n} |\, \overline{y}_i - f^\star(x_i) \,|^2 + \frac{\|\mathbf{S}\| \log m}{n}.$$

Bias-Variance Decomposition

Impact of $(\widetilde{y}_1, \cdots, \widetilde{y}_n)^\top \sim N(\overline{\mathbf{y}}, \mathbf{S})$ on model selection:

$$\mathscr{E}(\hat{f}) \lesssim \min_{j \in [m]} \mathscr{E}(f_j) + \frac{1}{n} \sum_{i=1}^{n} |\, \overline{y}_i - f^\star(x_i) \,|^2 + \frac{\|\mathbf{S}\| \log m}{n}.$$

MSE of $(\widetilde{y}_1, \cdots, \widetilde{y}_n)^\top$:

$$\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} |\, \widetilde{y}_i - f^\star(x_i) \,|^2 \right) = \frac{1}{n} \sum_{i=1}^{n} |\, \overline{y}_i - f^\star(x_i) \,|^2 + \frac{\mathrm{Tr}(\mathbf{S})}{n}.$$

- True labels: $\overline{y}_i = f^\star(x_i), \quad \mathbf{S} = \sigma^2 \mathbf{I}.$

Bias-Variance Decomposition

Impact of $(\widetilde{y}_1, \cdots, \widetilde{y}_n)^\top \sim N(\overline{\mathbf{y}}, \mathbf{S})$ on model selection:

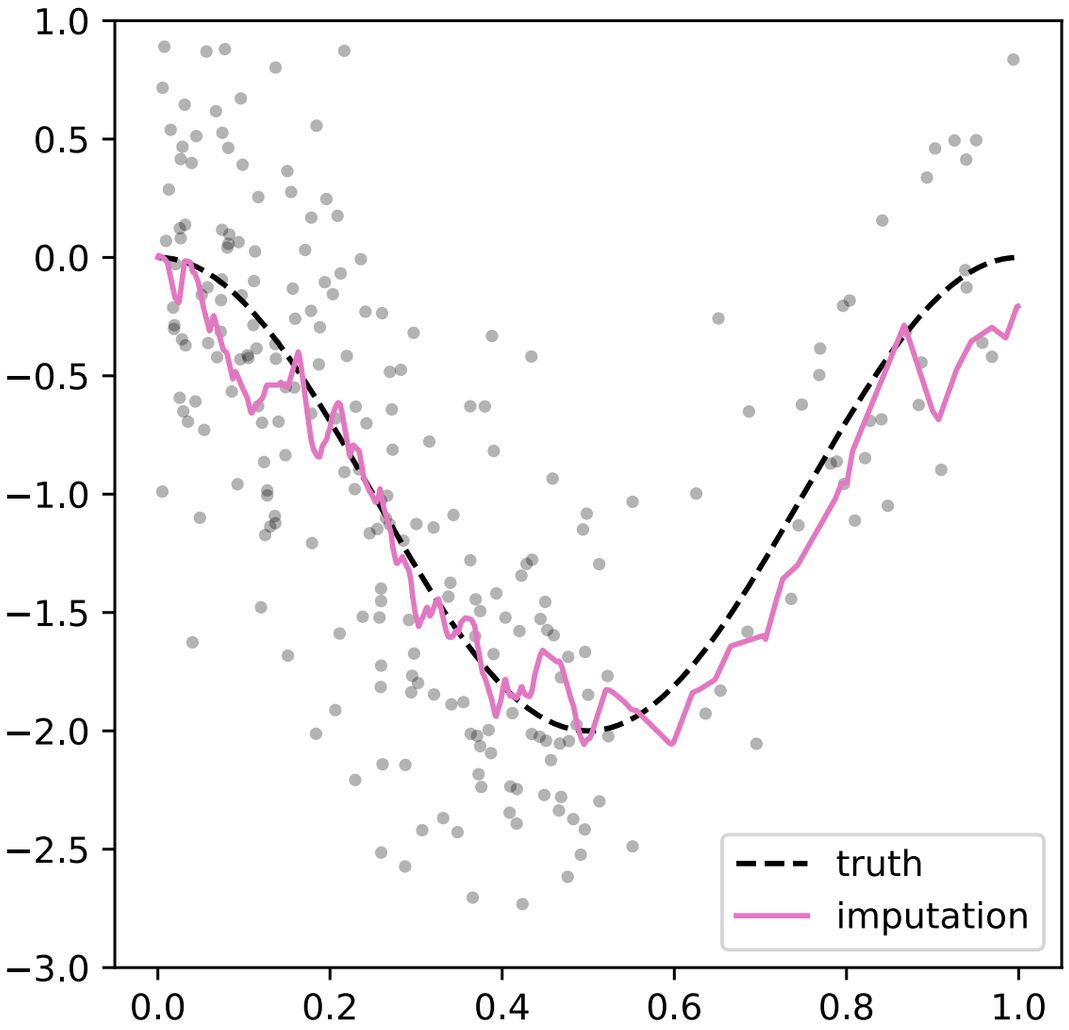
$$\mathscr{E}(\hat{f}) \lesssim \min_{j\in[m]} \mathscr{E}(f_j) + \frac{1}{n}\sum_{i=1}^{n} |\,\overline{y}_i - f^\star(x_i)\,|^2 + \frac{\|\mathbf{S}\|\log m}{n}.$$

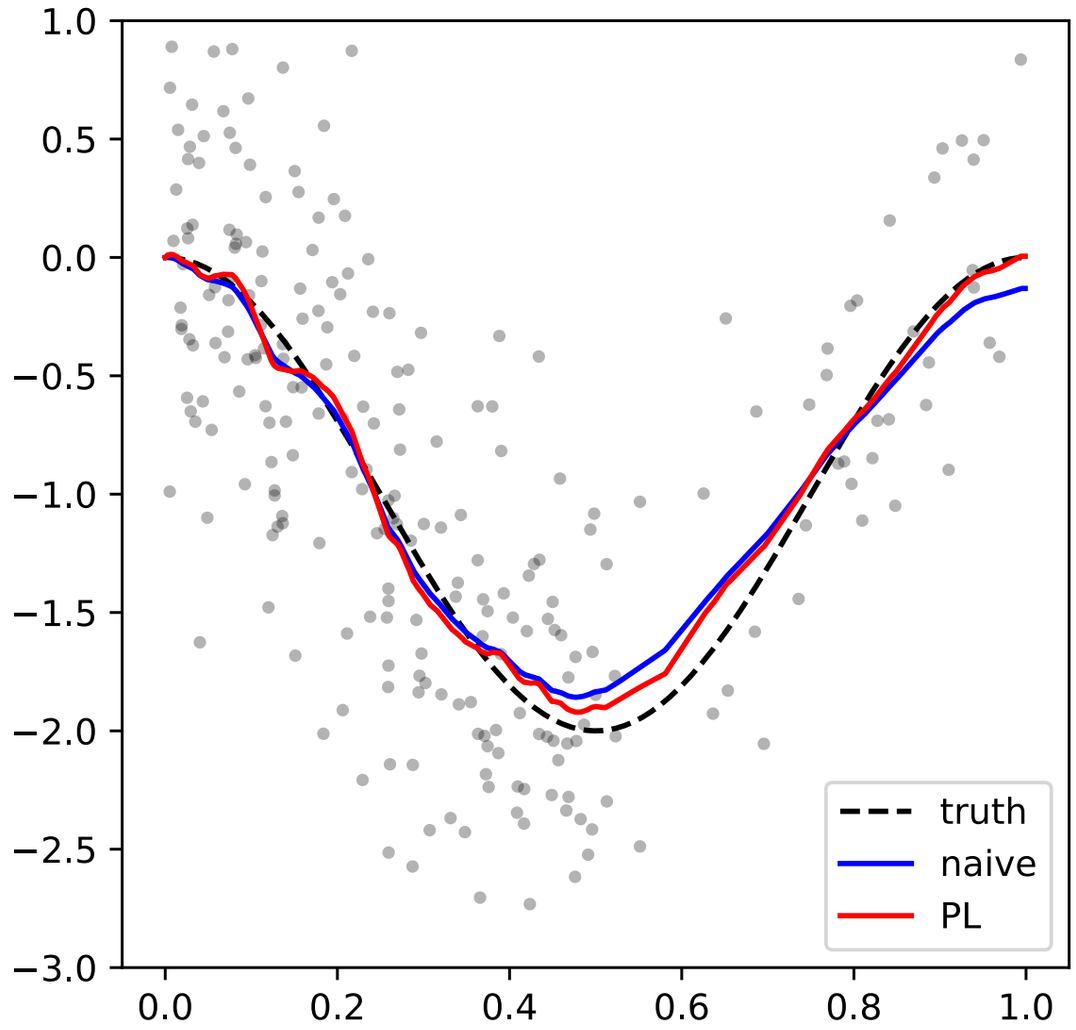
MSE of $(\widetilde{y}_1, \cdots, \widetilde{y}_n)^\top$:

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} |\,\widetilde{y}_i - f^\star(x_i)\,|^2\right) = \frac{1}{n}\sum_{i=1}^{n} |\,\overline{y}_i - f^\star(x_i)\,|^2 + \frac{\mathrm{Tr}(\mathbf{S})}{n}.$$

- For model selection, variance has a much weaker effect.
- **Good pseudo-labels: low bias and reasonable variance.**

Example: Kernel Ridge Regression





Target: $\mathcal{Q}_X = U[0.5,1]$

Outline

- Introduction: covariate shift adaptation
- Methodology: validation with pseudo-labels
- Model selection theory: general results
- **Adaptivity guarantees: kernel ridge regression**

Kernel Ridge Regression

Kernel ridge regression:

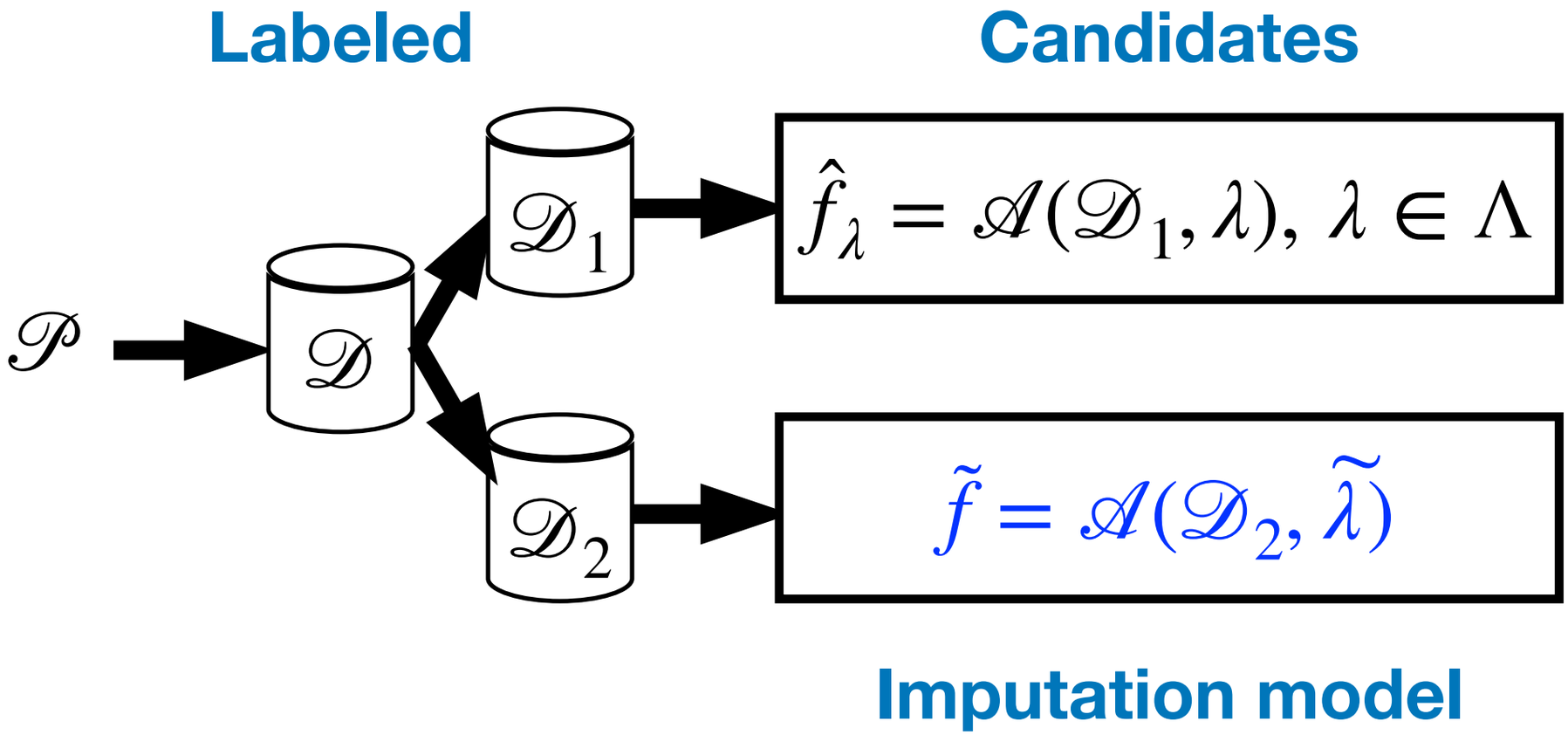
$$\mathscr{A}(\mathscr{D}, \lambda) = \text{argmin}_{f \in \mathscr{F}} \left\{ \frac{1}{|\mathscr{D}|} \sum_{(x,y) \in \mathscr{D}} |f(x) - y|^2 + \lambda \|f\|_{\mathscr{F}}^2 \right\}.$$

$\mathscr{F}$: **reproducing kernel Hilbert space (RKHS)** containing $f^\star$.

For some Hilbert space $\mathbb{H}$ and mapping $\phi : \mathscr{X} \to \mathbb{H}$,

KRR = ridge regression in $\{f_\theta(\,\cdot\,) = \langle \phi(\,\cdot\,), \theta \rangle : \theta \in \mathbb{H}\}$.

Undersmoothing for Imputation



Requirement on $\tilde{f}$: small bias, reasonable variance.

- $\tilde{\lambda} \asymp n_P^{-1}$, $\Lambda \asymp \{n_P^{-1}, 2n_P^{-1}, 2^2 n_P^{-1}, \cdots, 1\}$.

Main Result

Theorem (informal)

$\hat{f} \approx$ KRR on n_{e} labeled samples from $\mathcal{Q}$, with optimal λ.

- n_{e}: effective sample size, gauges transfer from $\mathcal{P}$ to $\mathcal{Q}$.
- Adaptation to the spectrum of $\mathcal{Q}$ and the covariate shift.

Excess Risk Bounds

Theorem

Let $\mathbf{\Sigma}_Q = \mathbb{E}_Q[\phi(x) \otimes \phi(x)]$, and $\{\mu_j\}_{j=1}^{\infty}$ be its eigenvalues.

- If $\operatorname{rank}(\mathbf{\Sigma}_Q) \leq D$, then $R(\hat{f}) - R(f^\star) \lesssim D n_{\mathrm{e}}^{-1} + n_Q^{-1}$;
- If $\mu_j \lesssim e^{-cj}$, then $R(\hat{f}) - R(f^\star) \lesssim n_{\mathrm{e}}^{-1} + n_Q^{-1}$;
- If $\mu_j \lesssim j^{-2\alpha}$, then $R(\hat{f}) - R(f^\star) \lesssim n_{\mathrm{e}}^{-\frac{2\alpha}{2\alpha+1}} + n_Q^{-1}$.

Excess Risk Bounds

Theorem

Let $\boldsymbol{\Sigma}_Q = \mathbb{E}_Q[\phi(x) \otimes \phi(x)]$, and $\{\mu_j\}_{j=1}^{\infty}$ be its eigenvalues.

- If $\operatorname{rank}(\boldsymbol{\Sigma}_Q) \leq D$, then $R(\hat{f}) - R(f^\star) \lesssim D n_{\mathrm{e}}^{-1} + n_Q^{-1}$;
- If $\mu_j \lesssim e^{-cj}$, then $R(\hat{f}) - R(f^\star) \lesssim n_{\mathrm{e}}^{-1} + n_Q^{-1}$;
- If $\mu_j \lesssim j^{-2\alpha}$, then $R(\hat{f}) - R(f^\star) \lesssim n_{\mathrm{e}}^{-\frac{2\alpha}{2\alpha+1}} + n_Q^{-1}$.

- First term: **optimal** rate for n_{e} labeled samples from $\mathcal{Q}$.
- Second term: **minor** cost of using limited samples from $\mathcal{Q}$.

Effective Sample Size

Definition

$$n_{\mathsf{e}} = \sup\{t \le n_P : \ \ t\boldsymbol{\Sigma}_Q \le n_P\boldsymbol{\Sigma}_P + \mathbf{I}\}.$$

The followings have the same excess risk bound:

- KRR on n_P **labeled source samples**, with $\mathcal{Q}$-optimal λ;
- KRR on n_{e} **labeled target samples**, with $\mathcal{Q}$-optimal λ.

Pseudo-labeling helps select a near-optimal λ.

Adaptivity

Theorem (informal)

$\hat{f} \approx$ optimal KRR on n_{e} labeled samples from $\mathcal{Q}$.

$$n_{\mathsf{e}} = \sup\{t \le n_P : \ t\boldsymbol{\Sigma}_Q \le n_P\boldsymbol{\Sigma}_P + \mathbf{I}\}.$$

Example: $\dfrac{\mathrm{d}\mathcal{Q}_X}{\mathrm{d}\mathcal{P}_X} \le B \quad \Rightarrow \quad \boldsymbol{\Sigma}_Q \le B\boldsymbol{\Sigma}_P \quad \Rightarrow \quad n_{\mathsf{e}} \ge n_P/B.$

Adaptivity

Theorem (informal)

$\hat{f} \approx$ optimal KRR on n_{e} labeled samples from $\mathcal{Q}$.

$$n_{\mathsf{e}} = \sup\{t \le n_P : \ t\boldsymbol{\Sigma}_Q \le n_P \boldsymbol{\Sigma}_P + \mathbf{I}\}.$$

Example: $\dfrac{\mathrm{d}\mathcal{Q}_X}{\mathrm{d}\mathcal{P}_X} \le B \quad \Rightarrow \quad \boldsymbol{\Sigma}_Q \le B\boldsymbol{\Sigma}_P \quad \Rightarrow \quad n_{\mathsf{e}} \ge n_P/B.$

- Optimal excess risk bound. **Adaptive to B & spectral decay**.
- Matching lower bound: Ma, Pathak and Wainwright, 2023.

Adaptivity

Theorem (informal)

$\hat{f} \approx$ optimal KRR on n_{e} labeled samples from $\mathcal{Q}$.

$$n_{\mathrm{e}} = \sup\{t \le n_P : \ t\boldsymbol{\Sigma}_Q \le n_P\boldsymbol{\Sigma}_P + \mathbf{I}\}.$$

Example: $\mathcal{P}_X = U[0,1]$, $\mathcal{Q}_X = \delta_{x_0}$, 1st-order Sobolev space

- $n_{\mathrm{e}} \asymp ?$

Adaptivity

Theorem (informal)

$\hat{f} \approx$ optimal KRR on n_{e} labeled samples from $\mathcal{Q}$.

$$n_{\mathsf{e}} = \sup\{t \le n_P : \ t\boldsymbol{\Sigma}_Q \le n_P\boldsymbol{\Sigma}_P + \mathbf{I}\}.$$

Example: $\mathcal{P}_X = U[0,1],\ \mathcal{Q}_X = \delta_{x_0}$, 1st-order Sobolev space

- $n_{\mathsf{e}} \asymp n_P^{1/2} \Rightarrow\ R(\hat{f}) - R(f^\star) \lesssim n_{\mathsf{e}}^{-1} \asymp n_P^{-1/2}$, optimal.
- **Adaptive to singularity!**
- Each $\hat{f}_\lambda$ is **sub-optimal** for either $\mathcal{P}_X$ or $\mathcal{Q}_X$.

Summary

- Covariate shift adaptation given unlabeled target data
- **Method: hold-out validation with pseudo-labels**
- General requirement: low bias, reasonable variance
- Analysis of kernel ridge regression: Adaptivity

Wang. Pseudo-Labeling for Kernel Ridge Regression under Covariate Shift. arXiv:2302.10160, 2023.

Supported by NSF DMS-2210907, Columbia University DSI SF-181.

Q & A

Thank you!

Validation with Pseudo-Labels

$\{(x_i, y_i)\}_{i=1}^n$, x_i deterministic, $\mathbf{y} = (y_1, \cdots, y_n)^\top \sim N(\bar{\mathbf{y}}, \mathbf{S})$.

Given $\{f_j\}_{j=1}^m$, select $\hat{j} \in \arg\min_{j \in [m]} \left\{ \frac{1}{n} \sum_{i=1}^n |f(x_i) - y_i|^2 \right\}$.

Define $\mathscr{E}(f) = \frac{1}{n} \sum_{i=1}^n |f(x_i) - f^*(x_i)|^2$.

Theorem

$$\sqrt{\mathscr{E}(\hat{f})} - \min_{j \in [m]} \sqrt{\mathscr{E}(f_j)} \lesssim \sqrt{\frac{1}{n} \sum_{i=1}^n |\bar{y}_i - f^*(x_i)|^2} + \sqrt{\frac{\|\mathbf{S}\| \log m}{n}}.$$

Challenges

Example: $\mathscr{X} = [0,1]$, $\mathscr{P}_X = U[0,1]$, Sobolev kernel.

- $\mathscr{Q}_X = \mathscr{P}_X$ (no shift): $\lambda \asymp n^{-2/3}$ is minimax optimal.
- $\mathscr{Q}_X = \delta_{x_0}$ (e.g., conditional ATE): $\lambda \asymp n^{-1}$.

Speckman (1979), Li (1982), Donoho (1994), Wainwright (2019).