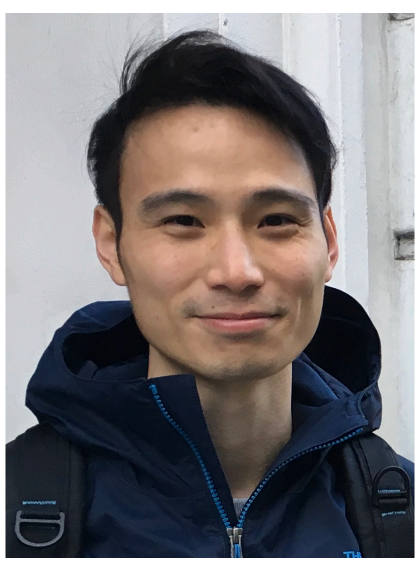
Sequential Domain Adaptation by Synthesizing Distributionally Robust Experts

Bahar Taşkesen

Risk Analytics and Optimization Chair
École Polytechnique Fédérale de Lausanne
rao.epfl.ch

Joint work with

Man-Chung Yue | Jose Blanchet | Daniel Kuhn | Viet Anh Nguyen



EPFL

The Hong Kong Polytechnic University
香港理工大學

Stanford University

Domain Adaptation

https://greenlogic.com.au/blog/what-is-transfer-learning/

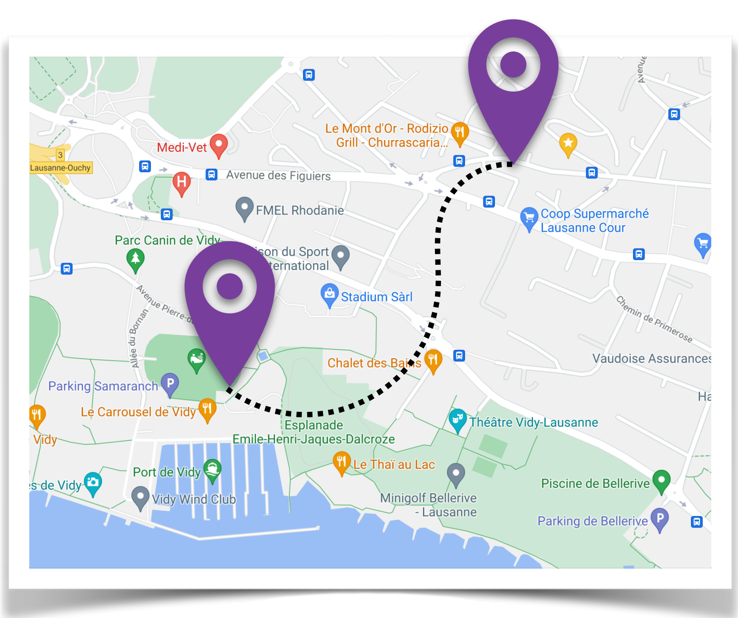
Supervised Domain Adaptation

https://greenlogic.com.au/blog/what-is-transfer-learning/

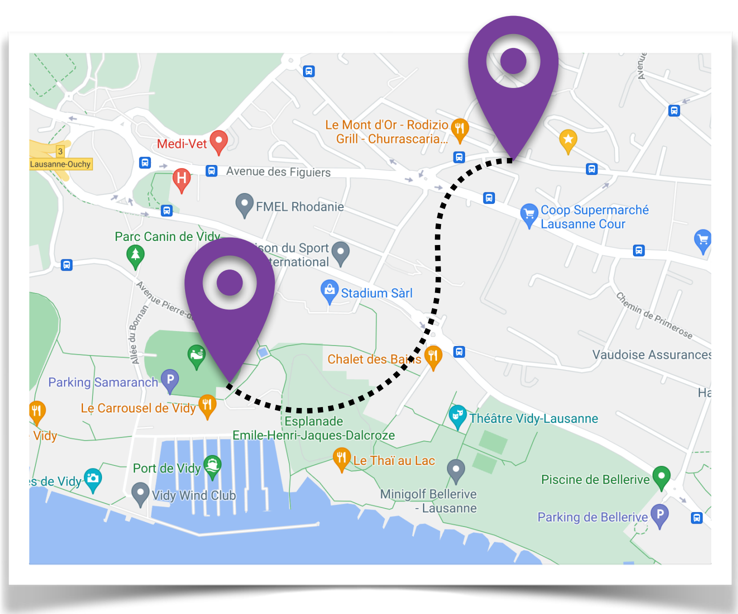
Supervised Domain Adaptation

https://greenlogic.com.au/blog/what-is-transfer-learning/

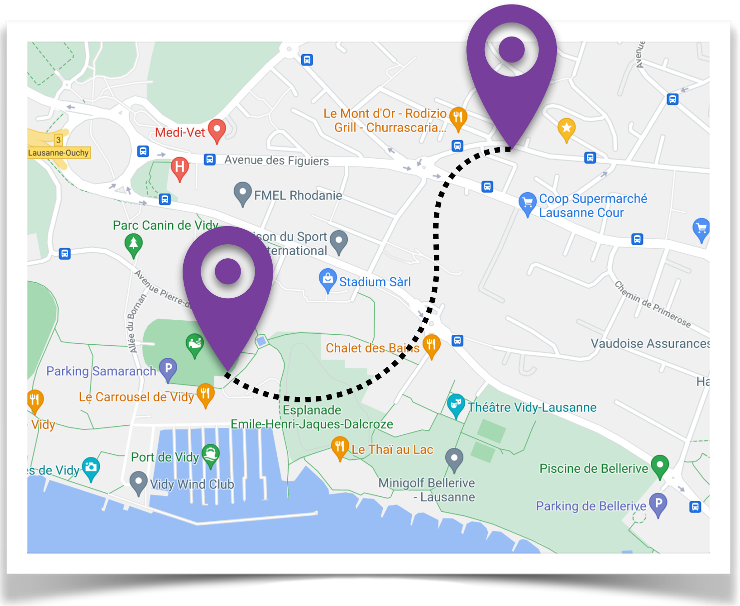
Supervised Domain Adaptation

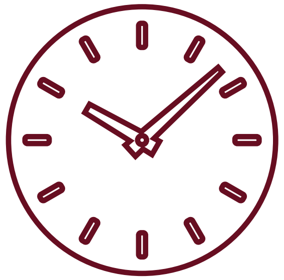


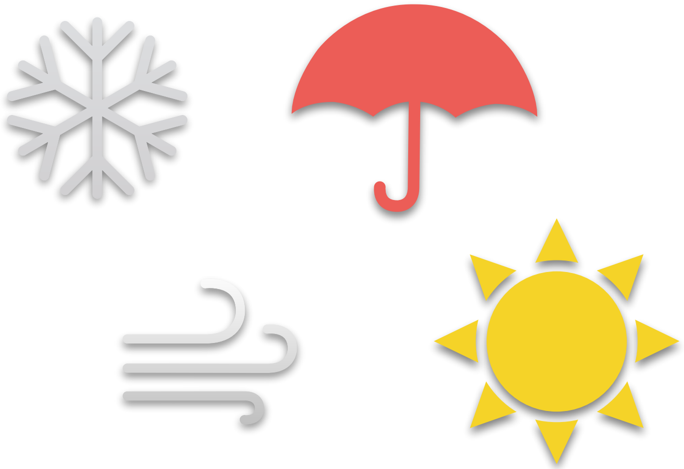
Supervised Domain Adaptation



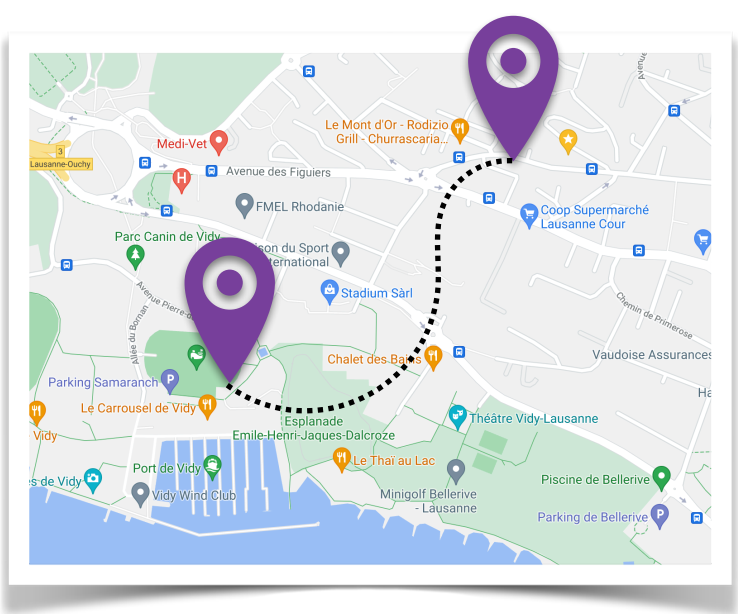
Supervised Domain Adaptation





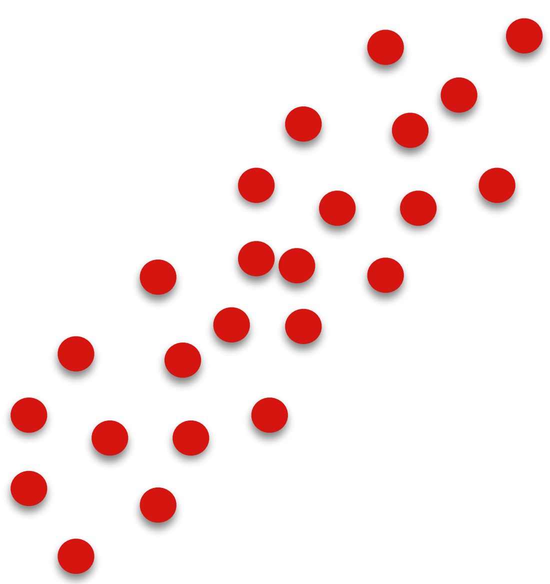


Supervised Domain Adaptation



Ridge Regression

X : d-dimensional covariate Y : univariate response N : number of samples

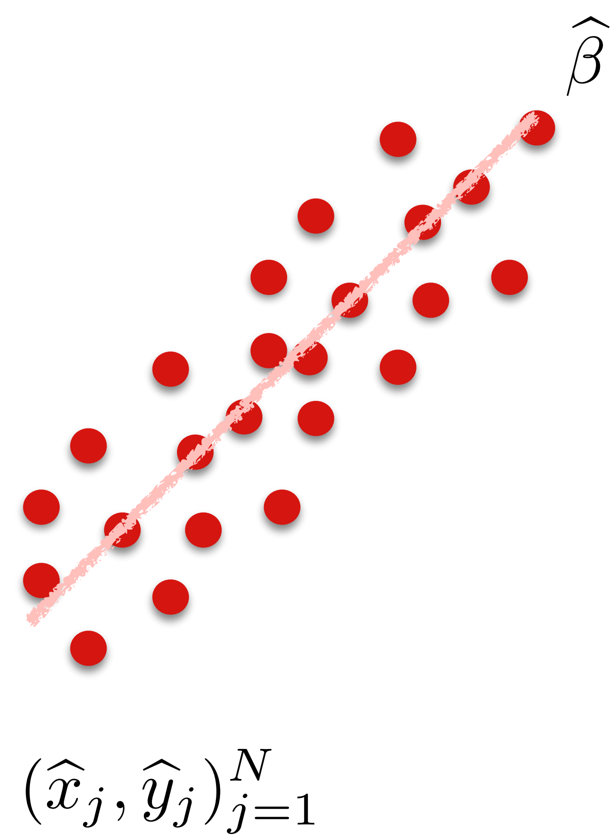


$(\widehat{x}_j, \widehat{y}_j)_{j=1}^N$

Ridge Regression

X : d-dimensional covariate $\quad$ Y : univariate response $\quad$ N : number of samples

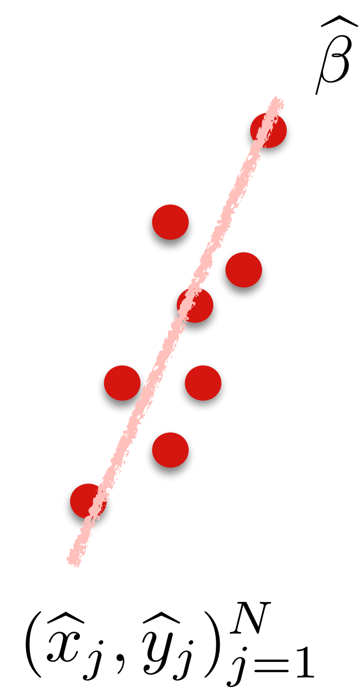
$$\min_{\beta \in \mathbb{R}^d} \frac{1}{N} \sum_{j=1}^{N} (\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta \|\beta\|_2^2 \quad \Rightarrow \quad \widehat{\beta} = \left(\frac{1}{N} \sum_{j=1}^{N} \widehat{x}_j \widehat{x}_j^\top + \eta I_d \right)^{-1} \left(\frac{1}{N} \sum_{j=1}^{N} \widehat{x}_j \widehat{y}_j \right)$$



Ridge Regression - Scarce data

X : d-dimensional covariate Y : univariate response N : number of samples

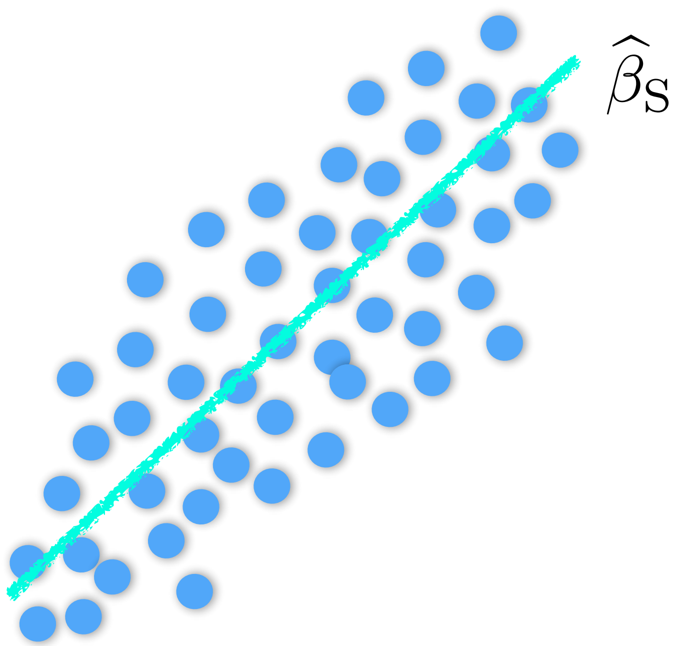
$$\min_{\beta \in \mathbb{R}^d} \frac{1}{N} \sum_{j=1}^{N} (\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta \|\beta\|_2^2 \quad \Rightarrow \quad \widehat{\beta} = \left(\frac{1}{N} \sum_{j=1}^{N} \widehat{x}_j \widehat{x}_j^\top + \eta I_d \right)^{-1} \left(\frac{1}{N} \sum_{j=1}^{N} \widehat{x}_j \widehat{y}_j \right)$$



Supervised Domain Adaptation

$$\min_{\beta \in \mathbb{R}^d} \frac{1}{N} \sum_{j=1}^{N} (\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta \|\beta\|_2^2$$

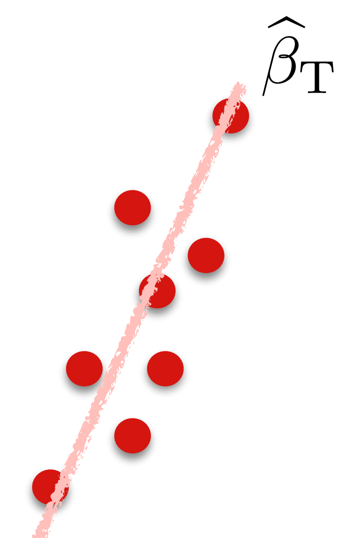
Source



Abundant **labelled** data

$(\widehat{x}_i, \widehat{y}_i)_{i=1}^{N_{\mathrm{S}}}$

Target



Scarce **labelled** data

$(\widehat{x}_j, \widehat{y}_j)_{j=1}^{N_{\mathrm{T}}}$

Supervised Domain Adaptation

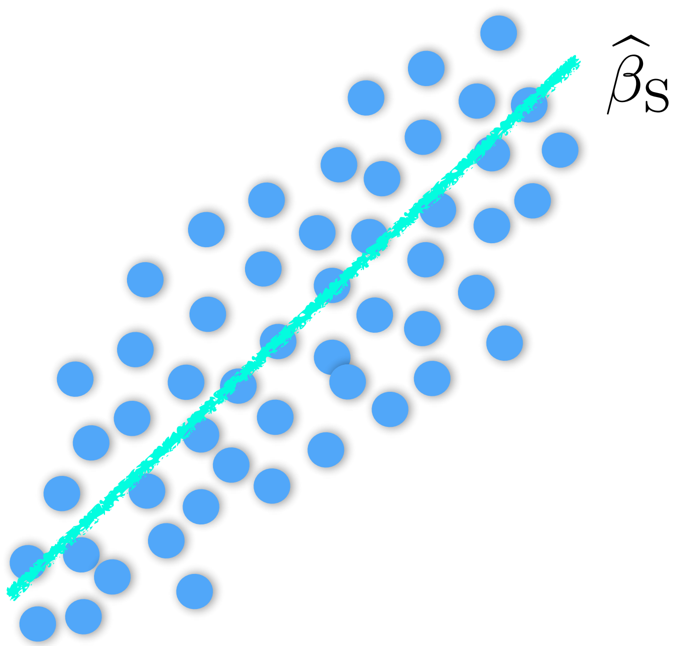
$$\min_{\beta \in \mathbb{R}^d} \frac{1}{N} \sum_{j=1}^{N} (\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta \|\beta\|_2^2$$

Source

Target

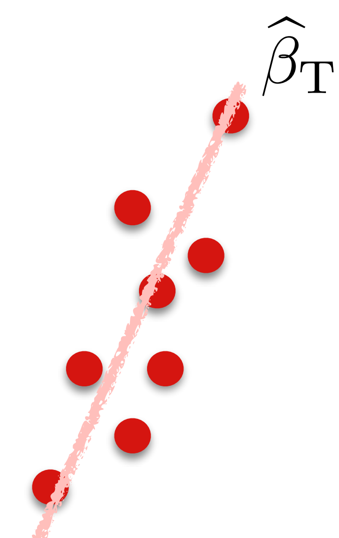
<u>Aim</u>

- Model works well on target domain



Abundant **labelled** data

$(\widehat{x}_i, \widehat{y}_i)_{i=1}^{N_{\mathrm{S}}}$



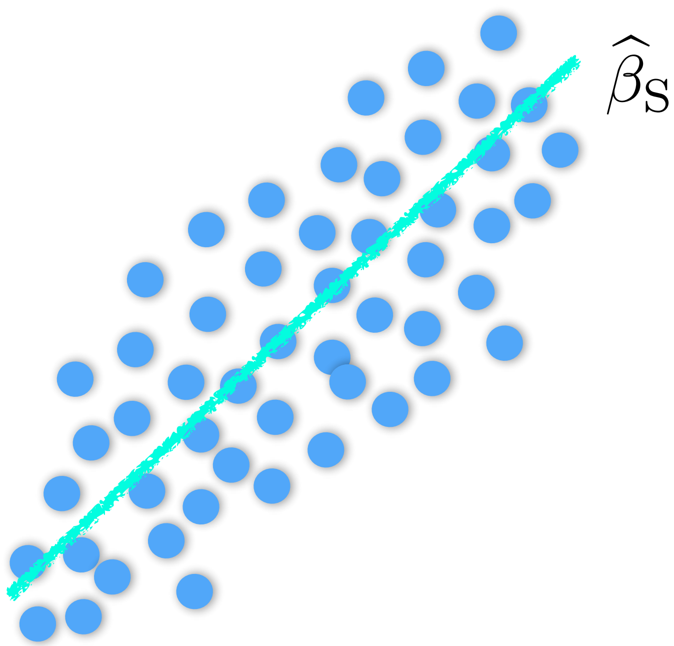
Scarce **labelled** data

$(\widehat{x}_j, \widehat{y}_j)_{j=1}^{N_{\mathrm{T}}}$

Supervised Domain Adaptation

$$\min_{\beta \in \mathbb{R}^d} \frac{1}{N} \sum_{j=1}^{N} (\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta \|\beta\|_2^2$$

Source



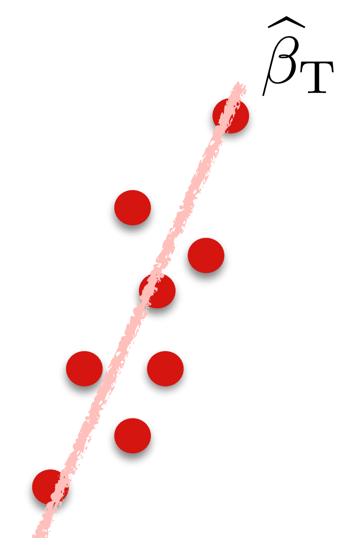
Abundant **labelled** data

$(\widehat{x}_i, \widehat{y}_i)_{i=1}^{N_{\mathrm{S}}}$

Assumptions

- Source and target are close
- Unseen target data arrive sequentially

Target



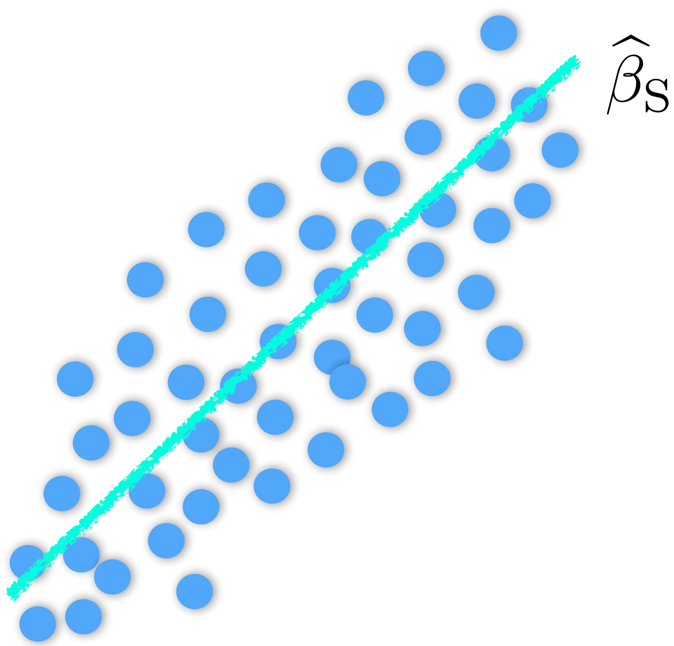
Scarce **labelled** data

$(\widehat{x}_j, \widehat{y}_j)_{j=1}^{N_{\mathrm{T}}}$

Supervised Domain Adaptation

$$\min_{\beta \in \mathbb{R}^d} \frac{1}{N} \sum_{j=1}^{N} (\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta \|\beta\|_2^2$$

Source



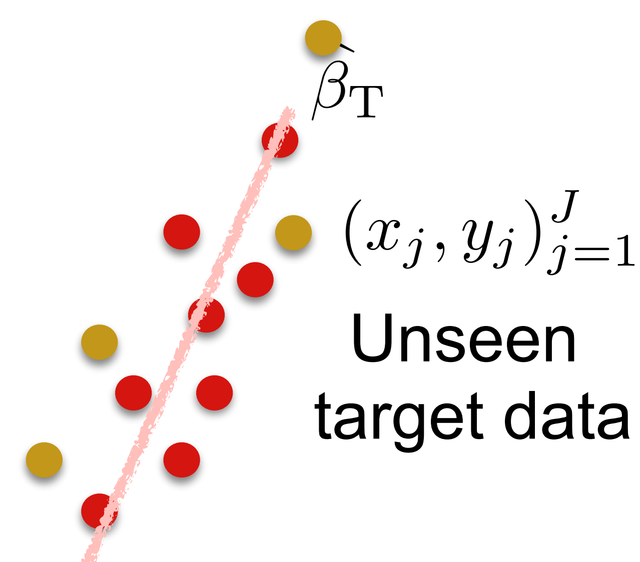
Abundant **labelled** data

$(\widehat{x}_i, \widehat{y}_i)_{i=1}^{N_{\mathrm{S}}}$

Assumptions

- Source and target are close
- Unseen target data arrive sequentially

Target



Scarce **labelled** data

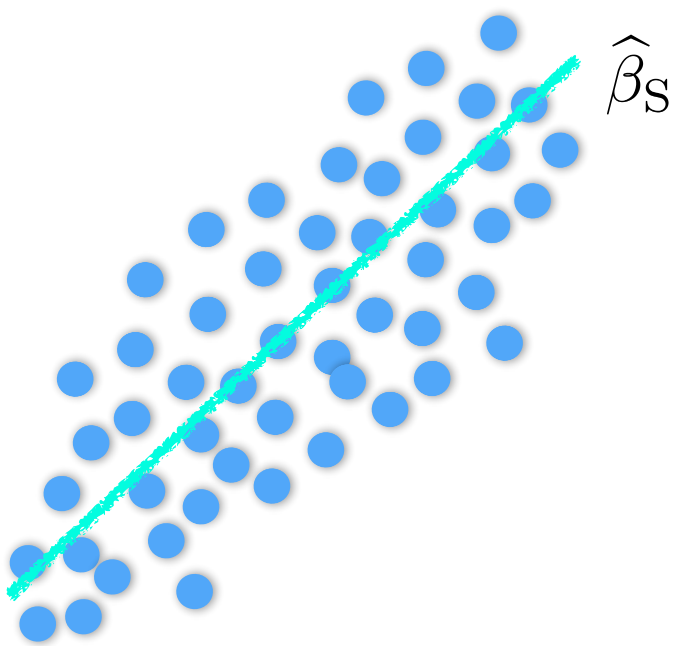
$(\widehat{x}_j, \widehat{y}_j)_{j=1}^{N_{\mathrm{T}}}$

Supervised Domain Adaptation

$$\min_{\beta \in \mathbb{R}^d} \frac{1}{N} \sum_{j=1}^{N} (\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta \|\beta\|_2^2$$

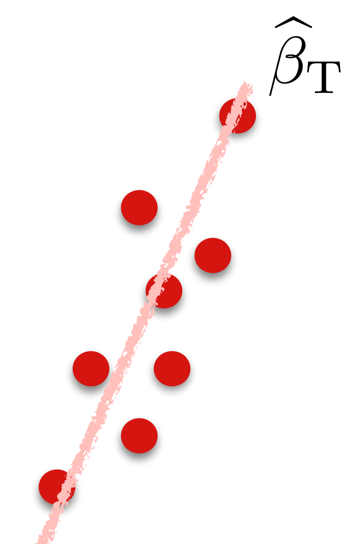
Source

Target



<u>Challenges</u>

1. Exploit source data
2. Tune hyperparameters



Abundant **labelled** data

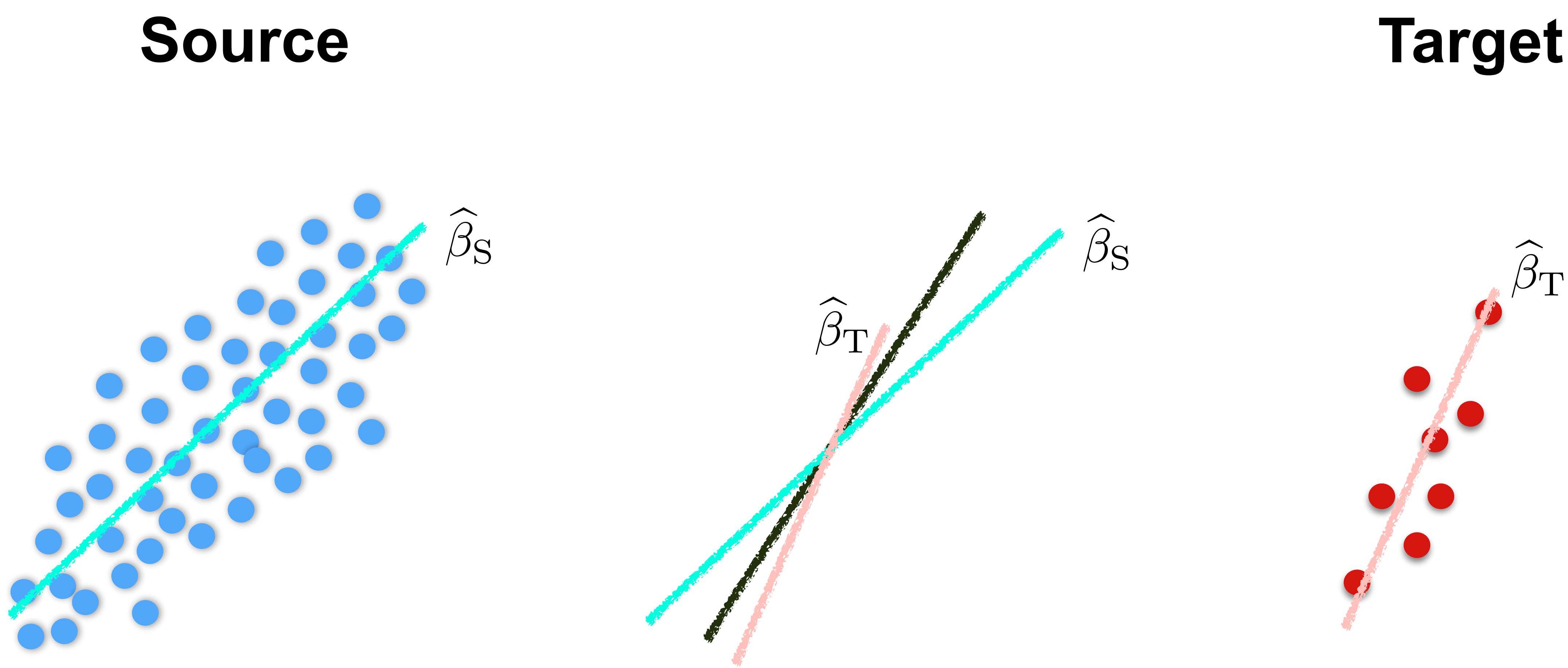
$(\widehat{x}_i, \widehat{y}_i)_{i=1}^{N_{\mathrm{S}}}$

Scarce **labelled** data

$(\widehat{x}_j, \widehat{y}_j)_{j=1}^{N_{\mathrm{T}}}$

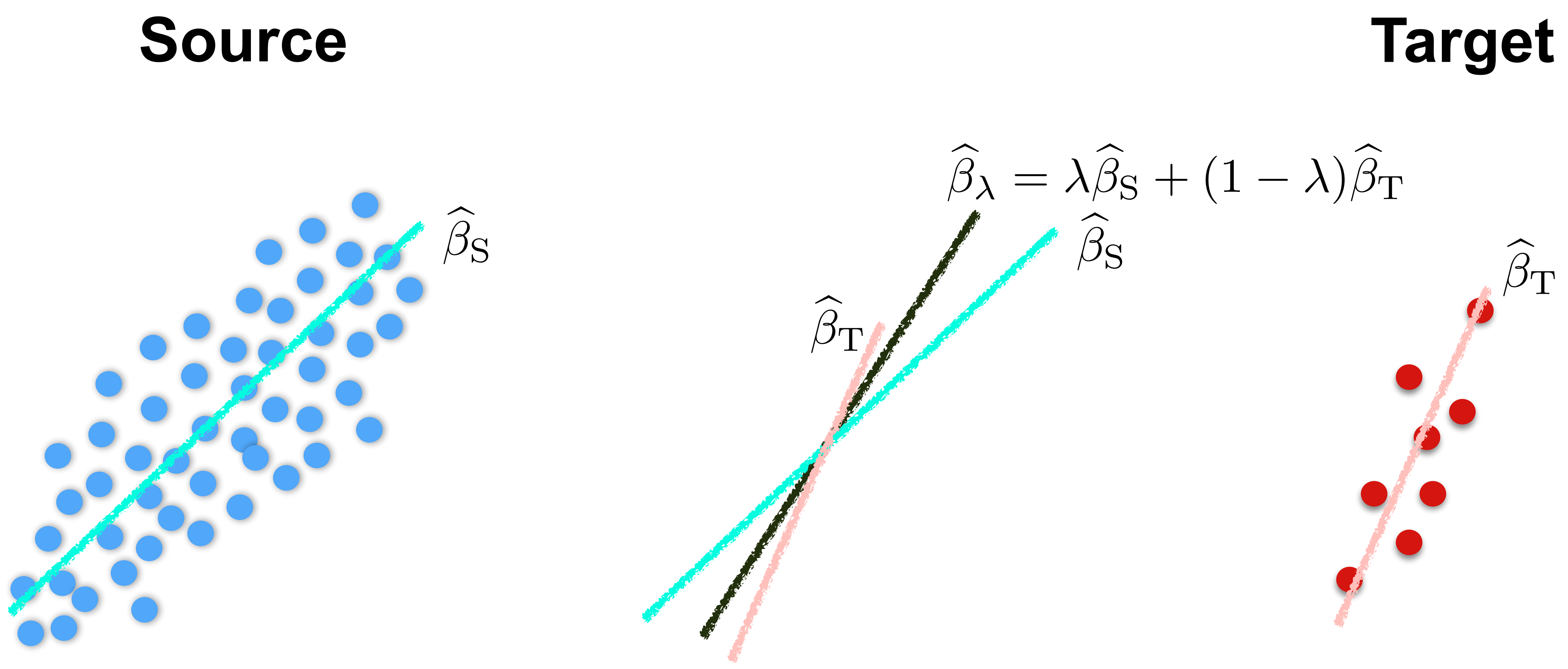
Exploit Source Data

1) Convex Combination Strategy (Baseline)



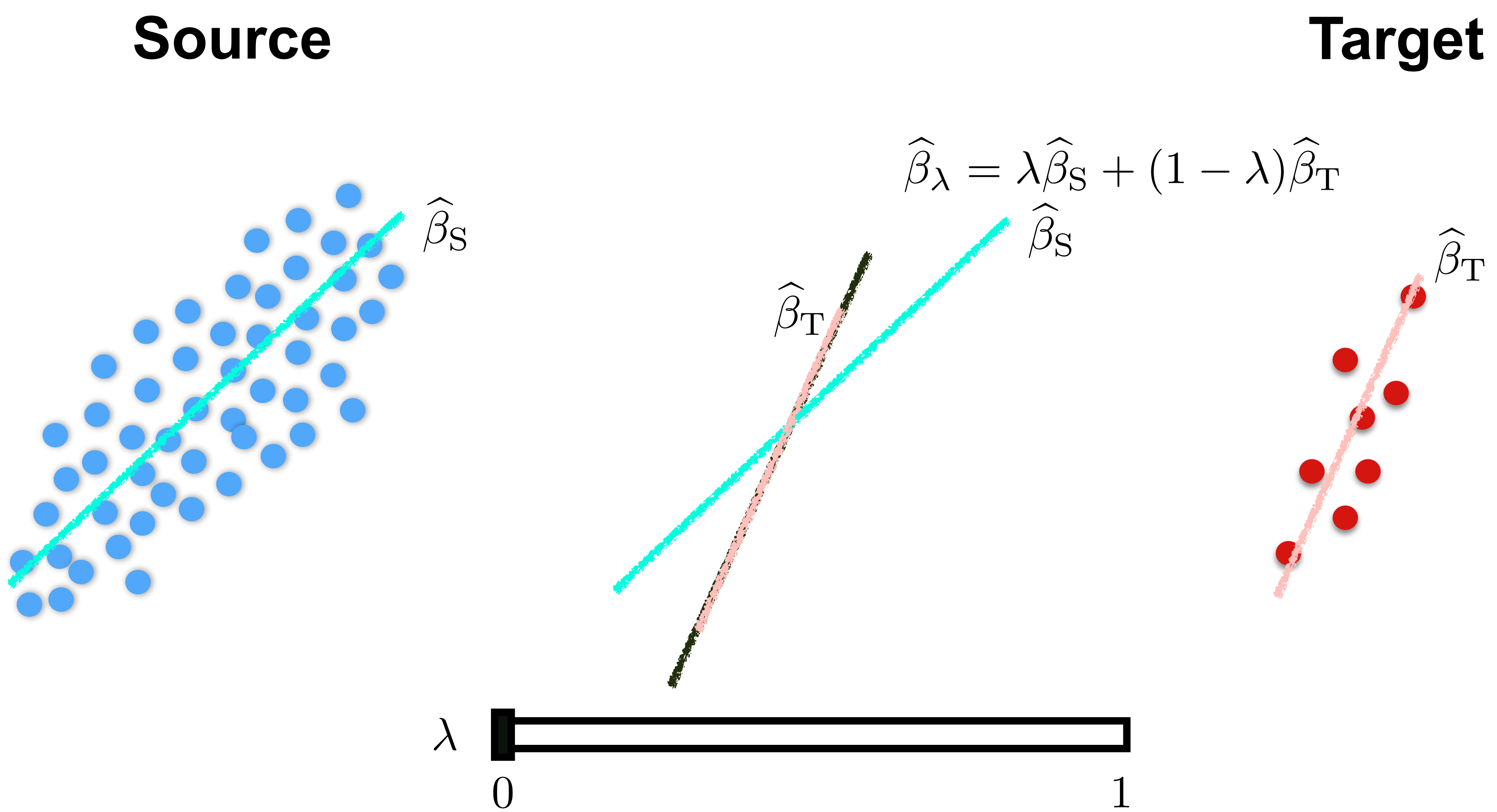
Exploit Source Data

1) Convex Combination Strategy (Baseline)



Exploit Source Data

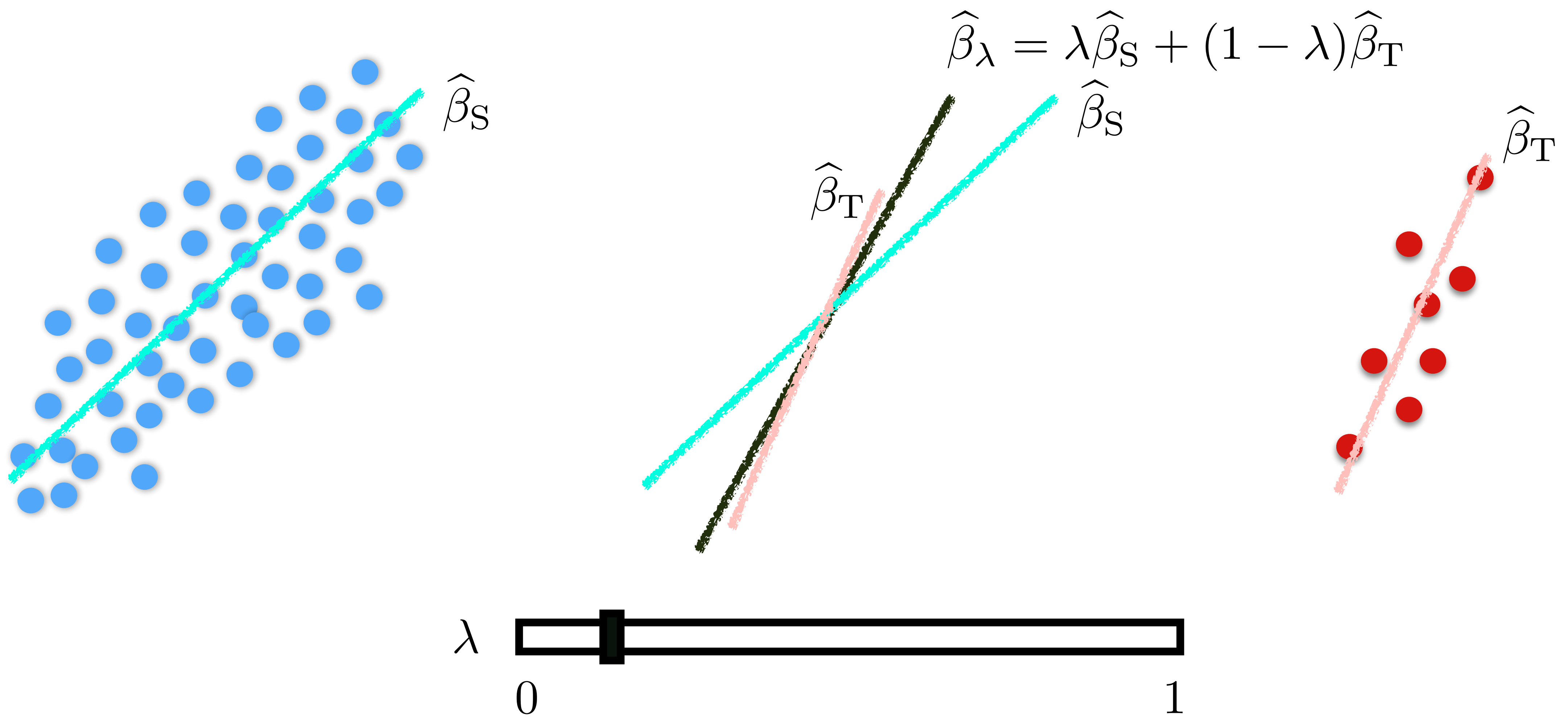
1) Convex Combination Strategy (Baseline)



Exploit Source Data

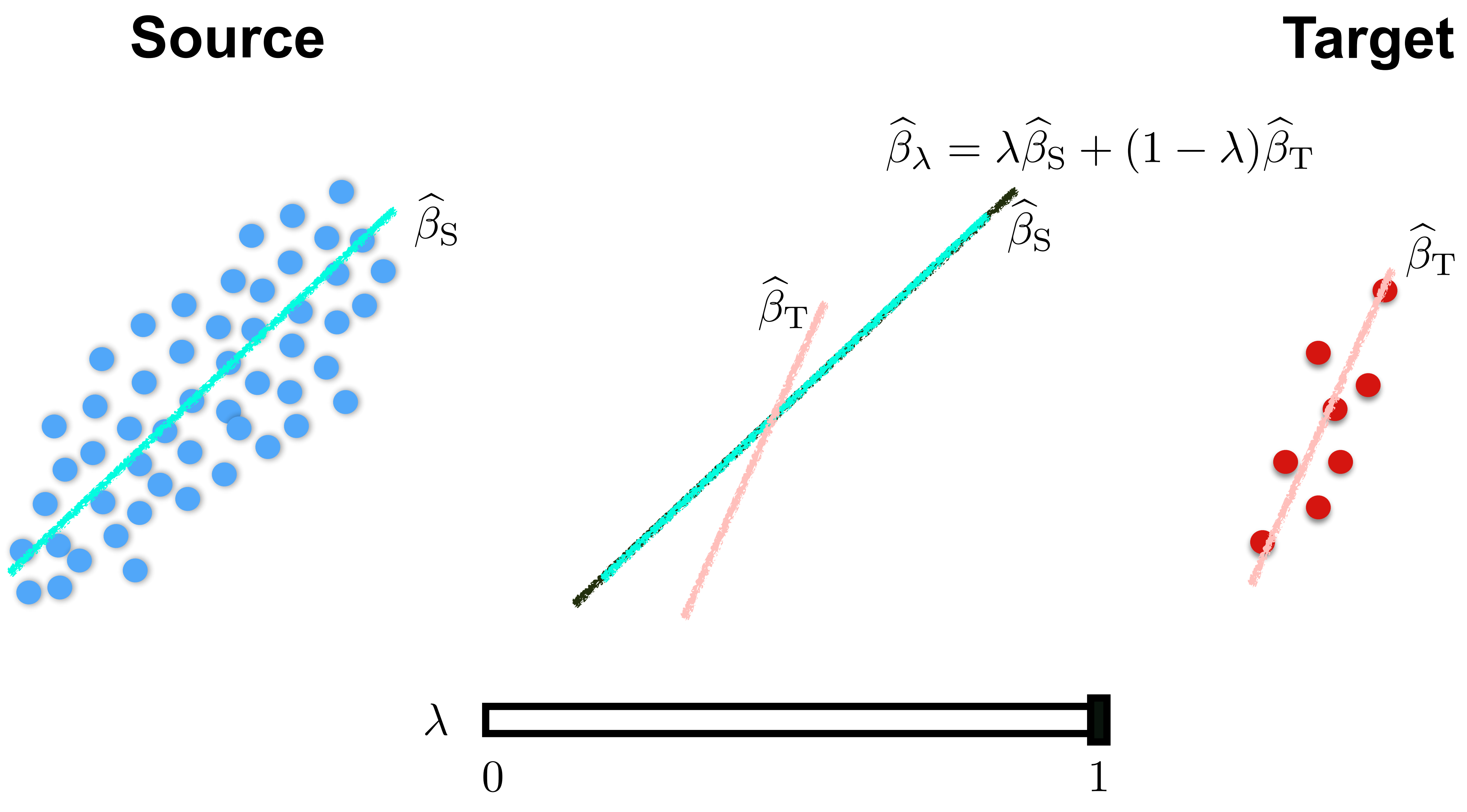
1) Convex Combination Strategy (Baseline)

Source **Target**



Exploit Source Data

1) Convex Combination Strategy (Baseline)



How to pick λ?

Exploit Source Data

2) Reweighting Strategy (RWS)[1)]

1) Garcke & Vanck, ECML PKDD, 2014

Exploit Source Data

2) Reweighting Strategy (RWS)[1)]

$$\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{N_{\mathrm{S}}} w_{h,i} (\beta^\top \widehat{x}_i - \widehat{y}_i)^2 + \sum_{j=1}^{N_{\mathrm{T}}} (\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta \|\beta\|_2^2$$

1) Garcke & Vanck, ECML PKDD, 2014

Exploit Source Data

2) Reweighting Strategy (RWS)[1]

$$\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{N_{\mathrm{S}}} w_{h,i} (\beta^\top \widehat{x}_i - \widehat{y}_i)^2 + \sum_{j=1}^{N_{\mathrm{T}}} (\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta \|\beta\|_2^2$$

$$w_{h,i} = \sum_{l=1}^{N_{\mathrm{S}}} \alpha_l \exp\left(- \frac{\|\widehat{x}_i - \widehat{x}_l\|_2^2 + (\widehat{y}_i - \widehat{y}_l)^2}{h^2} \right)$$

---

[1] Garcke & Vanck, ECML PKDD, 2014

Exploit Source Data

2) Reweighting Strategy (RWS)[1]

$$\min_{\beta\in\mathbb{R}^d} \sum_{i=1}^{N_{\mathrm{S}}} w_{h,i}(\beta^\top \widehat{x}_i - \widehat{y}_i)^2 + \sum_{j=1}^{N_{\mathrm{T}}}(\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta\|\beta\|_2^2$$

$$w_{h,i} = \sum_{l=1}^{N_{\mathrm{S}}} \alpha_l \exp\left(-\frac{\|\widehat{x}_i - \widehat{x}_l\|_2^2 + (\widehat{y}_i - \widehat{y}_l)^2}{h^2}\right)$$

How to pick h*?*

[1] Garcke & Vanck, ECML PKDD, 2014

Tune Hyperparameters - Synethesizing Experts

Expert - 1: β_1 

Expert - 2: β_2 

.

.

.

Expert - $|\mathcal{E}|$: $\beta_{|\mathcal{E}|}$ 

Tune Hyperparameters - Synthesizing Experts

Bernstein Online Aggregation (BOA)[1)]

Expert - 1: β_1

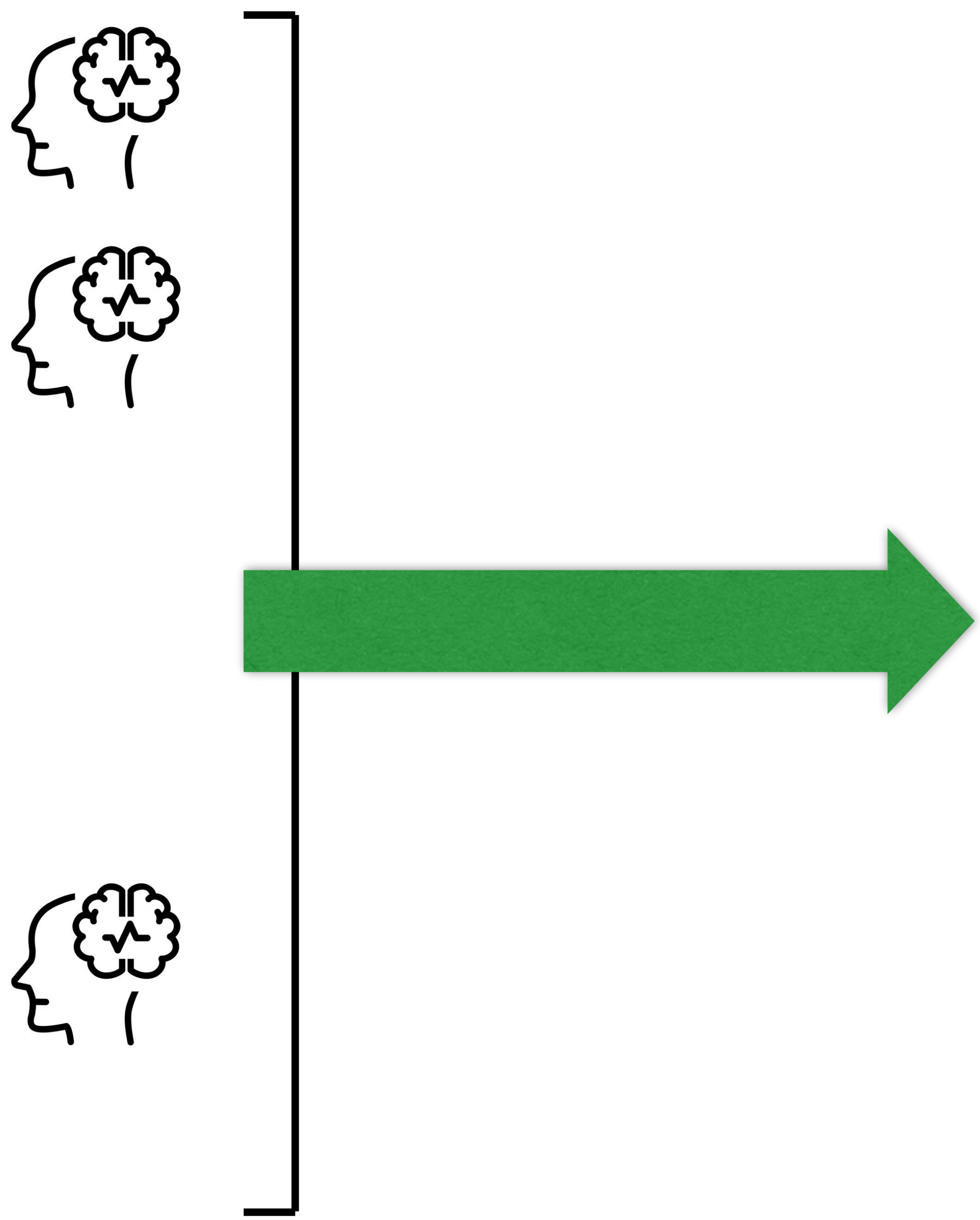
Expert - 2: β_2

.

.

.

Expert - $|\mathcal{E}|$: $\beta_{|\mathcal{E}|}$



Tune Hyperparameters - Synethesizing Experts

Bernstein Online Aggregation (BOA)[1)]

Expert - 1: β_1

Expert - 2: β_2

.

.

.

Expert - $|\mathcal{E}|$: $\beta_{|\mathcal{E}|}$

unseen target data

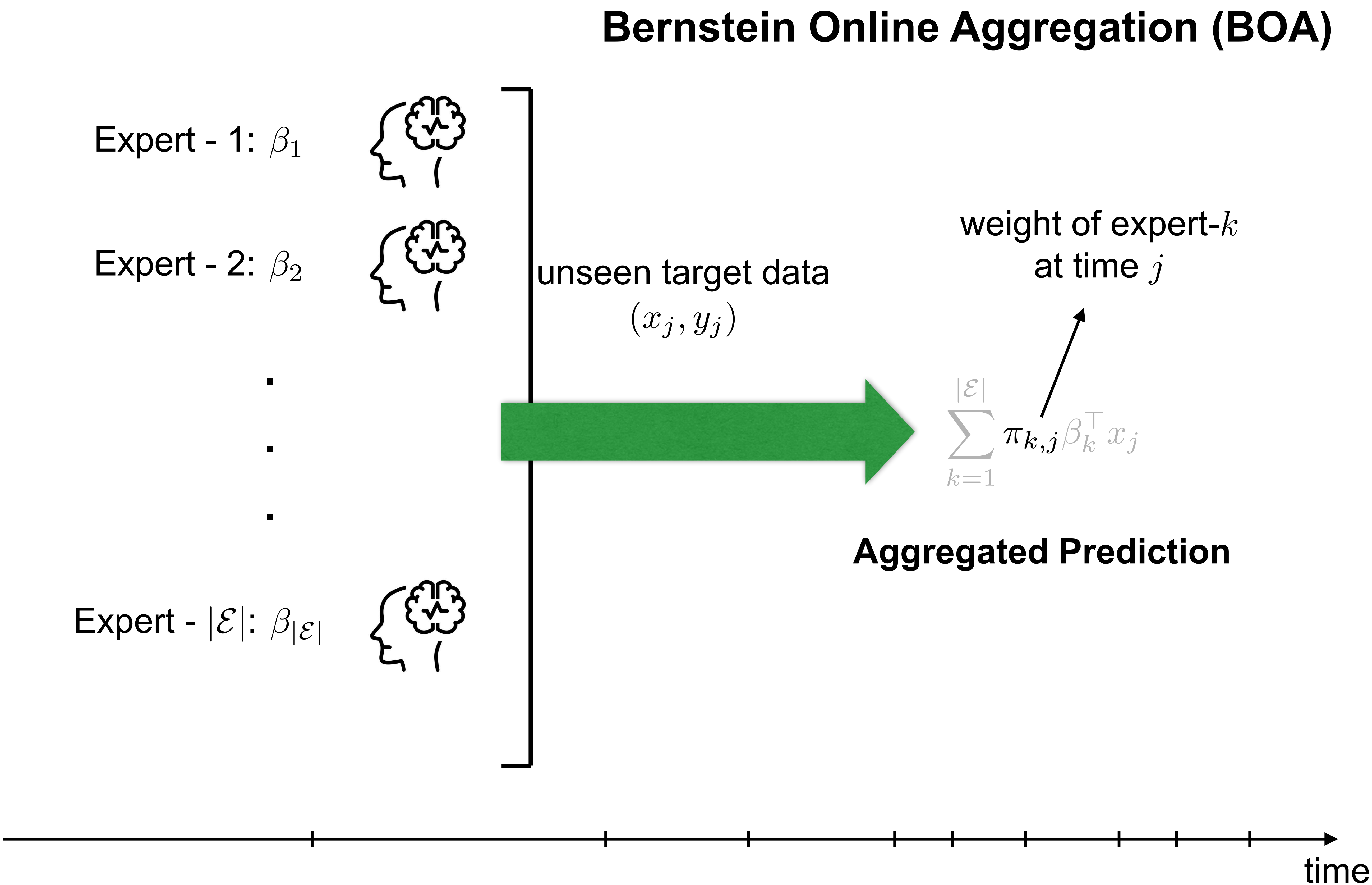
(x_j, y_j)

$$\sum_{k=1}^{|\mathcal{E}|} \pi_{k,j} \beta_k^\top x_j$$

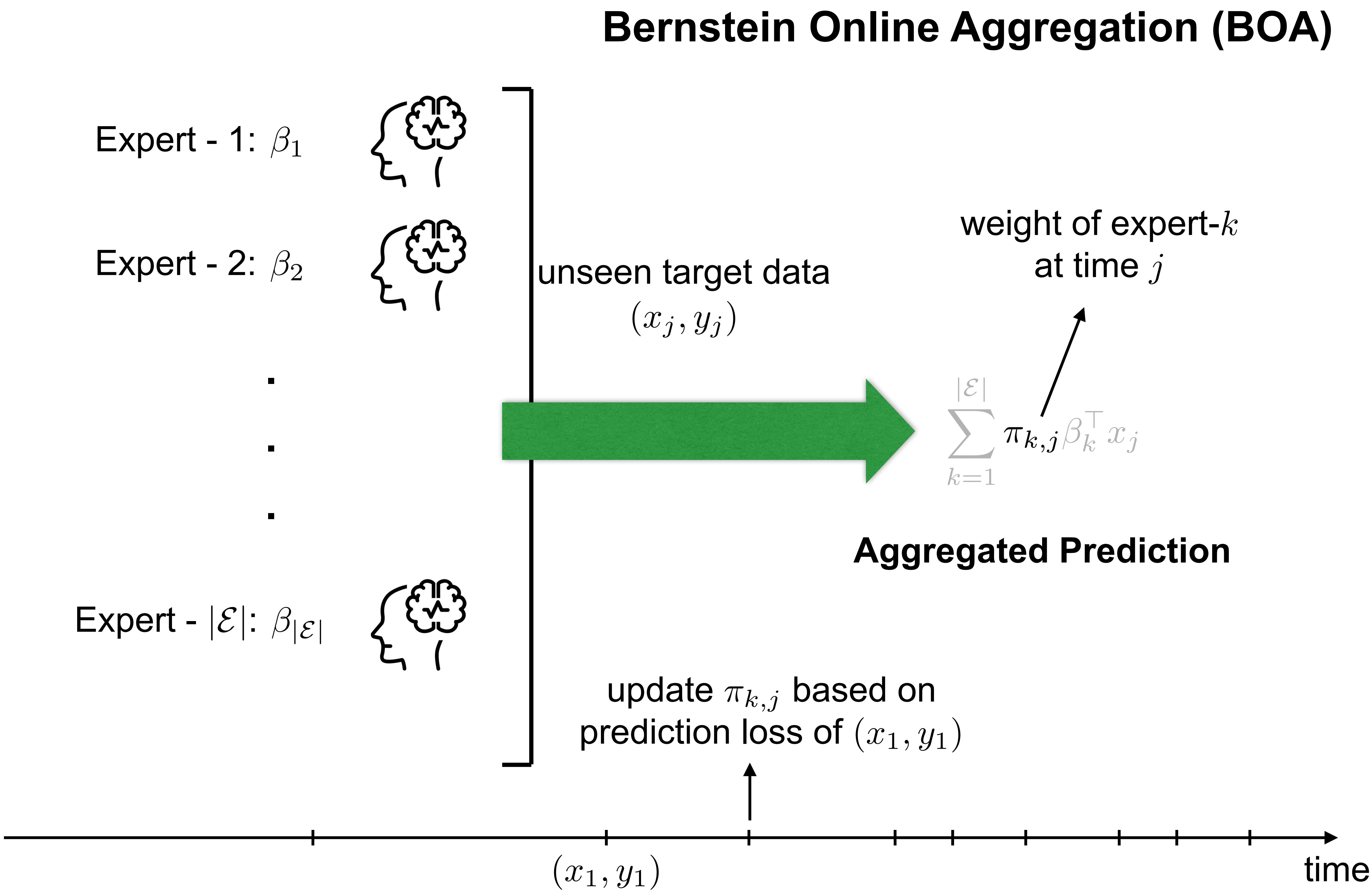
Aggregated Prediction

1) Wintenberger, Machine Learning, 2017

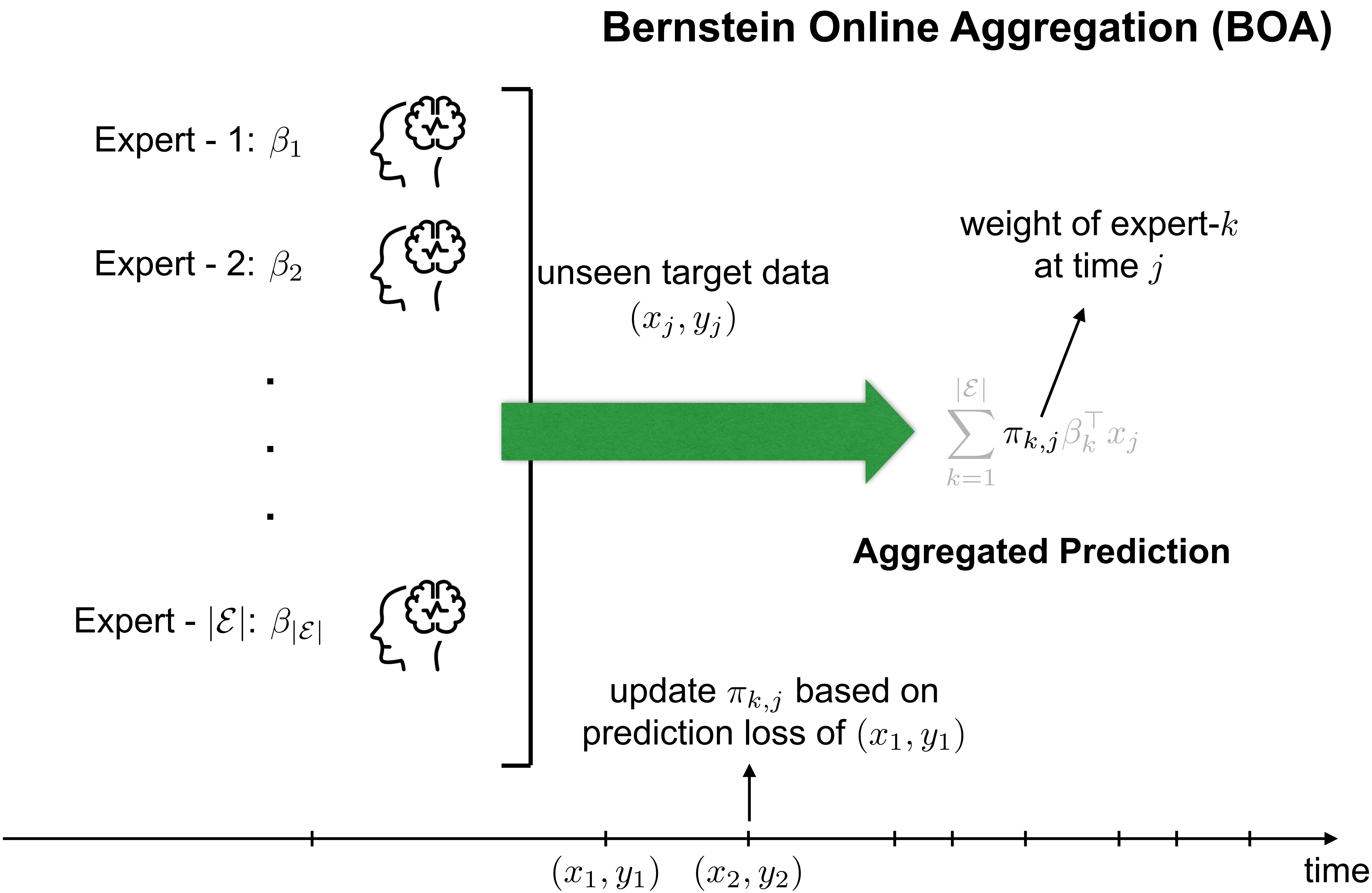
Synethesizing Experts



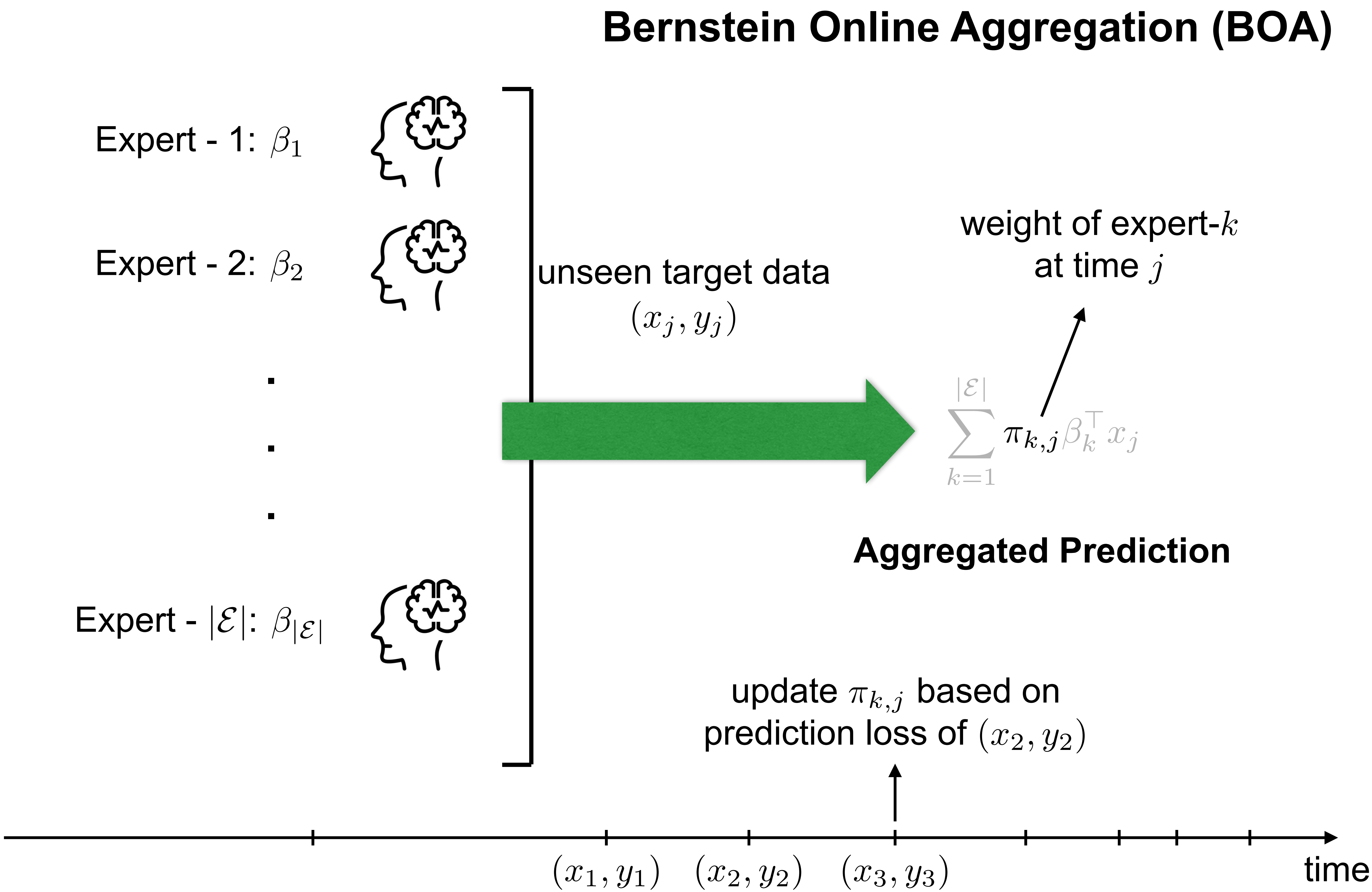
Synthesizing Experts



Synthesizing Experts

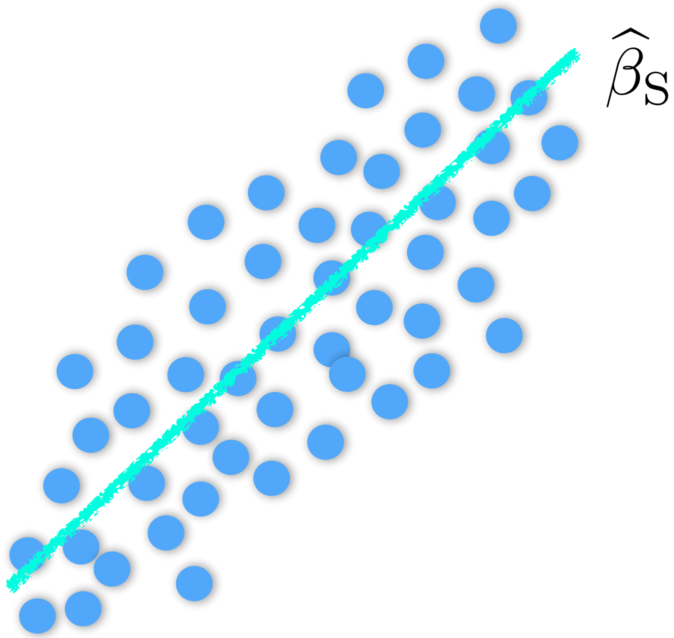


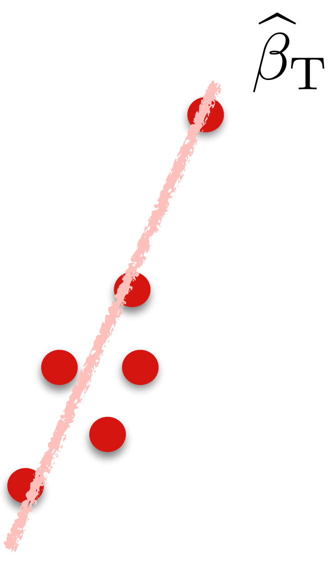
Synthesizing Experts



Distributionally Robust Expert Generation

Distributionally Robust Expert Generation





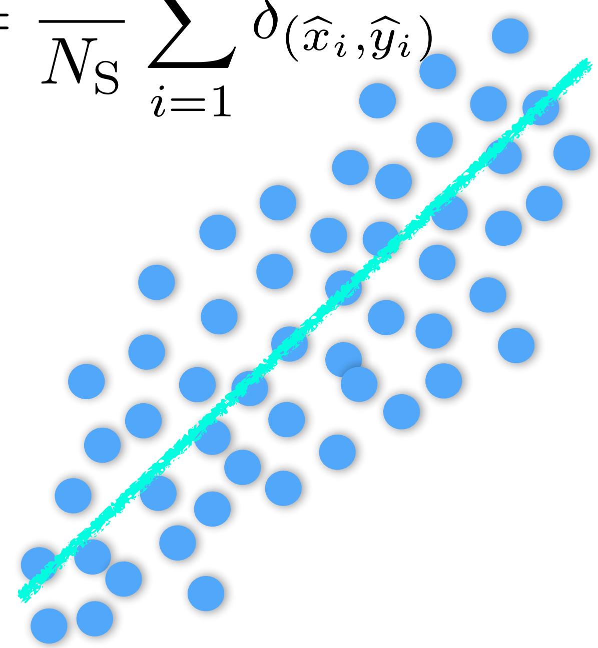
Distributionally Robust Expert Generation

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]$$

Source

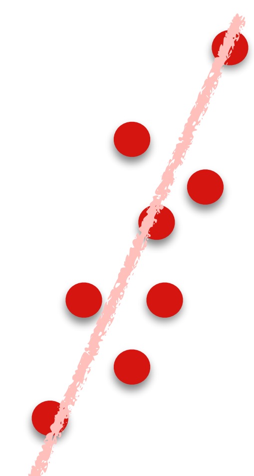
Target

$$\widehat{\mathbb{P}}_{\mathrm{S}} = \frac{1}{N_{\mathrm{S}}} \sum_{i=1}^{N_{\mathrm{S}}} \delta_{(\widehat{x}_i, \widehat{y}_i)}$$



$\widehat{\beta}_{\mathrm{S}}$

$\widehat{\beta}_{\mathrm{T}}$



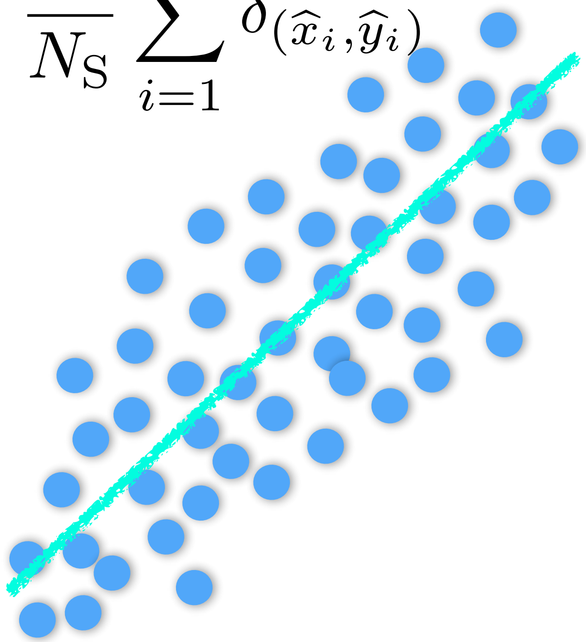
$$\widehat{\mathbb{P}}_{\mathrm{T}} = \frac{1}{N_{\mathrm{T}}} \sum_{j=1}^{N_{\mathrm{T}}} \delta_{(\widehat{x}_j, \widehat{y}_j)}$$

Distributionally Robust Expert Generation

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]$$

Source

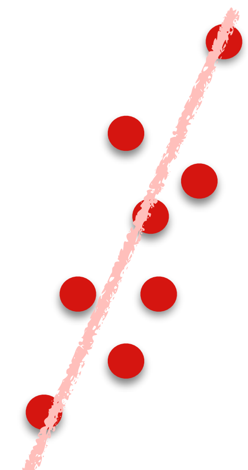
$$\widehat{\mathbb{P}}_{\mathrm{S}} = \frac{1}{N_{\mathrm{S}}} \sum_{i=1}^{N_{\mathrm{S}}} \delta_{(\widehat{x}_i, \widehat{y}_i)}$$



$\widehat{\beta}_{\mathrm{S}}$

Target

$\widehat{\beta}_{\mathrm{T}}$



$$\widehat{\mathbb{P}}_{\mathrm{T}} = \frac{1}{N_{\mathrm{T}}} \sum_{j=1}^{N_{\mathrm{T}}} \delta_{(\widehat{x}_j, \widehat{y}_j)}$$

1) **Distributional probing**

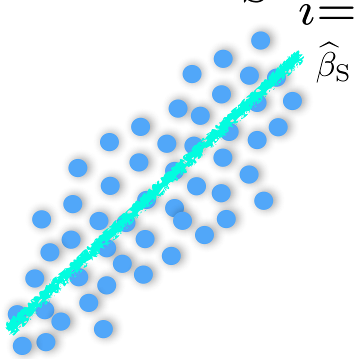
2) **Robust Estimation**

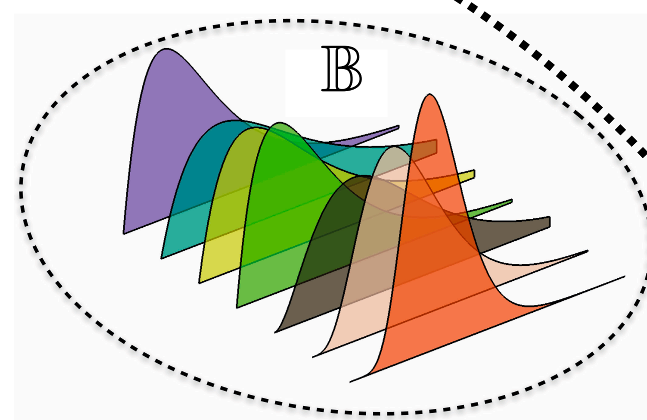
Distributionally Robust Expert Generation

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]$$

Source

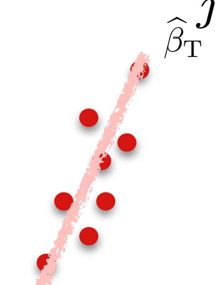
$$\widehat{\mathbb{P}}_{\mathrm{S}} = \frac{1}{N_{\mathrm{S}}} \sum_{i=1}^{N_{\mathrm{S}}} \delta_{(\widehat{x}_i, \widehat{y}_i)}$$





Target

$$\widehat{\mathbb{P}}_{\mathrm{T}} = \frac{1}{N_{\mathrm{T}}} \sum_{j=1}^{N_{\mathrm{T}}} \delta_{(\widehat{x}_j, \widehat{y}_j)}$$



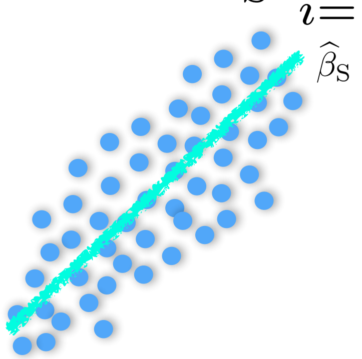
1) Distributional probing

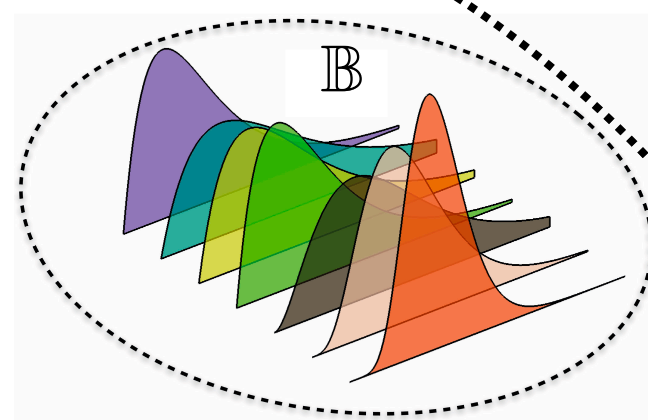
Distributionally Robust Expert Generation

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]$$

Source

$$\widehat{\mathbb{P}}_{\mathrm{S}} = \frac{1}{N_{\mathrm{S}}} \sum_{i=1}^{N_{\mathrm{S}}} \delta_{(\widehat{x}_i, \widehat{y}_i)}$$

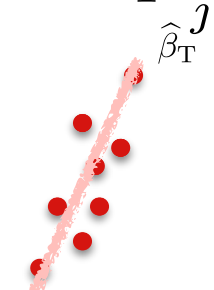




$$\mathbb{B} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu, \Sigma), (\mu, \Sigma) \in \mathbb{U}\}$$

Target

$$\widehat{\mathbb{P}}_{\mathrm{T}} = \frac{1}{N_{\mathrm{T}}} \sum_{j=1}^{N_{\mathrm{T}}} \delta_{(\widehat{x}_j, \widehat{y}_j)}$$



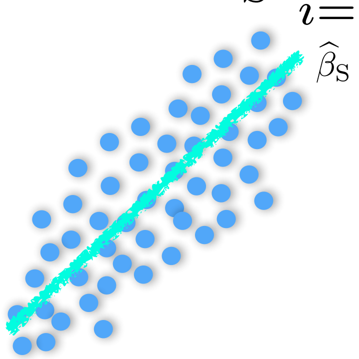
1) Distributional probing

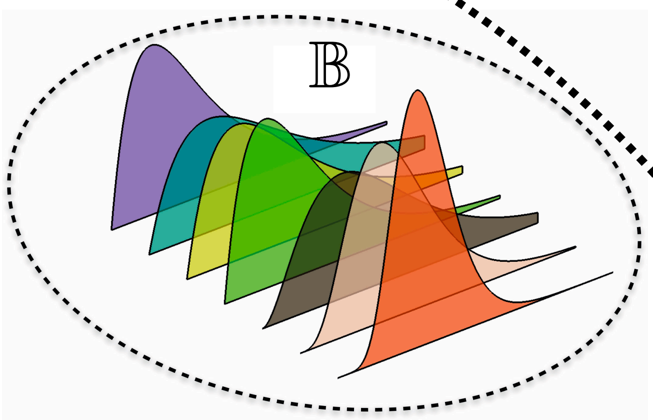
Distributionally Robust Expert Generation

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]$$

Source

$$\widehat{\mathbb{P}}_{\mathrm{S}} = \frac{1}{N_{\mathrm{S}}} \sum_{i=1}^{N_{\mathrm{S}}} \delta_{(\widehat{x}_i, \widehat{y}_i)}$$





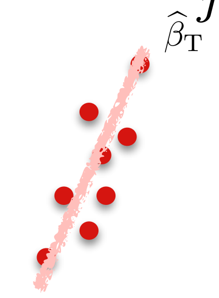
$p = d + 1$

$$\mathbb{B} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu, \Sigma), (\mu, \Sigma) \in \mathbb{U}\}$$

moment information set

Target

$$\widehat{\mathbb{P}}_{\mathrm{T}} = \frac{1}{N_{\mathrm{T}}} \sum_{j=1}^{N_{\mathrm{T}}} \delta_{(\widehat{x}_j, \widehat{y}_j)}$$

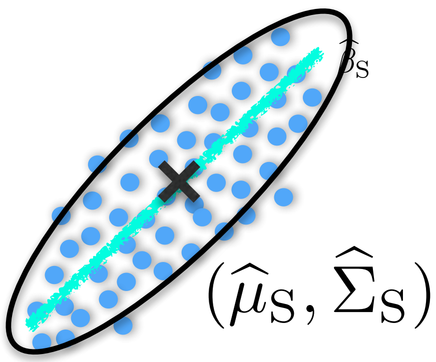


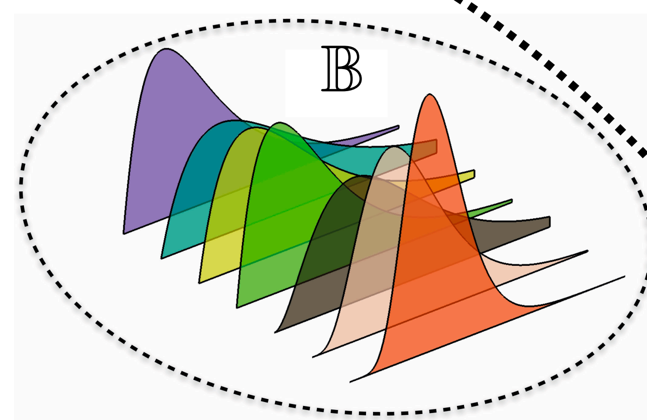
1) Distributional probing

Distributionally Robust Expert Generation

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]$$

Source



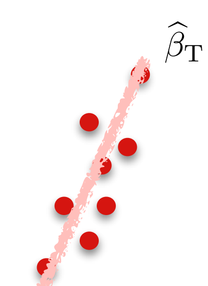


Target

$p = d + 1$

$$\mathbb{B} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu, \Sigma), (\mu, \Sigma) \in \mathbb{U}\}$$

moment
information set

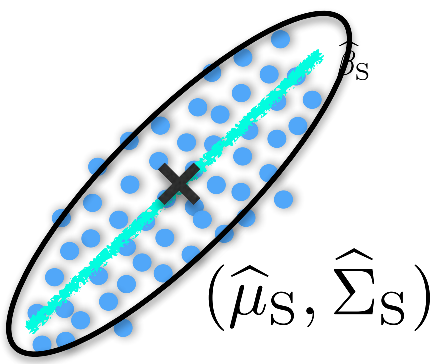


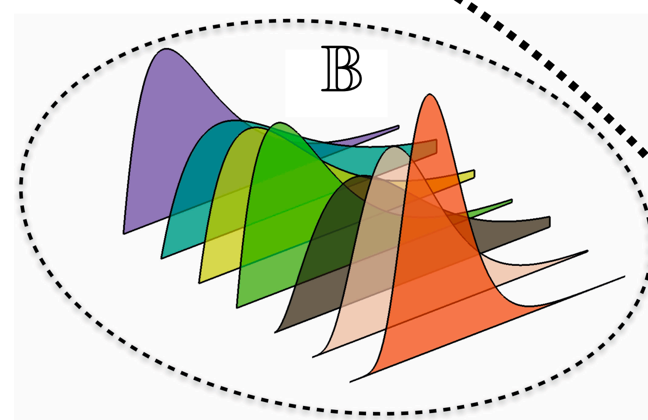
1) Distributional probing

Distributionally Robust Expert Generation

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]$$

Source



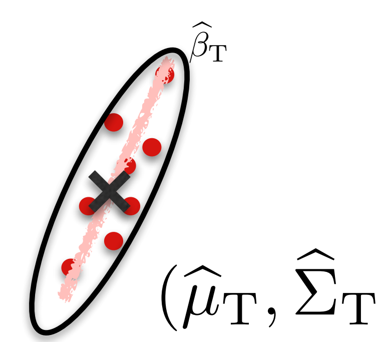


$p = d + 1$

$$\mathbb{B} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu, \Sigma), (\mu, \Sigma) \in \mathbb{U}\}$$

moment information set

Target

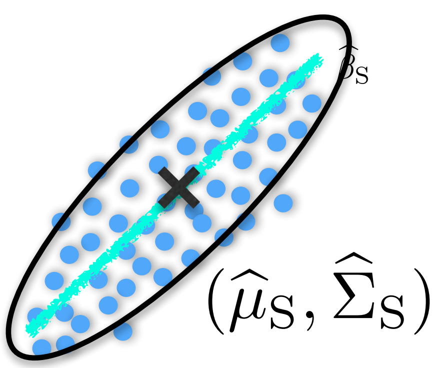


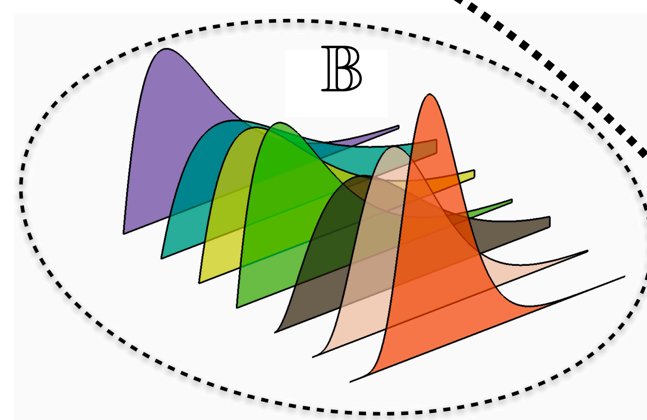
1) Distributional probing

Distributionally Robust Expert Generation

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]$$

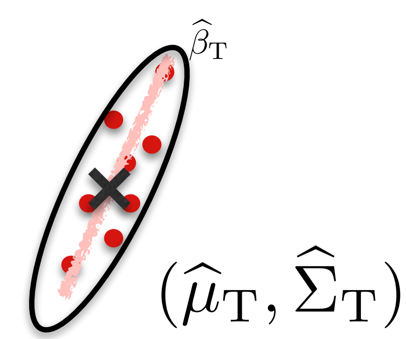
Source





Target

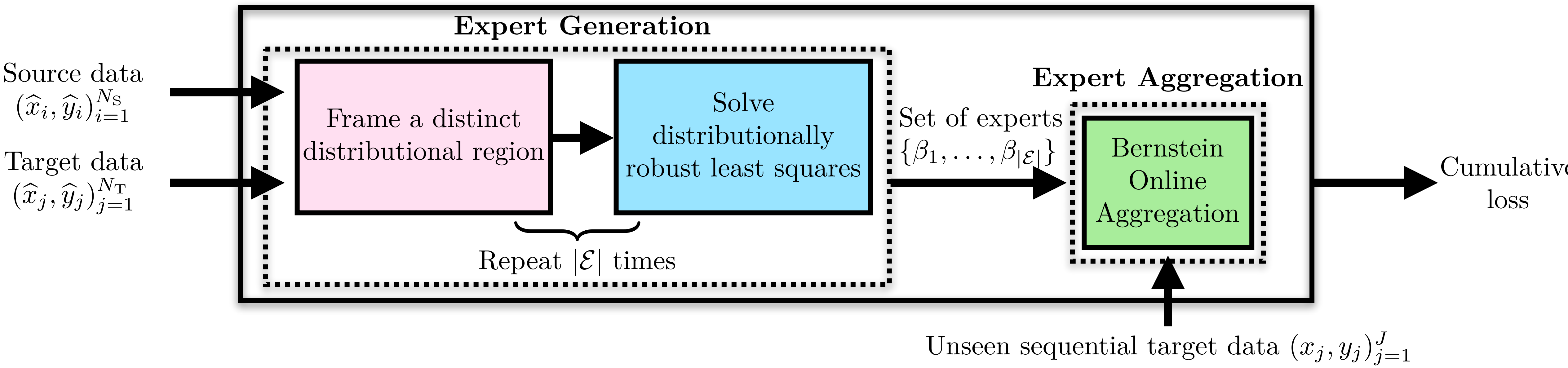
$$\mathbb{B} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu, \Sigma), (\mu, \Sigma) \in \mathbb{U}\}$$



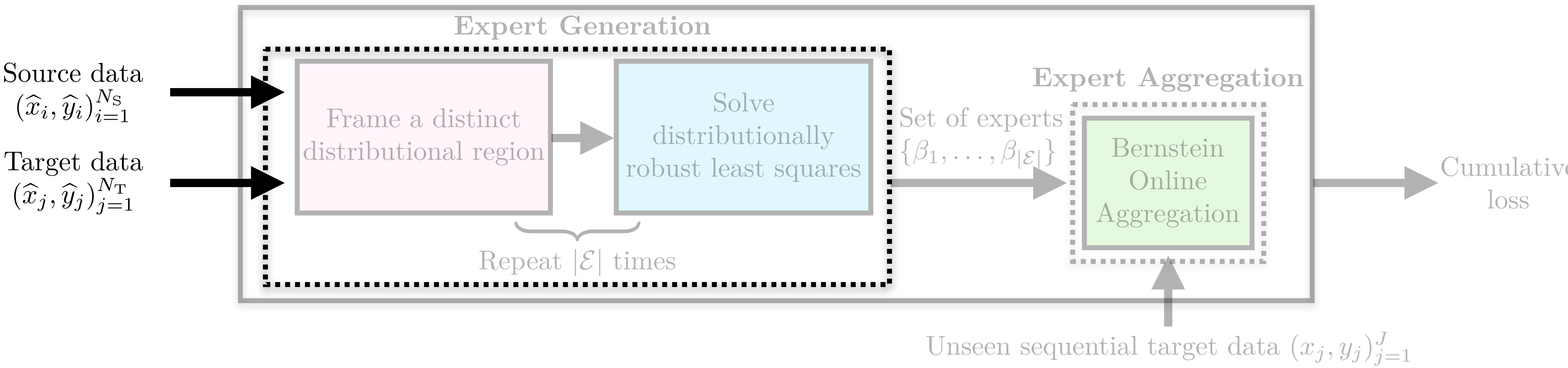
1) Distributional probing

2) Robust Estimation

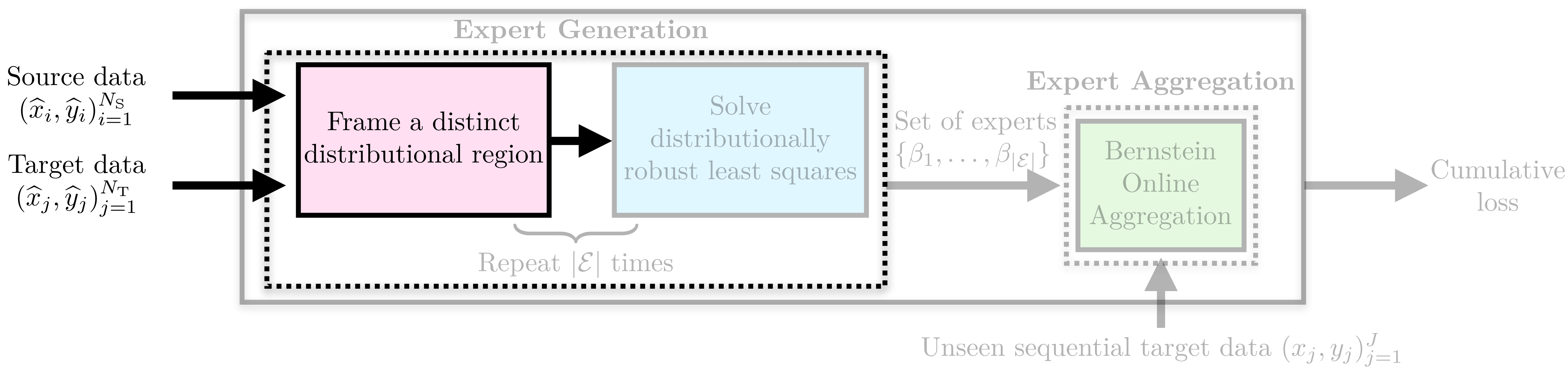
Summary of Framework



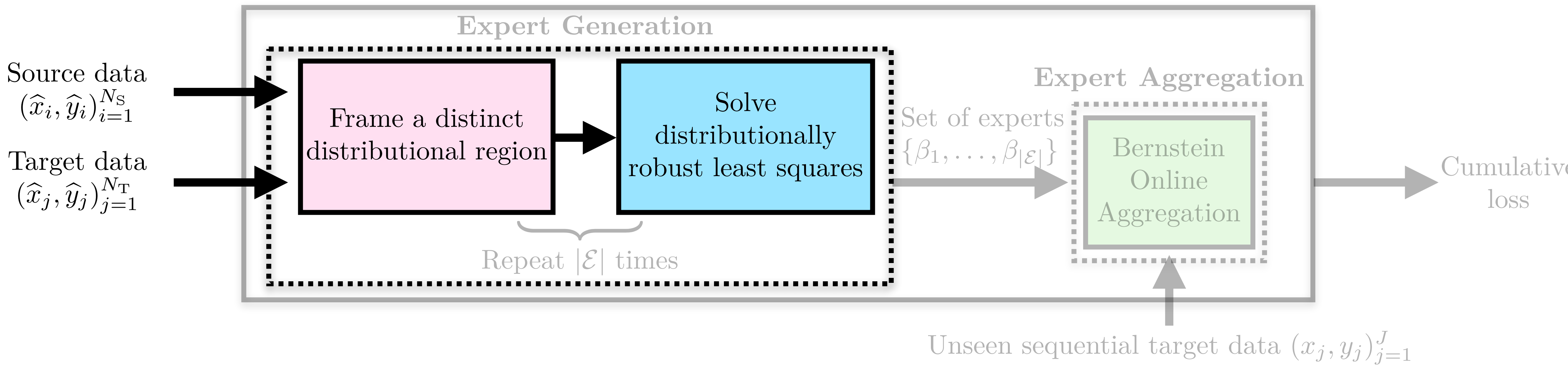
Summary of Framework



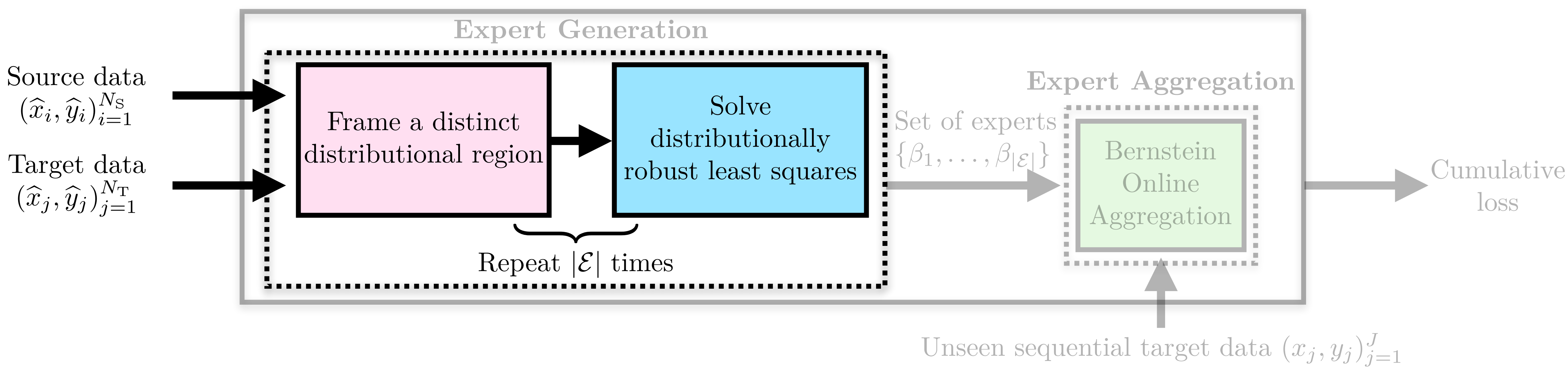
Summary of Framework



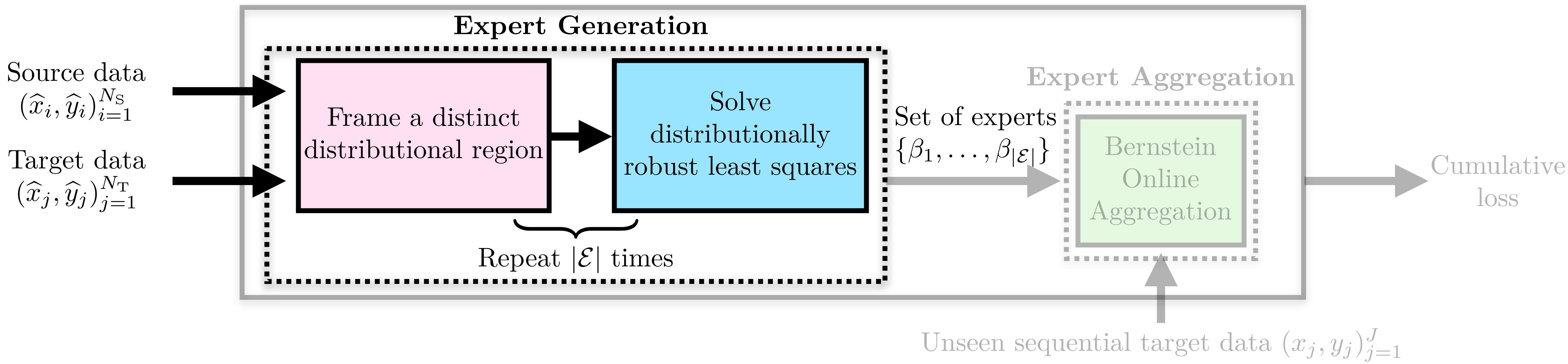
Summary of Framework



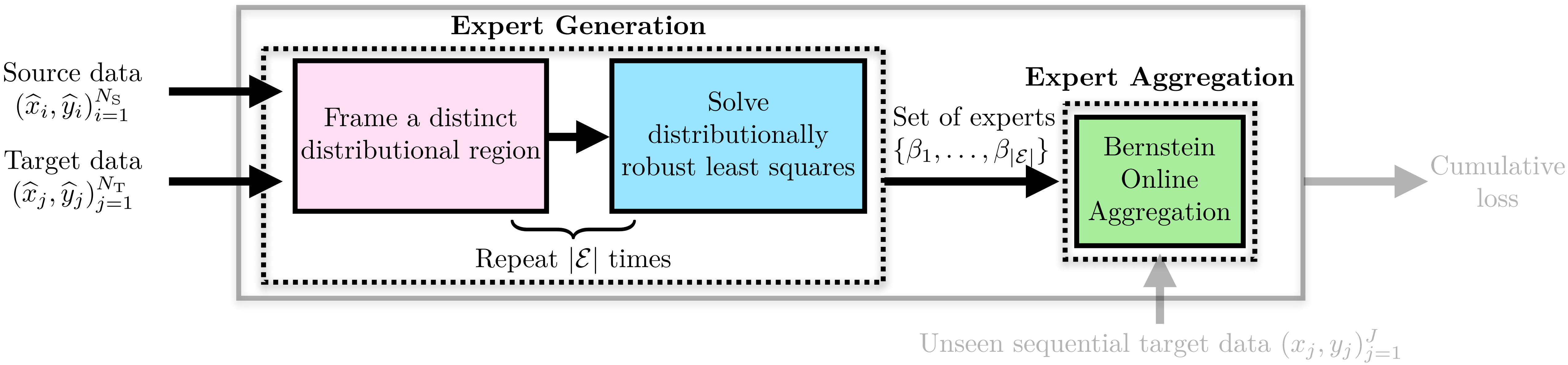
Summary of Framework



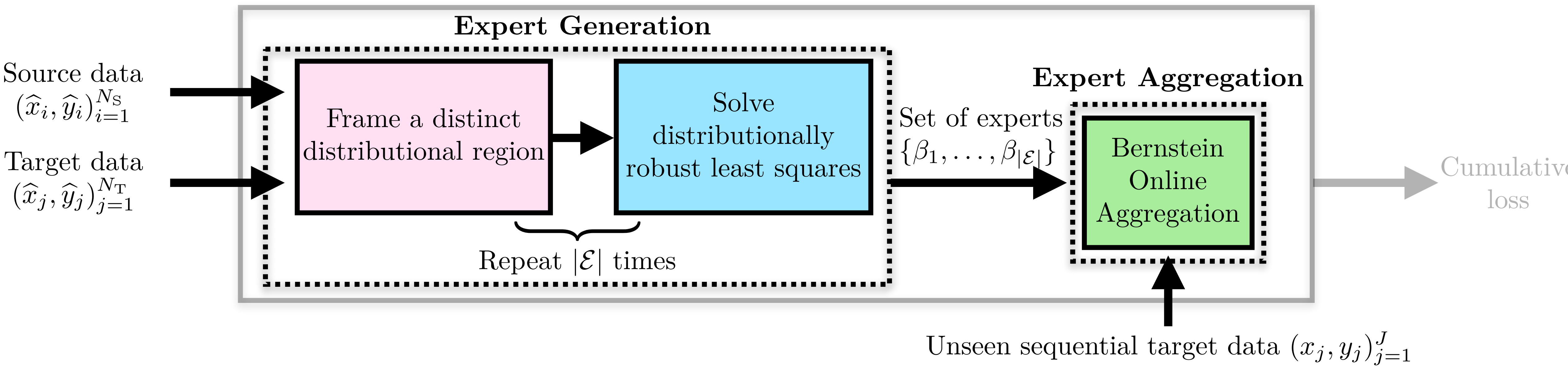
Summary of Framework



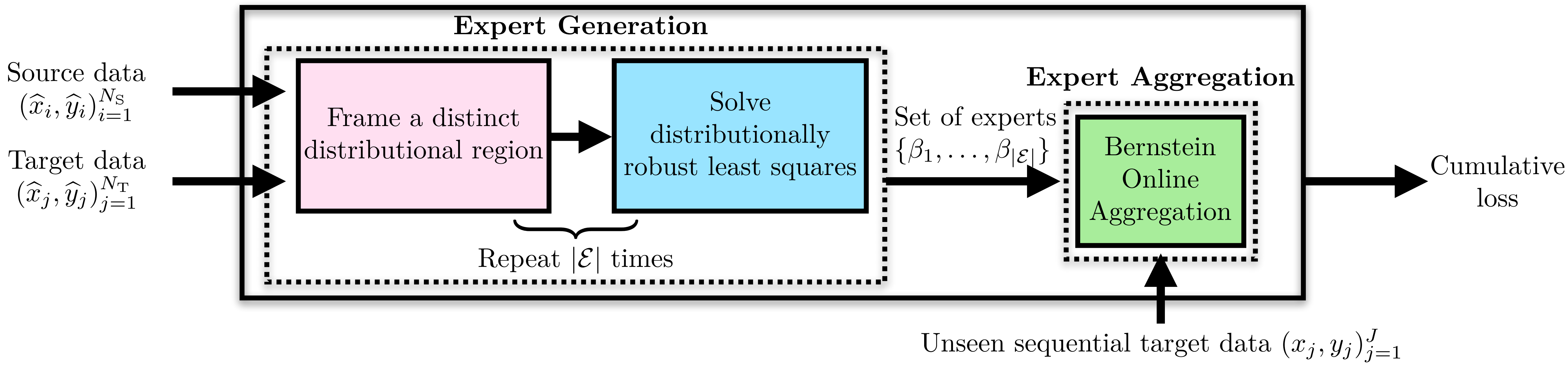
Summary of Framework



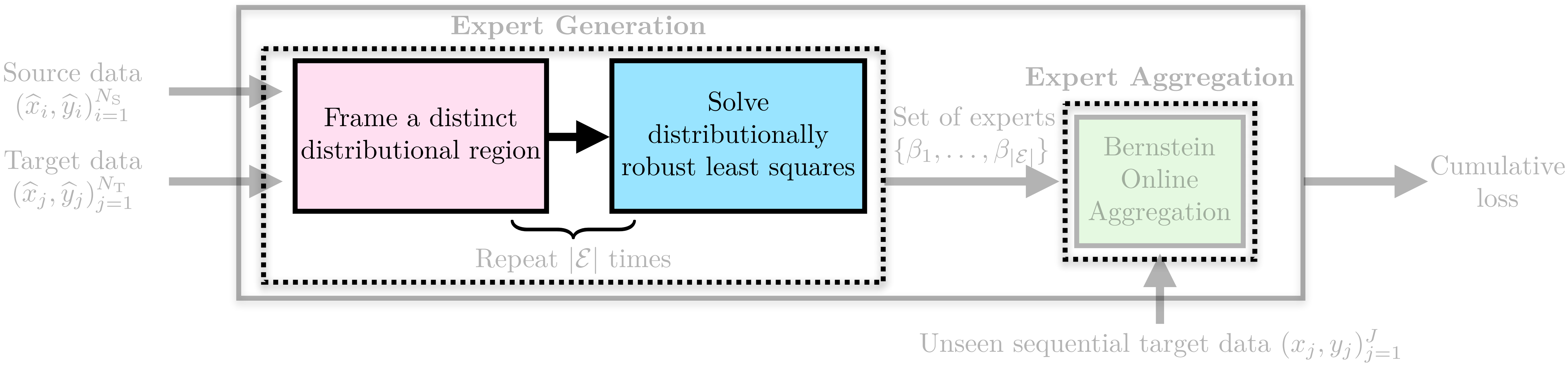
Summary of Framework



Summary of Framework



Summary of Framework

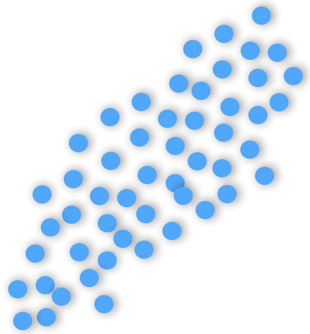


1) "Interpolate, then Robustify" (IR)

2) "Surround, then Intersect" (SI)

1) “Interpolate, then Robustify” (IR) Strategy

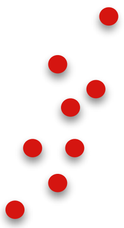
Source



$(\widehat{x}_i, \widehat{y}_i)_{i=1}^{N_{\mathrm{S}}}$

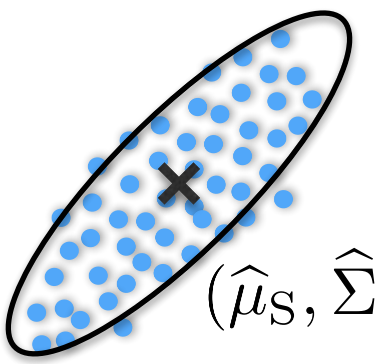
Target

$(\widehat{x}_j, \widehat{y}_j)_{j=1}^{N_{\mathrm{T}}}$



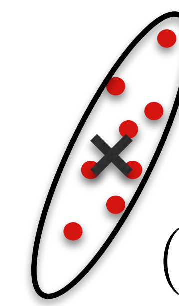
1) “Interpolate, then Robustify” (IR) Strategy

Source



$(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})$

Target



$(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$

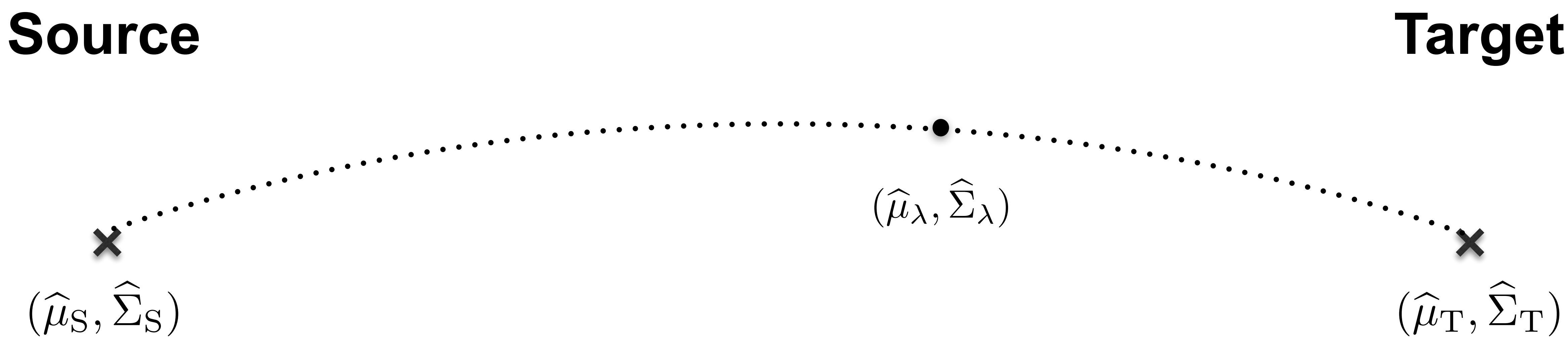
$$\widehat{\mu}_{\mathrm{S}} = \frac{1}{N_{\mathrm{S}}} \sum_{i=1}^{N_{\mathrm{S}}} \begin{pmatrix} \widehat{x}_i \\ \widehat{y}_i \end{pmatrix}$$

$$\widehat{\Sigma}_{\mathrm{S}} = \frac{1}{N_{\mathrm{S}}} \sum_{i=1}^{N_{\mathrm{S}}} \begin{pmatrix} \widehat{x}_i \\ \widehat{y}_i \end{pmatrix} \begin{pmatrix} \widehat{x}_i \\ \widehat{y}_i \end{pmatrix}^{\top} - \widehat{\mu}_{\mathrm{S}} \widehat{\mu}_{\mathrm{S}}^{\top}$$

$$\widehat{\mu}_{\mathrm{T}} = \frac{1}{N_{\mathrm{T}}} \sum_{j=1}^{N_{\mathrm{J}}} \begin{pmatrix} \widehat{x}_j \\ \widehat{y}_j \end{pmatrix}$$

$$\widehat{\Sigma}_{\mathrm{T}} = \frac{1}{N_{\mathrm{T}}} \sum_{j=1}^{N_{\mathrm{T}}} \begin{pmatrix} \widehat{x}_j \\ \widehat{y}_j \end{pmatrix} \begin{pmatrix} \widehat{x}_j \\ \widehat{y}_j \end{pmatrix}^{\top} - \widehat{\mu}_{\mathrm{T}} \widehat{\mu}_{\mathrm{T}}^{\top}$$

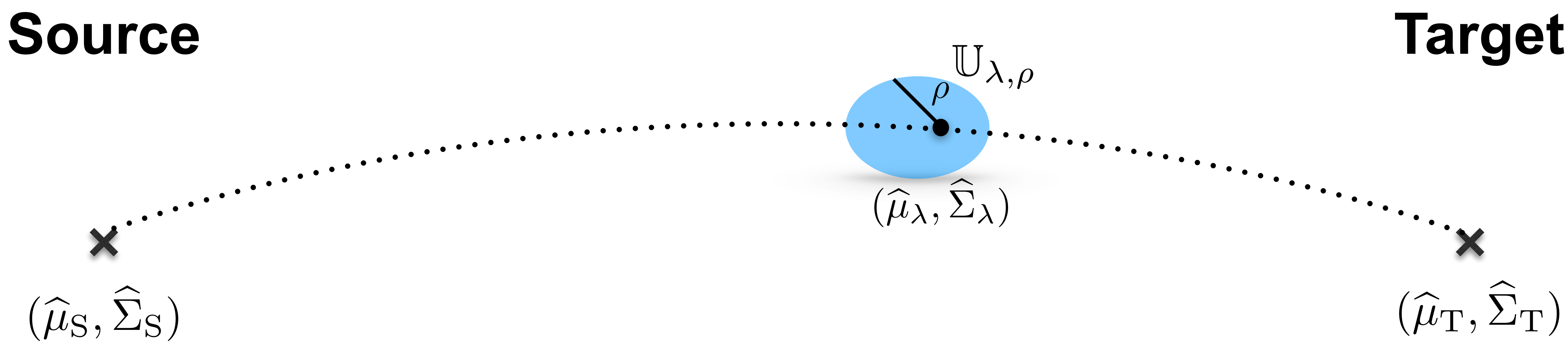
1) “Interpolate, then Robustify” (IR) Strategy



ψ-barycenter

$$(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda) = \underset{\mu \in \mathbb{R}^p, \Sigma \in \mathbb{S}_+^p}{\operatorname{argmin}} \; \lambda \psi((\mu, \Sigma) \| (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) + (1 - \lambda) \psi((\mu, \Sigma) \| (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}))$$

1) "Interpolate, then Robustify" (IR) Strategy



<u>ψ-barycenter</u>

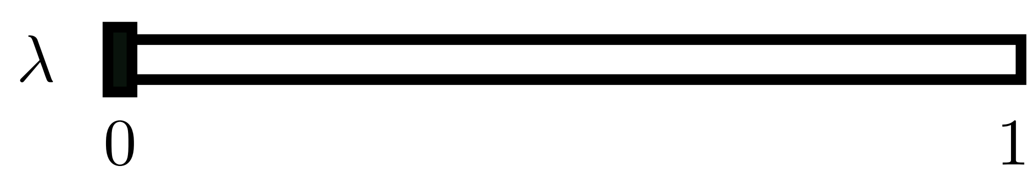
$$(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda) = \operatorname*{argmin}_{\mu \in \mathbb{R}^p, \Sigma \in \mathbb{S}_+^p} \lambda \psi((\mu, \Sigma) \| (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) + (1-\lambda) \psi((\mu, \Sigma) \| (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}))$$

<u>Moment information set</u>

$$\mathbb{U}_{\lambda,\rho} = \{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p : \psi((\mu, \Sigma) \| (\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)) \leq \rho\}$$

1) “Interpolate, then Robustify” (IR) Strategy

Source **Target**



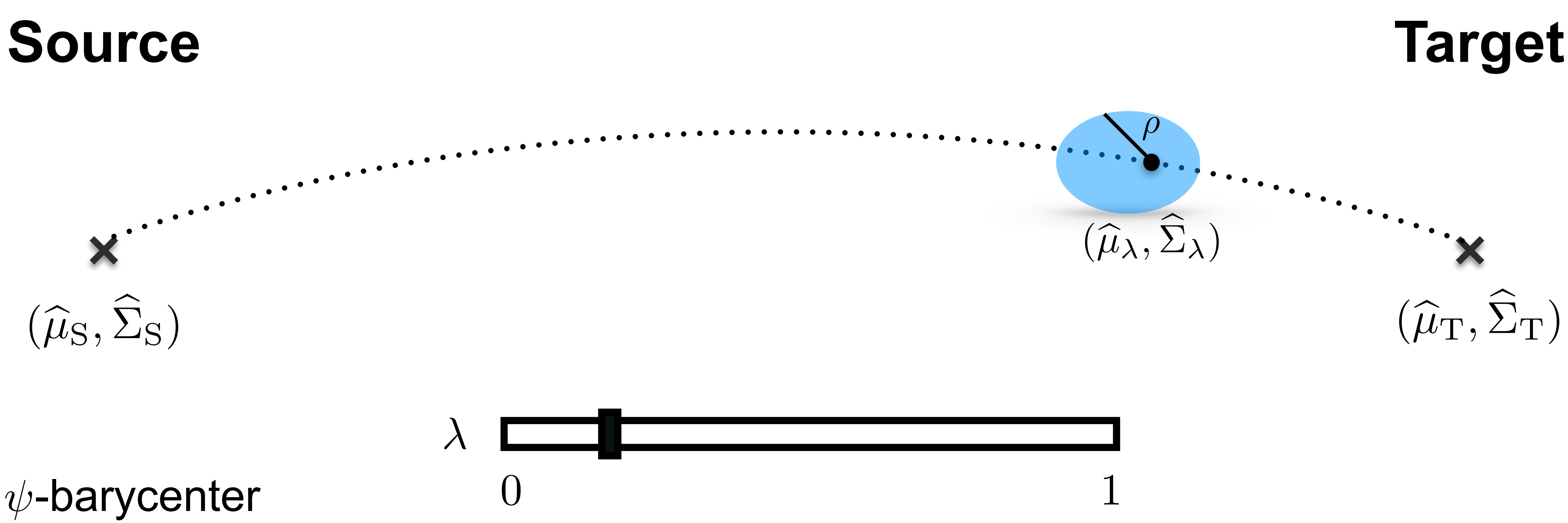
<u>ψ-barycenter</u>

$$(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda) = \underset{\mu \in \mathbb{R}^p, \Sigma \in \mathbb{S}_+^p}{\operatorname{argmin}} \; \lambda \psi((\mu, \Sigma) \| (\widehat{\mu}_\mathrm{S}, \widehat{\Sigma}_\mathrm{S})) + (1-\lambda) \psi((\mu, \Sigma) \| (\widehat{\mu}_\mathrm{T}, \widehat{\Sigma}_\mathrm{T}))$$

<u>Moment information set</u>

$$\mathbb{U}_{\lambda,\rho} = \{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p : \psi((\mu, \Sigma) \| (\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)) \le \rho\}$$

1) "Interpolate, then Robustify" (IR) Strategy



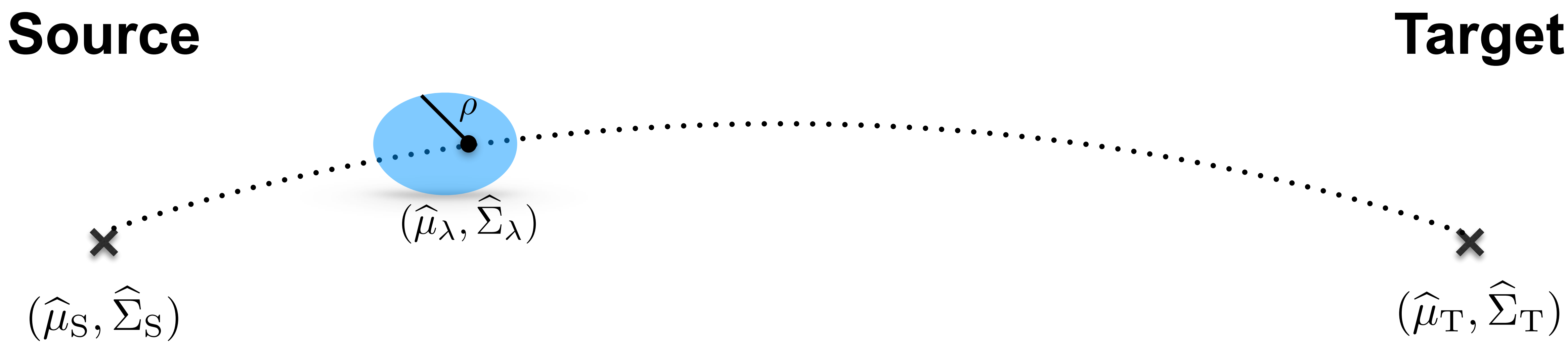
ψ-barycenter

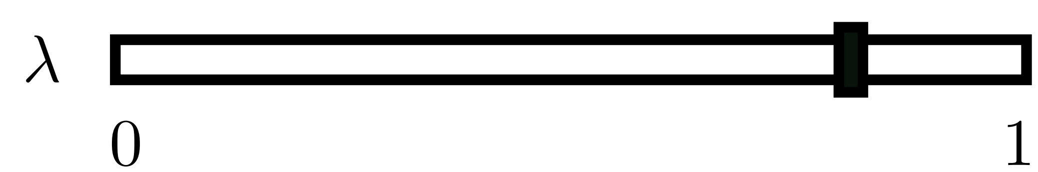
$$(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda) = \operatorname*{argmin}_{\mu \in \mathbb{R}^p, \Sigma \in \mathbb{S}_+^p} \lambda \psi((\mu, \Sigma) \| (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) + (1-\lambda)\psi((\mu, \Sigma) \| (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}))$$

Moment information set

$$\mathbb{U}_{\lambda,\rho} = \{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p : \psi((\mu, \Sigma) \| (\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)) \leq \rho\}$$

1) “Interpolate, then Robustify” (IR) Strategy





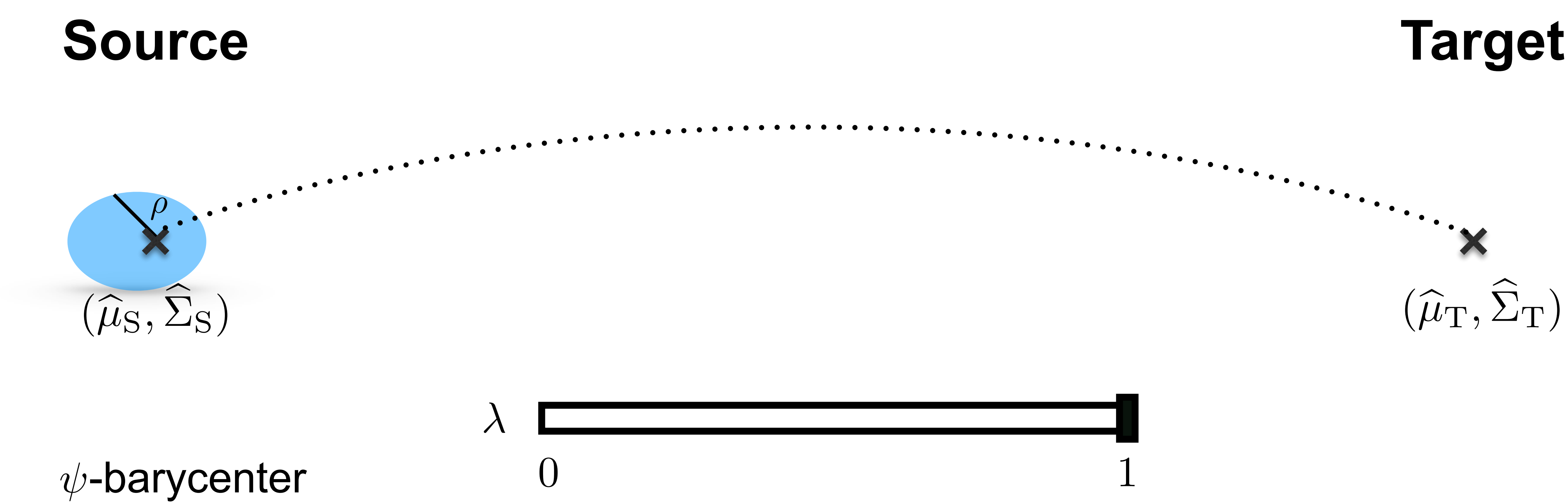
<u>ψ-barycenter</u>

$$(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda) = \operatorname*{argmin}_{\mu \in \mathbb{R}^p, \Sigma \in \mathbb{S}_+^p} \lambda \psi((\mu, \Sigma) \| (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) + (1-\lambda) \psi((\mu, \Sigma) \| (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}))$$

<u>Moment information set</u>

$$\mathbb{U}_{\lambda,\rho} = \{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p : \psi((\mu, \Sigma) \| (\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)) \leq \rho\}$$

1) "Interpolate, then Robustify" (IR) Strategy



<ins>ψ-barycenter</ins>

$$(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda) = \underset{\mu \in \mathbb{R}^p, \Sigma \in \mathbb{S}_+^p}{\operatorname{argmin}} \ \lambda \psi((\mu, \Sigma) \| (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) + (1-\lambda)\psi((\mu, \Sigma) \| (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}))$$

<ins>Moment information set</ins>

$$\mathbb{U}_{\lambda,\rho} = \{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p : \psi((\mu, \Sigma) \| (\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)) \le \rho\}$$

1) "Interpolate, then Robustify" (IR) Strategy

i) Kullback-Leibler (KL)-type Divergence

$$\mathbb{D}((\mu,\Sigma)\|(\widehat{\mu},\widehat{\Sigma})) = (\widehat{\mu}-\mu)^\top \widehat{\Sigma}^{-1}(\widehat{\mu}-\mu) + \mathrm{Tr}[\Sigma\widehat{\Sigma}^{-1}] - \log\det(\Sigma\widehat{\Sigma}^{-1}) - \mathrm{p}$$

KL barycenter:

$$\widehat{\Sigma}_\lambda = (\lambda\widehat{\Sigma}_{\mathrm{S}}^{-1} + (1-\lambda)\widehat{\Sigma}_{\mathrm{T}}^{-1})^{-1} \qquad \widehat{\mu}_\lambda = \widehat{\Sigma}_\lambda(\lambda\widehat{\Sigma}_{\mathrm{S}}^{-1}\widehat{\mu}_{\mathrm{S}} + (1-\lambda)\widehat{\Sigma}_{\mathrm{T}}^{-1}\widehat{\mu}_{\mathrm{T}})$$

Moment information set:

$$\mathbb{U}_{\lambda,\rho} = \{(\mu,\Sigma)\in\mathbb{R}^p\times\mathbb{S}_+^p : \mathbb{D}((\mu,\Sigma)\|(\widehat{\mu}_\lambda,\widehat{\Sigma}_\lambda))\le\rho\}$$

Distribution set:

$$\mathbb{B}_{\lambda,\rho} = \{\mathbb{Q}\in\mathcal{M}(\mathbb{R}^p) : \mathbb{Q}\sim(\mu,\Sigma), (\mu,\Sigma)\in\mathbb{U}_{\lambda,\rho}\}$$

Robust Estimation:

$$\min_{\beta\in\mathbb{R}^d}\left\{\sup_{\mathbb{Q}\in\mathbb{B}_{\lambda,\rho}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]\right\}$$

convex, continuously differentiable, and locally smooth

Global optimality via adaptive gradient descent algorithm

1) “Interpolate, then Robustify” (IR) Strategy

ii) Wasserstein-type Divergence

$$\mathbb{W}((\mu,\Sigma)\|(\widehat{\mu},\widehat{\Sigma})) = \|\mu-\widehat{\mu}\|_2^2 + \mathrm{Tr}[\Sigma+\widehat{\Sigma}-2(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}]$$

Wasserstein interpolation:

$$\widehat{\mu}_\lambda = \lambda\widehat{\mu}_{\mathrm{S}} + (1-\lambda)\widehat{\mu}_{\mathrm{T}} \qquad \widehat{\Sigma}_\lambda = (\lambda I_p + (1-\lambda)L)\widehat{\Sigma}_{\mathrm{S}}(\lambda I_p + (1-\lambda)L)$$

$$L = \widehat{\Sigma}_{\mathrm{T}}^{\frac{1}{2}}(\widehat{\Sigma}_{\mathrm{T}}^{\frac{1}{2}}\widehat{\Sigma}_{\mathrm{S}}\widehat{\Sigma}_{\mathrm{T}}^{\frac{1}{2}})^{-\frac{1}{2}}\widehat{\Sigma}_{\mathrm{T}}^{\frac{1}{2}}$$

Moment information set:

$$\mathbb{U}_{\lambda,\rho} = \{(\mu,\Sigma)\in\mathbb{R}^p\times\mathbb{S}_+^p : \mathbb{W}((\mu,\Sigma)\|(\widehat{\mu}_\lambda,\widehat{\Sigma}_\lambda))\le\rho\}$$

Distribution set:

$$\mathbb{B}_{\lambda,\rho} = \{\mathbb{Q}\in\mathcal{M}(\mathbb{R}^p) : \mathbb{Q}\sim(\mu,\Sigma),(\mu,\Sigma)\in\mathbb{U}_{\lambda,\rho}\}$$

Robust Estimation:

$$\min_{\beta\in\mathbb{R}^d}\sup_{\mathbb{Q}\in\mathbb{B}_{\lambda,\rho}}\mathbb{E}_{\mathbb{Q}}[(\beta^\top X-Y)^2] \equiv \min_{\beta\in\mathbb{R}^d}\left\|(\widehat{\Sigma}_\lambda+\widehat{\mu}_\lambda\widehat{\mu}_\lambda^\top)^{\frac{1}{2}}\begin{bmatrix}\beta\\-1\end{bmatrix}\right\|_2 + \sqrt{\rho}\left\|\begin{bmatrix}\beta\\-1\end{bmatrix}\right\|_2$$

SOCP

2) "Surround, then Intersect" (SI) Strategy

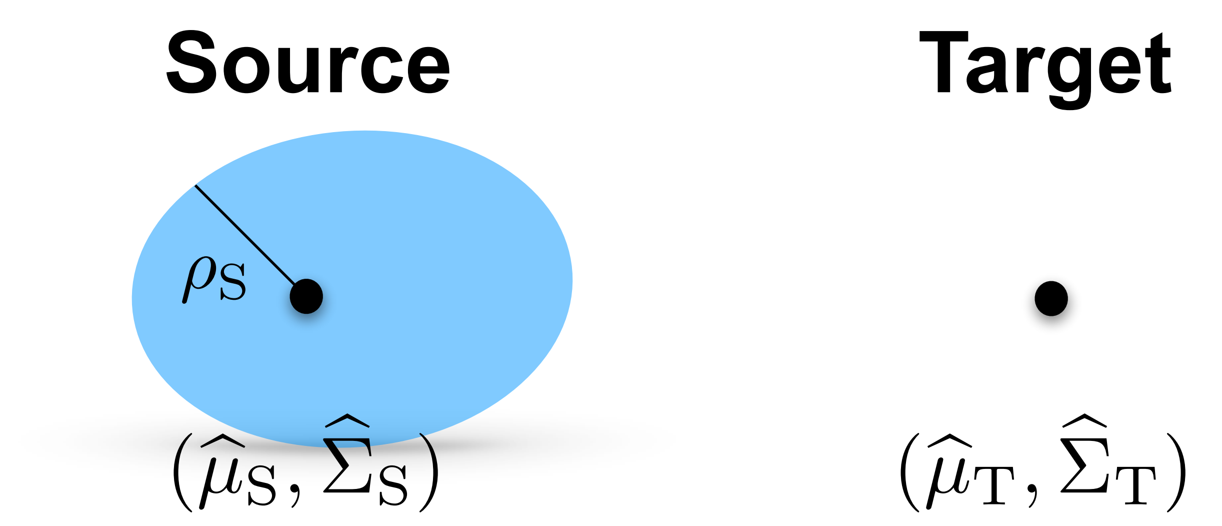
Source **Target**

$(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})$ $(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$

Moment information set:

$$\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} \triangleq \left\{ (\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p \text{ such that: } \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) \leq \rho_{\mathrm{S}}, \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})) \leq \rho_{\mathrm{T}}, \Sigma + \mu\mu^\top \succeq \varepsilon I_p \right\}$$

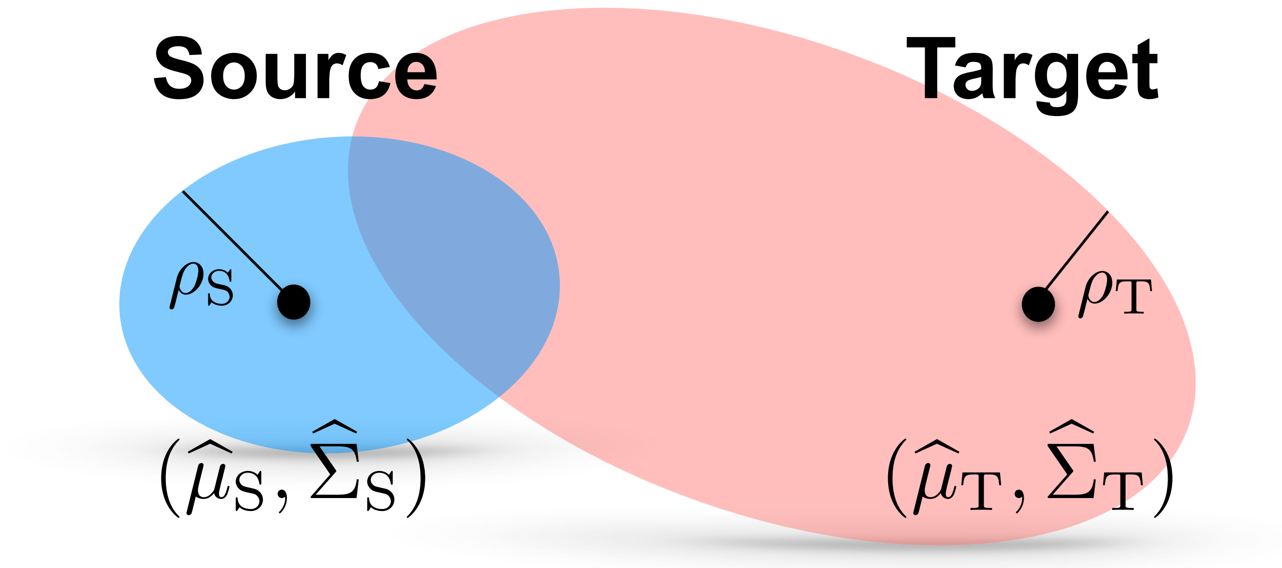
2) "Surround, then Intersect" (SI) Strategy



Moment information set:

$$\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} \triangleq \left\{ (\mu,\Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p \text{ such that: } \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{S}},\widehat{\Sigma}_{\mathrm{S}})) \le \rho_{\mathrm{S}}, \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{T}},\widehat{\Sigma}_{\mathrm{T}})) \le \rho_{\mathrm{T}}, \Sigma + \mu\mu^\top \succeq \varepsilon I_p \right\}$$

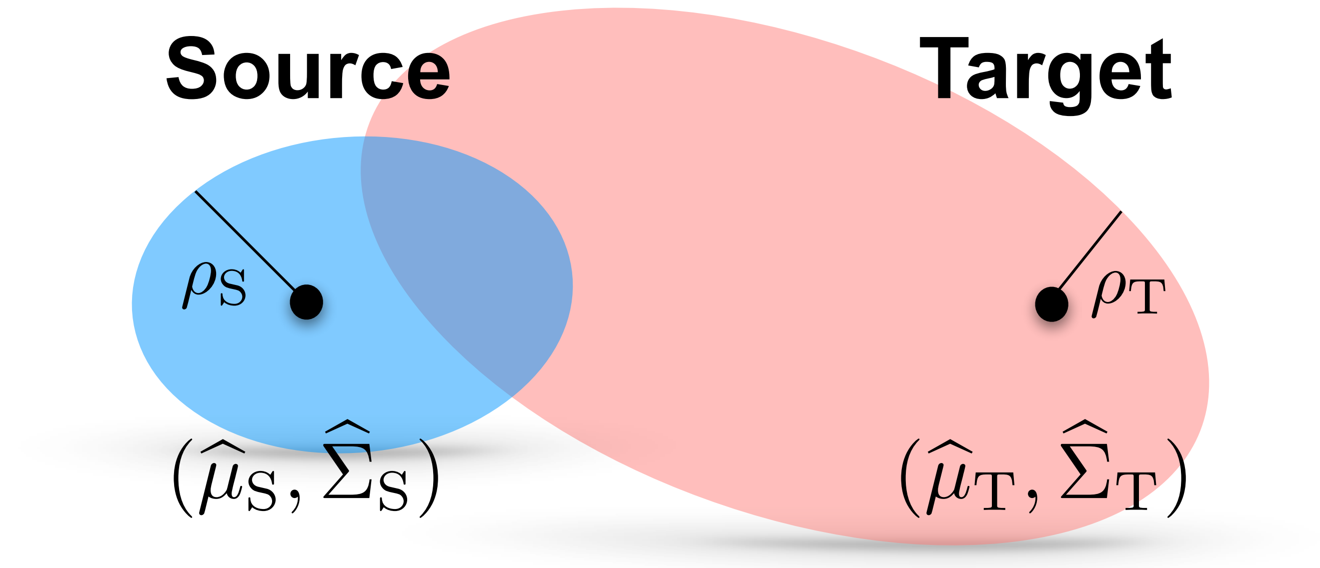
2) "Surround, then Intersect" (SI) Strategy



Moment information set:

$$\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} \triangleq \left\{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}^p_+ \text{ such that: } \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) \leq \rho_{\mathrm{S}}, \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})) \leq \rho_{\mathrm{T}}, \Sigma + \mu\mu^\top \succeq \varepsilon I_p \right\}$$

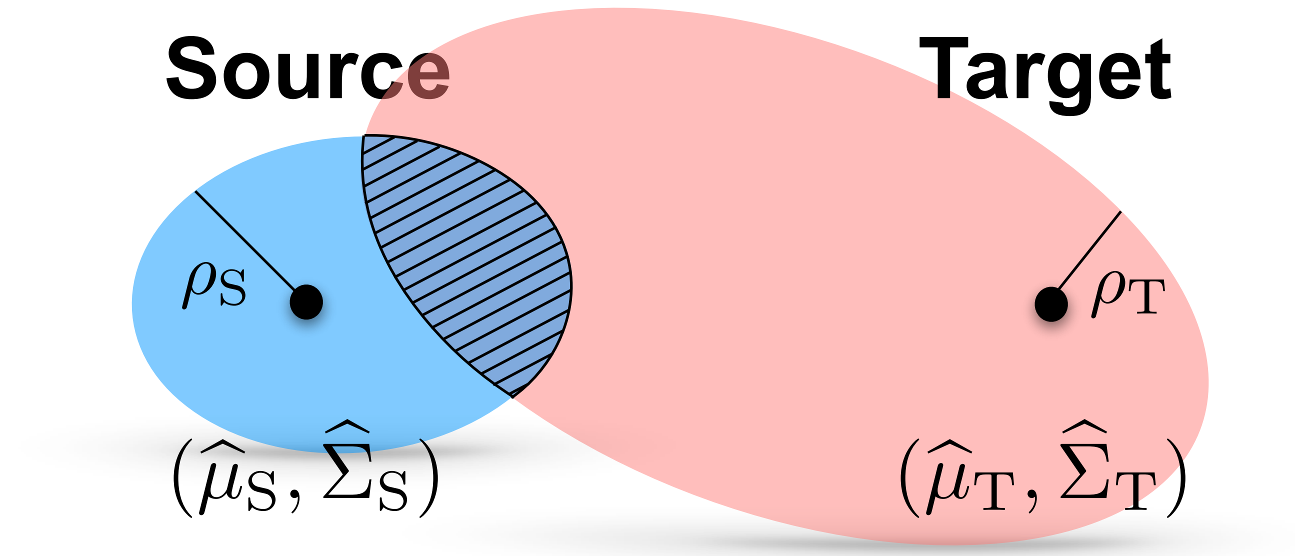
2) “Surround, then Intersect” (SI) Strategy



Moment information set:

$$\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} \triangleq \left\{ (\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p \text{ such that: } \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) \le \rho_{\mathrm{S}}, \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})) \le \rho_{\mathrm{T}}, \Sigma + \mu\mu^\top \succeq \varepsilon I_p \right\}$$

2) "Surround, then Intersect" (SI) Strategy



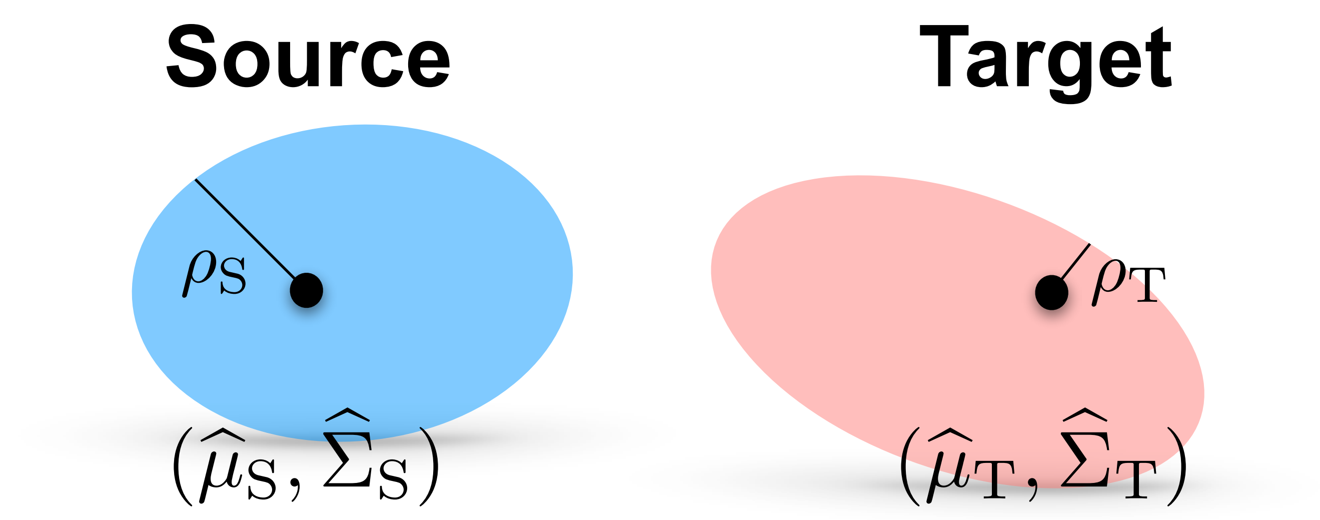
Moment information set:

$$\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} \triangleq \left\{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p \text{ such that: } \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) \leq \rho_{\mathrm{S}}, \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})) \leq \rho_{\mathrm{T}}, \Sigma + \mu\mu^\top \succeq \varepsilon I_p\right\}$$

Distribution set:

$$\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu, \Sigma), (\mu, \Sigma) \in \mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}\}$$

2) "Surround, then Intersect" (SI) Strategy



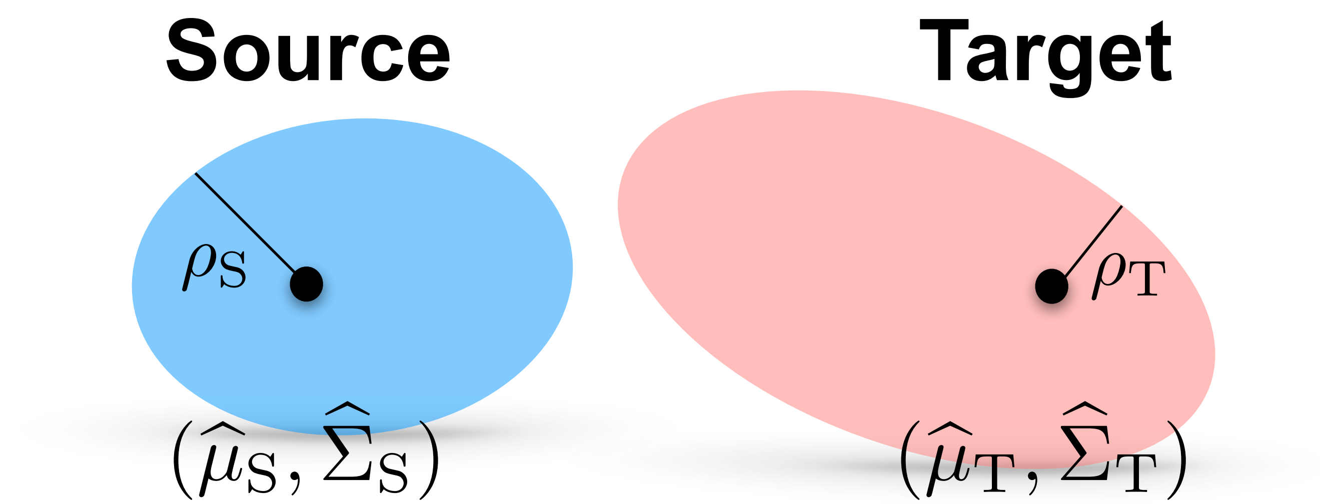
Moment information set:

$$\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} \triangleq \left\{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p \text{ such that: } \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) \le \rho_{\mathrm{S}}, \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})) \le \rho_{\mathrm{T}}, \Sigma + \mu\mu^\top \succeq \varepsilon I_p \right\}$$

Distribution set:

$$\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu, \Sigma), (\mu, \Sigma) \in \mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}\}$$

2) "Surround, then Intersect" (SI) Strategy



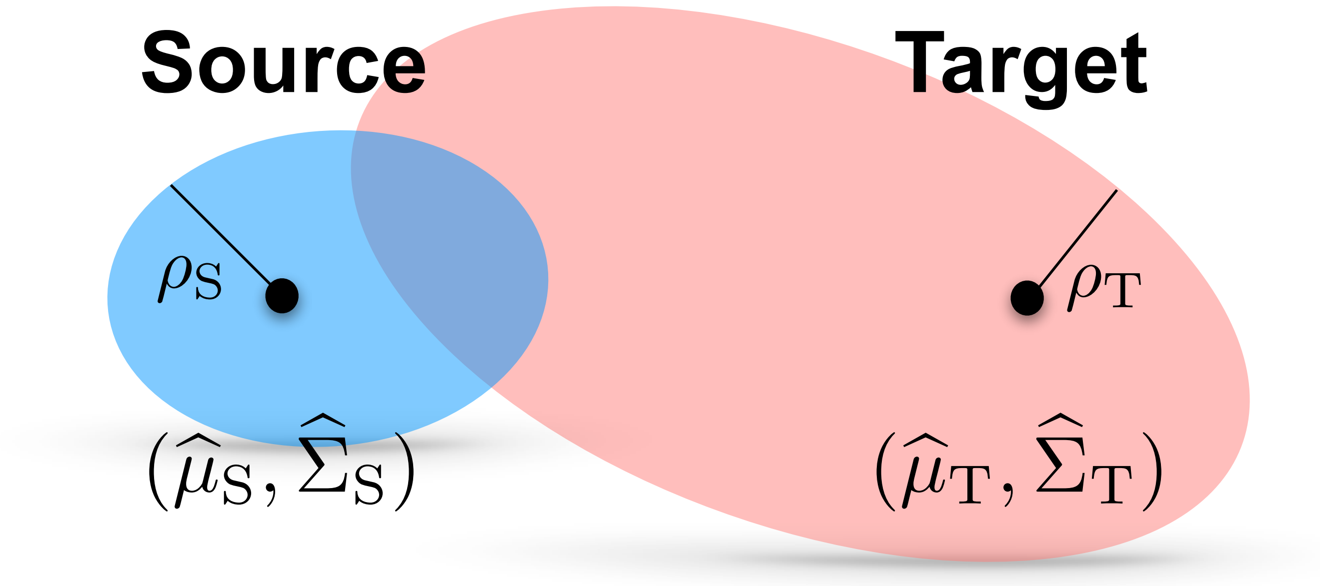
Moment information set:

$$\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} \triangleq \left\{(\mu,\Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p \text{ such that: } \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{S}},\widehat{\Sigma}_{\mathrm{S}})) \le \rho_{\mathrm{S}}, \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{T}},\widehat{\Sigma}_{\mathrm{T}})) \le \rho_{\mathrm{T}}, \Sigma + \mu\mu^\top \succeq \varepsilon I_p \right\}$$

Distribution set:

$$\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu,\Sigma), (\mu,\Sigma) \in \mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}\}$$

2) "Surround, then Intersect" (SI) Strategy



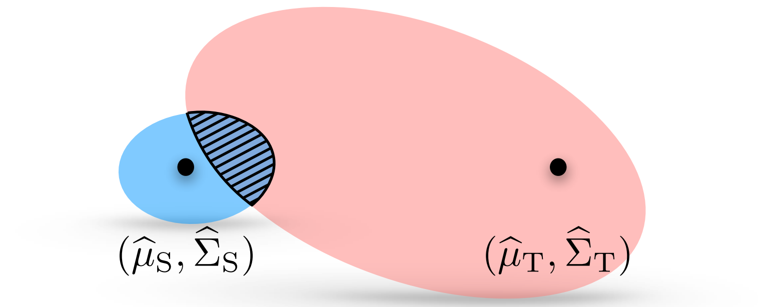
Moment information set:

$$\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} \triangleq \left\{ (\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p \text{ such that: } \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) \le \rho_{\mathrm{S}}, \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})) \le \rho_{\mathrm{T}}, \Sigma + \mu\mu^\top \succeq \varepsilon I_p \right\}$$

Distribution set:

$$\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} = \{ \mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu, \Sigma), (\mu, \Sigma) \in \mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} \}$$

2) "Surround, then Intersect" (SI) Strategy



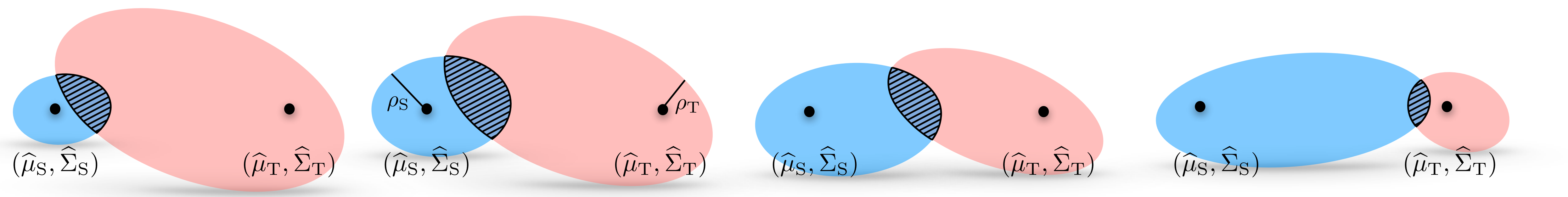
Moment information set:

$$\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} \triangleq \left\{(\mu,\Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p \text{ such that: } \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{S}},\widehat{\Sigma}_{\mathrm{S}})) \le \rho_{\mathrm{S}}, \psi((\mu,\Sigma)\|(\widehat{\mu}_{\mathrm{T}},\widehat{\Sigma}_{\mathrm{T}})) \le \rho_{\mathrm{T}}, \Sigma + \mu\mu^\top \succeq \varepsilon I_p \right\}$$

Distribution set:

$$\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu,\Sigma), (\mu,\Sigma) \in \mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}\}$$

2) "Surround, then Intersect" (SI) Strategy



Moment information set:

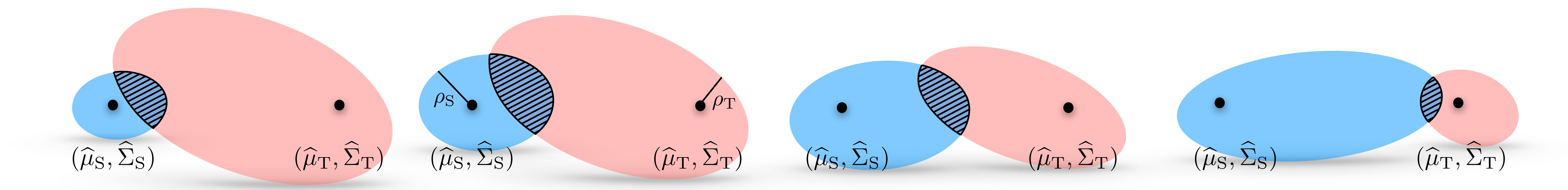
$$\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} \triangleq \left\{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p \text{ such that: } \psi((\mu, \Sigma)\|(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) \leq \rho_{\mathrm{S}}, \psi((\mu, \Sigma)\|(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})) \leq \rho_{\mathrm{T}}, \Sigma + \mu\mu^\top \succeq \varepsilon I_p \right\}$$

Distribution set:

$$\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu, \Sigma), (\mu, \Sigma) \in \mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}\}$$

,

2) "Surround, then Intersect" (SI) Strategy



i) Kullback-Leibler (KL)-type Divergence

Moment information set:

$$\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} \triangleq \left\{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p \text{ such that: } \mathbb{D}((\mu, \Sigma) \| (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) \le \rho_{\mathrm{S}}, \mathbb{D}((\mu, \Sigma) \| (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})) \le \rho_{\mathrm{T}}, \Sigma + \mu\mu^\top \succeq \varepsilon I_p\right\}$$

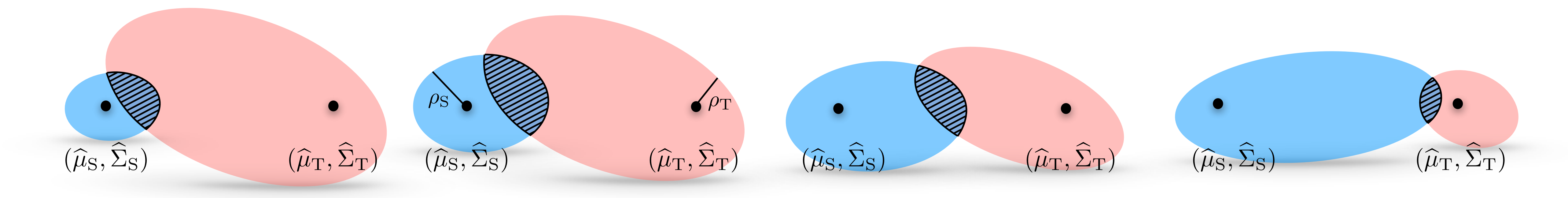
Distribution set:

$$\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu, \Sigma), (\mu, \Sigma) \in \mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}\}$$

Robust Estimation: $\beta^\star = (M_{XX}^\star)^{-1} M_{XY}^\star$ solves $\inf_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]$,

where $(M_{XX}^\star, M_{XY}^\star)$ is a solution of a convex semidefinite program.

2) "Surround, then Intersect" (SI) Strategy



ii) Wasserstein-type Divergence

Moment information set:

$$\mathbb{U}_{\rho_{\rm S},\rho_{\rm T}} \triangleq \left\{(\mu,\Sigma) \in \mathbb{R}^p \times \mathbb{S}^p_+ \text{ such that: } \mathbb{W}((\mu,\Sigma)\|(\widehat{\mu}_{\rm S},\widehat{\Sigma}_{\rm S})) \le \rho_{\rm S}, \mathbb{W}((\mu,\Sigma)\|(\widehat{\mu}_{\rm T},\widehat{\Sigma}_{\rm T})) \le \rho_{\rm T}, \Sigma + \mu\mu^\top \succeq \varepsilon I_p\right\}$$

Distribution set:

$$\mathbb{B}_{\rho_{\rm S},\rho_{\rm T}} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu,\Sigma), (\mu,\Sigma) \in \mathbb{U}_{\rho_{\rm S},\rho_{\rm T}}\}$$

Robust Estimation: $\beta^\star = (M^\star_{XX})^{-1} M^\star_{XY}$ solves $\inf_{\beta\in\mathbb{R}^d} \sup_{\mathbb{Q}\in\mathbb{B}_{\rho_{\rm S},\rho_{\rm T}}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]$,

where $(M^\star_{XX}, M^\star_{XY})$ is a solution of a linear semidefinite program.

Numerical Experiments

Data Set	Time	IR-KL	IR-WASS	SI-KL	SI-WASS	CC-L	CC-TL	CC-SL	CC-TE	CC-SE	RWS	LSE-T	LSE-T&S
Uber&Lyft	5	17.65	**1.00**	199.28	1.01	34.04	98.43	12.03	155.71	1.74	1.45	119.65	11.08
	10	13.67	**1.00**	111.52	1.01	30.85	99.22	11.40	161.72	1.58	1.34	137.15	6.32
	50	13.39	**1.00**	60.29	1.01	25.87	85.06	9.72	147.45	1.42	1.16	57.85	2.12
	100	15.24	**1.00**	59.06	1.01	26.01	85.77	9.91	148.49	1.41	1.12	31.25	1.57
US Births (2018)	5	79.83	1.02	44.71	**1.00**	64.99	257.60	25.13	432.09	2.07	4.50	727.88	39.17
	10	115.47	1.02	39.35	**1.00**	45.59	195.14	18.33	339.11	1.60	3.29	524.39	19.28
	50	107.40	1.01	40.04	**1.00**	42.74	192.46	13.12	361.51	1.31	2.00	191.27	5.20
	100	117.03	1.01	53.13	**1.00**	45.35	208.65	12.94	397.33	1.22	1.75	104.75	3.19
Life Expectancy	5	33.18	**1.00**	6.24	1.03	17.24	77.06	7.38	125.71	1.46	1.15	255.08	20.72
	10	25.59	**1.00**	5.45	1.02	12.49	60.19	5.50	104.00	1.40	1.15	167.15	10.73
	50	19.81	**1.00**	8.70	1.01	7.57	44.00	3.10	84.98	1.38	1.10	39.83	3.15
	100	19.02	**1.00**	8.25	1.005	6.82	41.40	2.68	83.60	1.38	1.08	20.42	2.10
House Prices in KC	5	1.58	**1.00**	1.21	1.01	3.98	8.87	2.12	13.31	1.29	1.23	11.75	3.70
	10	1.52	**1.00**	1.20	1.01	3.58	7.77	2.02	11.70	1.27	1.23	6.93	2.25
	50	1.34	**1.00**	1.31	1.01	2.79	6.52	1.86	10.37	1.27	1.20	3.91	1.30
	100	1.34	**1.00**	1.30	1.01	2.65	6.54	1.91	10.74	1.27	1.18	2.72	1.12
California Housing	5	63.33	1.05	3.31	**1.00**	27.63	102.82	9.60	181.52	1.35	1.17	96.43	54.34
	10	68.08	1.04	2.42	**1.00**	20.57	91.86	6.23	169.87	1.19	1.17	45.64	24.76
	50	70.08	1.01	1.97	**1.00**	11.79	81.72	2.49	170.18	1.05	1.13	10.17	5.63
	100	72.80	1.003	1.90	**1.00**	9.71	79.19	1.83	173.96	1.04	1.14	5.81	3.39

Numerical Experiments

Data Set	Time	IR-KL	IR-WASS	SI-KL	SI-WASS	CC-L	CC-TL	CC-SL	CC-TE	CC-SE	RWS	LSE-T	LSE-T&S
Uber&Lyft	5	17.65	**1.00**	199.28	1.01	34.04	98.43	12.03	155.71	1.74	1.45	119.65	11.08
	10	13.67	**1.00**	111.52	1.01	30.85	99.22	11.40	161.72	1.58	1.34	137.15	6.32
	50	13.39	**1.00**	60.29	1.01	25.87	85.06	9.72	147.45	1.42	1.16	57.85	2.12
	100	15.24	**1.00**	59.06	1.01	26.01	85.77	9.91	148.49	1.41	1.12	31.25	1.57
US Births (2018)	5	79.83	1.02	44.71	**1.00**	64.99	257.60	25.13	432.09	2.07	4.50	727.88	39.17
	10	115.47	1.02	39.35	**1.00**	45.59	195.14	18.33	339.11	1.60	3.29	524.39	19.28
	50	107.40	1.01	40.04	**1.00**	42.74	192.46	13.12	361.51	1.31	2.00	191.27	5.20
	100	117.03	1.01	53.13	**1.00**	45.35	208.65	12.94	397.33	1.22	1.75	104.75	3.19
Life Expectancy	5	33.18	**1.00**	6.24	1.03	17.24	77.06	7.38	125.71	1.46	1.15	255.08	20.72
	10	25.59	**1.00**	5.45	1.02	12.49	60.19	5.50	104.00	1.40	1.15	167.15	10.73
	50	19.81	**1.00**	8.70	1.01	7.57	44.00	3.10	84.98	1.38	1.10	39.83	3.15
	100	19.02	**1.00**	8.25	1.005	6.82	41.40	2.68	83.60	1.38	1.08	20.42	2.10
House Prices in KC	5	1.58	**1.00**	1.21	1.01	3.98	8.87	2.12	13.31	1.29	1.23	11.75	3.70
	10	1.52	**1.00**	1.20	1.01	3.58	7.77	2.02	11.70	1.27	1.23	6.93	2.25
	50	1.34	**1.00**	1.31	1.01	2.79	6.52	1.86	10.37	1.27	1.20	3.91	1.30
	100	1.34	**1.00**	1.30	1.01	2.65	6.54	1.91	10.74	1.27	1.18	2.72	1.12
California Housing	5	63.33	1.05	3.31	**1.00**	27.63	102.82	9.60	181.52	1.35	1.17	96.43	54.34
	10	68.08	1.04	2.42	**1.00**	20.57	91.86	6.23	169.87	1.19	1.17	45.64	24.76
	50	70.08	1.01	1.97	**1.00**	11.79	81.72	2.49	170.18	1.05	1.13	10.17	5.63
	100	72.80	1.003	1.90	**1.00**	9.71	79.19	1.83	173.96	1.04	1.14	5.81	3.39

Numerical Experiments

Data Set	Time	IR-KL	IR-WASS	SI-KL	SI-WASS	CC-L	CC-TL	CC-SL	CC-TE	CC-SE	RWS	LSE-T	LSE-T&S
Uber&Lyft	5	17.65	**1.00**	199.28	1.01	34.04	98.43	12.03	155.71	1.74	1.45	119.65	11.08
	10	13.67	**1.00**	111.52	1.01	30.85	99.22	11.40	161.72	1.58	1.34	137.15	6.32
	50	13.39	**1.00**	60.29	1.01	25.87	85.06	9.72	147.45	1.42	1.16	57.85	2.12
	100	15.24	**1.00**	59.06	1.01	26.01	85.77	9.91	148.49	1.41	1.12	31.25	1.57
US Births (2018)	5	79.83	1.02	44.71	**1.00**	64.99	257.60	25.13	432.09	2.07	4.50	727.88	39.17
	10	115.47	1.02	39.35	**1.00**	45.59	195.14	18.33	339.11	1.60	3.29	524.39	19.28
	50	107.40	1.01	40.04	**1.00**	42.74	192.46	13.12	361.51	1.31	2.00	191.27	5.20
	100	117.03	1.01	53.13	**1.00**	45.35	208.65	12.94	397.33	1.22	1.75	104.75	3.19
Life Expectancy	5	33.18	**1.00**	6.24	1.03	17.24	77.06	7.38	125.71	1.46	1.15	255.08	20.72
	10	25.59	**1.00**	5.45	1.02	12.49	60.19	5.50	104.00	1.40	1.15	167.15	10.73
	50	19.81	**1.00**	8.70	1.01	7.57	44.00	3.10	84.98	1.38	1.10	39.83	3.15
	100	19.02	**1.00**	8.25	1.005	6.82	41.40	2.68	83.60	1.38	1.08	20.42	2.10
House Prices in KC	5	1.58	**1.00**	1.21	1.01	3.98	8.87	2.12	13.31	1.29	1.23	11.75	3.70
	10	1.52	**1.00**	1.20	1.01	3.58	7.77	2.02	11.70	1.27	1.23	6.93	2.25
	50	1.34	**1.00**	1.31	1.01	2.79	6.52	1.86	10.37	1.27	1.20	3.91	1.30
	100	1.34	**1.00**	1.30	1.01	2.65	6.54	1.91	10.74	1.27	1.18	2.72	1.12
California Housing	5	63.33	1.05	3.31	**1.00**	27.63	102.82	9.60	181.52	1.35	1.17	96.43	54.34
	10	68.08	1.04	2.42	**1.00**	20.57	91.86	6.23	169.87	1.19	1.17	45.64	24.76
	50	70.08	1.01	1.97	**1.00**	11.79	81.72	2.49	170.18	1.05	1.13	10.17	5.63
	100	72.80	1.003	1.90	**1.00**	9.71	79.19	1.83	173.96	1.04	1.14	5.81	3.39

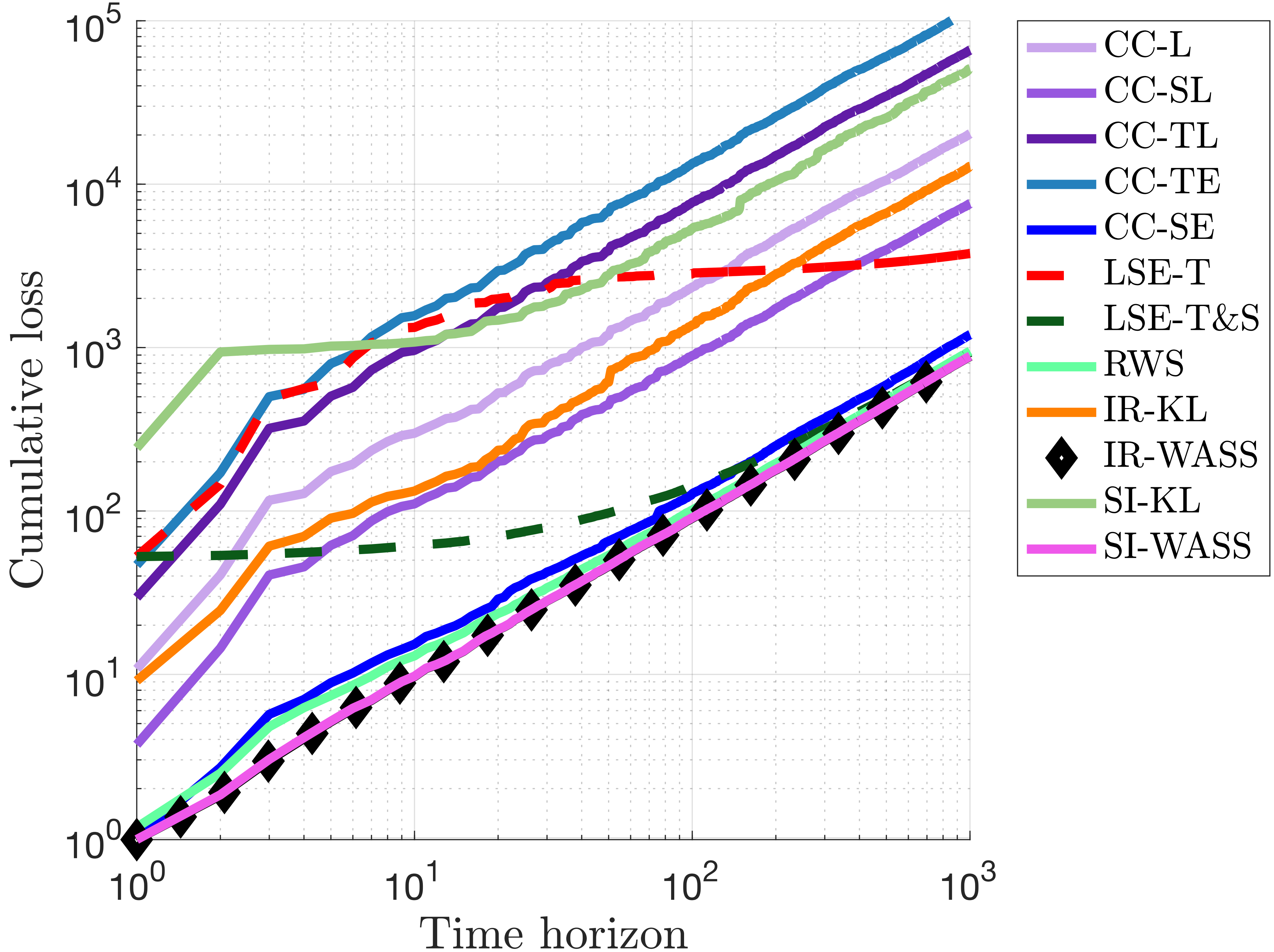
Numerical Experiments

Data Set	Time	IR-KL	IR-WASS	SI-KL	SI-WASS	CC-L	CC-TL	CC-SL	CC-TE	CC-SE	RWS	LSE-T	LSE-T&S
Uber&Lyft	5	17.65	**1.00**	199.28	1.01	34.04	98.43	12.03	155.71	1.74	1.45	119.65	11.08
	10	13.67	**1.00**	111.52	1.01	30.85	99.22	11.40	161.72	1.58	1.34	137.15	6.32
	50	13.39	**1.00**	60.29	1.01	25.87	85.06	9.72	147.45	1.42	1.16	57.85	2.12
	100	15.24	**1.00**	59.06	1.01	26.01	85.77	9.91	148.49	1.41	1.12	31.25	1.57
US Births (2018)	5	79.83	1.02	44.71	**1.00**	64.99	257.60	25.13	432.09	2.07	4.50	727.88	39.17
	10	115.47	1.02	39.35	**1.00**	45.59	195.14	18.33	339.11	1.60	3.29	524.39	19.28
	50	107.40	1.01	40.04	**1.00**	42.74	192.46	13.12	361.51	1.31	2.00	191.27	5.20
	100	117.03	1.01	53.13	**1.00**	45.35	208.65	12.94	397.33	1.22	1.75	104.75	3.19
Life Expectancy	5	33.18	**1.00**	6.24	1.03	17.24	77.06	7.38	125.71	1.46	1.15	255.08	20.72
	10	25.59	**1.00**	5.45	1.02	12.49	60.19	5.50	104.00	1.40	1.15	167.15	10.73
	50	19.81	**1.00**	8.70	1.01	7.57	44.00	3.10	84.98	1.38	1.10	39.83	3.15
	100	19.02	**1.00**	8.25	1.005	6.82	41.40	2.68	83.60	1.38	1.08	20.42	2.10
House Prices in KC	5	1.58	**1.00**	1.21	1.01	3.98	8.87	2.12	13.31	1.29	1.23	11.75	3.70
	10	1.52	**1.00**	1.20	1.01	3.58	7.77	2.02	11.70	1.27	1.23	6.93	2.25
	50	1.34	**1.00**	1.31	1.01	2.79	6.52	1.86	10.37	1.27	1.20	3.91	1.30
	100	1.34	**1.00**	1.30	1.01	2.65	6.54	1.91	10.74	1.27	1.18	2.72	1.12
California Housing	5	63.33	1.05	3.31	**1.00**	27.63	102.82	9.60	181.52	1.35	1.17	96.43	54.34
	10	68.08	1.04	2.42	**1.00**	20.57	91.86	6.23	169.87	1.19	1.17	45.64	24.76
	50	70.08	1.01	1.97	**1.00**	11.79	81.72	2.49	170.18	1.05	1.13	10.17	5.63
	100	72.80	1.003	1.90	**1.00**	9.71	79.19	1.83	173.96	1.04	1.14	5.81	3.39

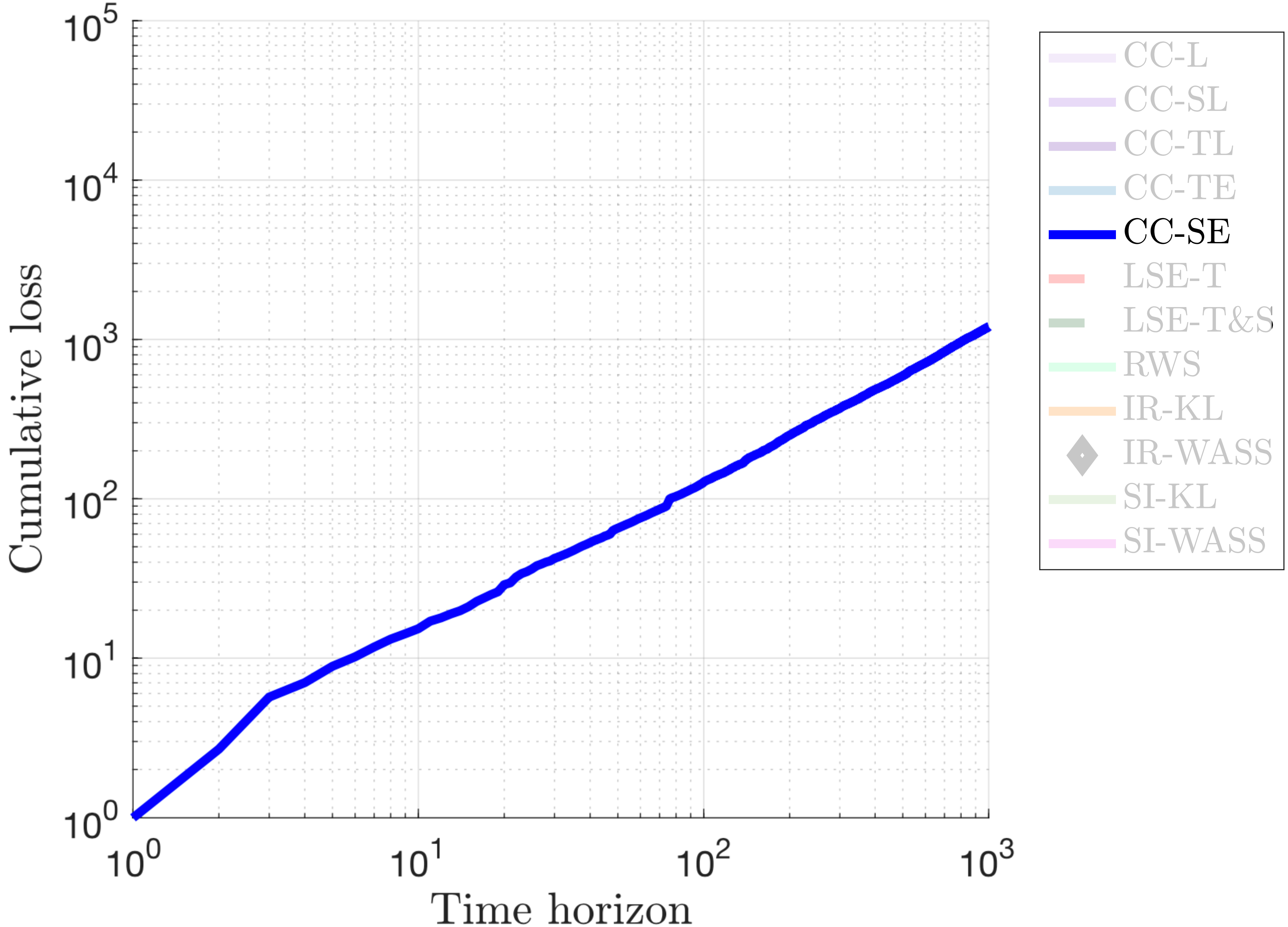
Numerical Experiments

Data Set	Time	IR-KL	IR-WASS	SI-KL	SI-WASS	CC-L	CC-TL	CC-SL	CC-TE	CC-SE	RWS	LSE-T	LSE-T&S
Uber&Lyft	5	17.65	**1.00**	199.28	1.01	34.04	98.43	12.03	155.71	1.74	1.45	119.65	11.08
	10	13.67	**1.00**	111.52	1.01	30.85	99.22	11.40	161.72	1.58	1.34	137.15	6.32
	50	13.39	**1.00**	60.29	1.01	25.87	85.06	9.72	147.45	1.42	1.16	57.85	2.12
	100	15.24	**1.00**	59.06	1.01	26.01	85.77	9.91	148.49	1.41	1.12	31.25	1.57
US Births (2018)	5	79.83	1.02	44.71	**1.00**	64.99	257.60	25.13	432.09	2.07	4.50	727.88	39.17
	10	115.47	1.02	39.35	**1.00**	45.59	195.14	18.33	339.11	1.60	3.29	524.39	19.28
	50	107.40	1.01	40.04	**1.00**	42.74	192.46	13.12	361.51	1.31	2.00	191.27	5.20
	100	117.03	1.01	53.13	**1.00**	45.35	208.65	12.94	397.33	1.22	1.75	104.75	3.19
Life Expectancy	5	33.18	**1.00**	6.24	1.03	17.24	77.06	7.38	125.71	1.46	1.15	255.08	20.72
	10	25.59	**1.00**	5.45	1.02	12.49	60.19	5.50	104.00	1.40	1.15	167.15	10.73
	50	19.81	**1.00**	8.70	1.01	7.57	44.00	3.10	84.98	1.38	1.10	39.83	3.15
	100	19.02	**1.00**	8.25	1.005	6.82	41.40	2.68	83.60	1.38	1.08	20.42	2.10
House Prices in KC	5	1.58	**1.00**	1.21	1.01	3.98	8.87	2.12	13.31	1.29	1.23	11.75	3.70
	10	1.52	**1.00**	1.20	1.01	3.58	7.77	2.02	11.70	1.27	1.23	6.93	2.25
	50	1.34	**1.00**	1.31	1.01	2.79	6.52	1.86	10.37	1.27	1.20	3.91	1.30
	100	1.34	**1.00**	1.30	1.01	2.65	6.54	1.91	10.74	1.27	1.18	2.72	1.12
California Housing	5	63.33	1.05	3.31	**1.00**	27.63	102.82	9.60	181.52	1.35	1.17	96.43	54.34
	10	68.08	1.04	2.42	**1.00**	20.57	91.86	6.23	169.87	1.19	1.17	45.64	24.76
	50	70.08	1.01	1.97	**1.00**	11.79	81.72	2.49	170.18	1.05	1.13	10.17	5.63
	100	72.80	1.003	1.90	**1.00**	9.71	79.19	1.83	173.96	1.04	1.14	5.81	3.39

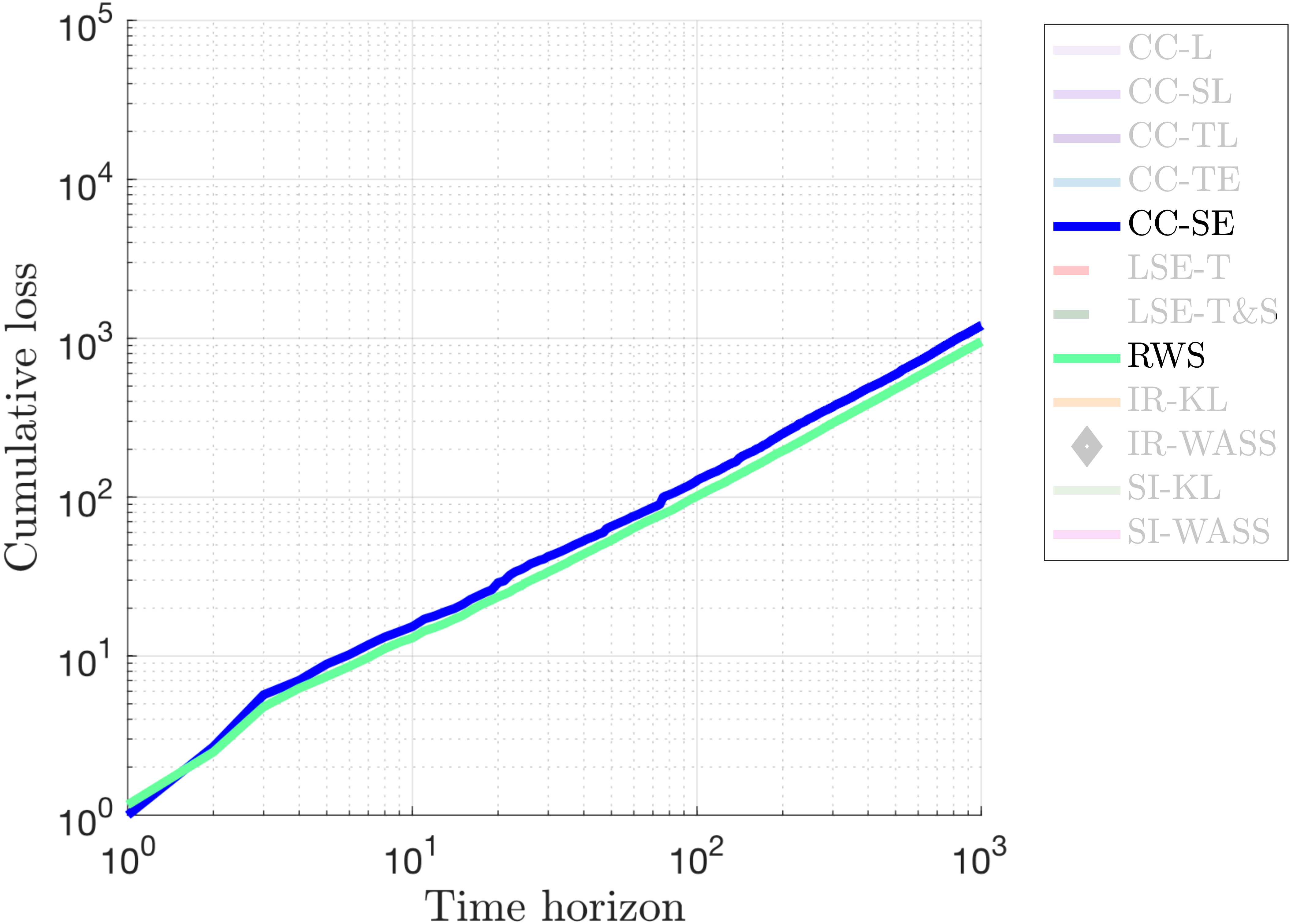
Numerical Experiments - Uber&Lyft



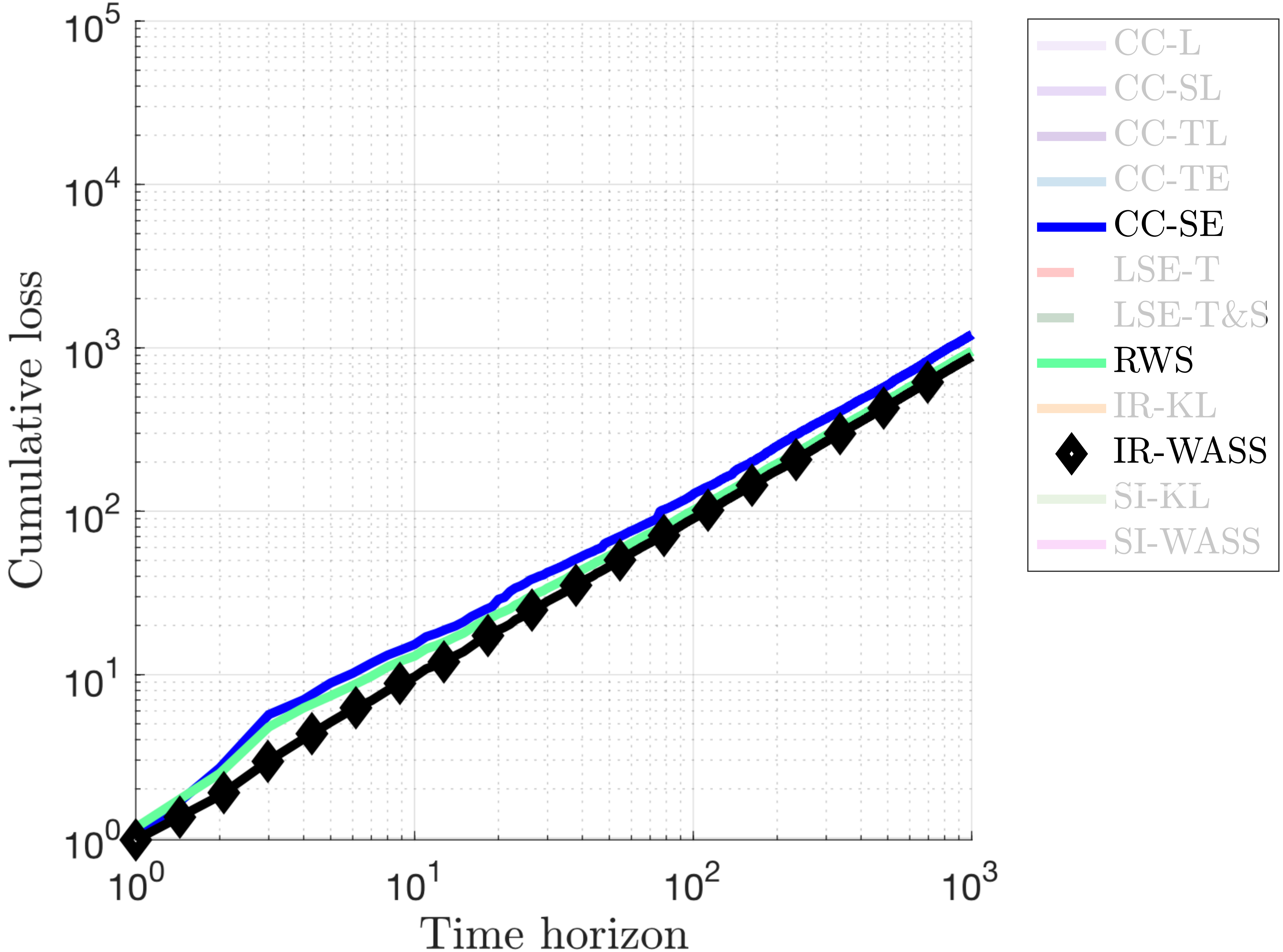
Numerical Experiments



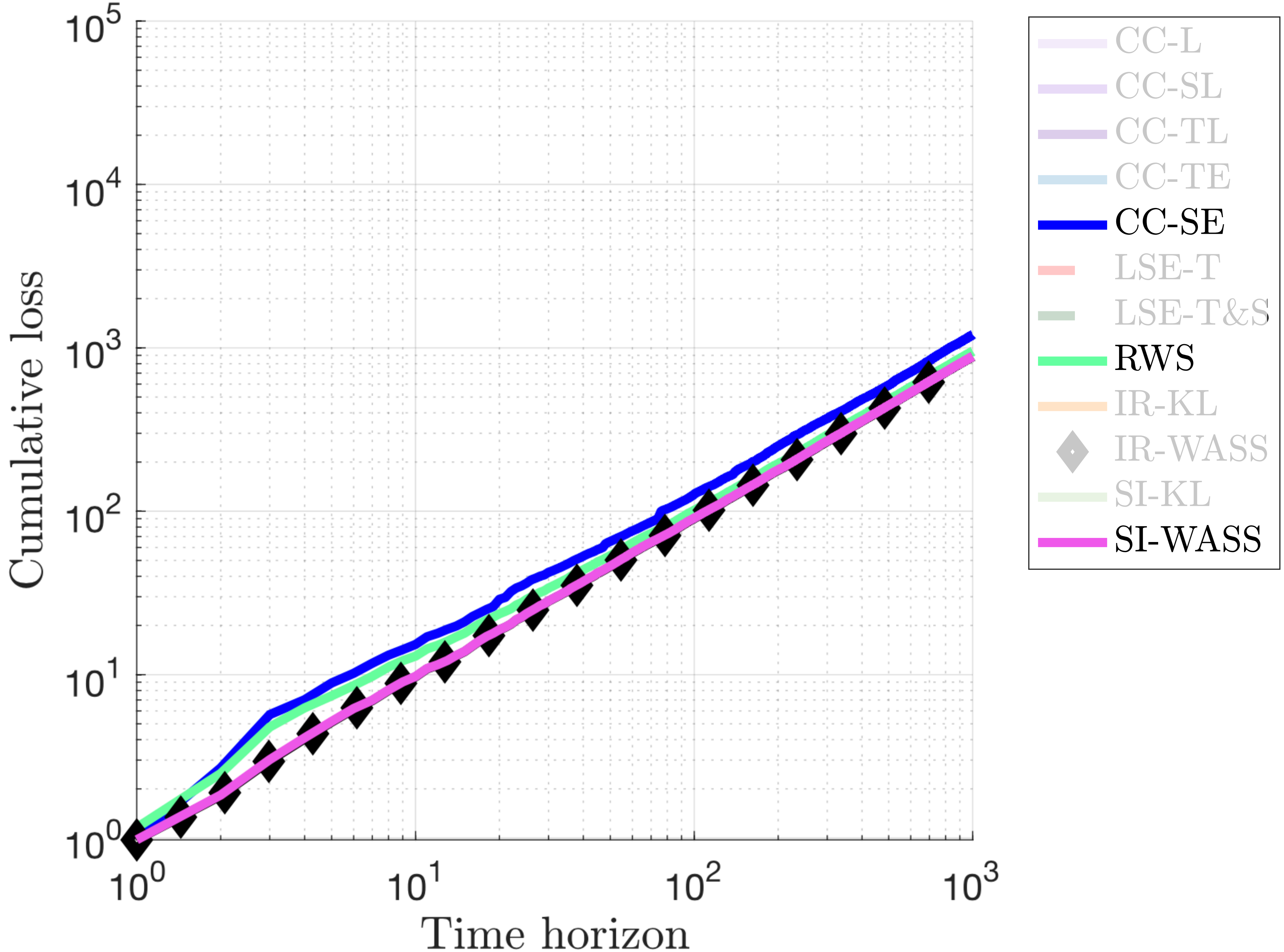
Numerical Experiments



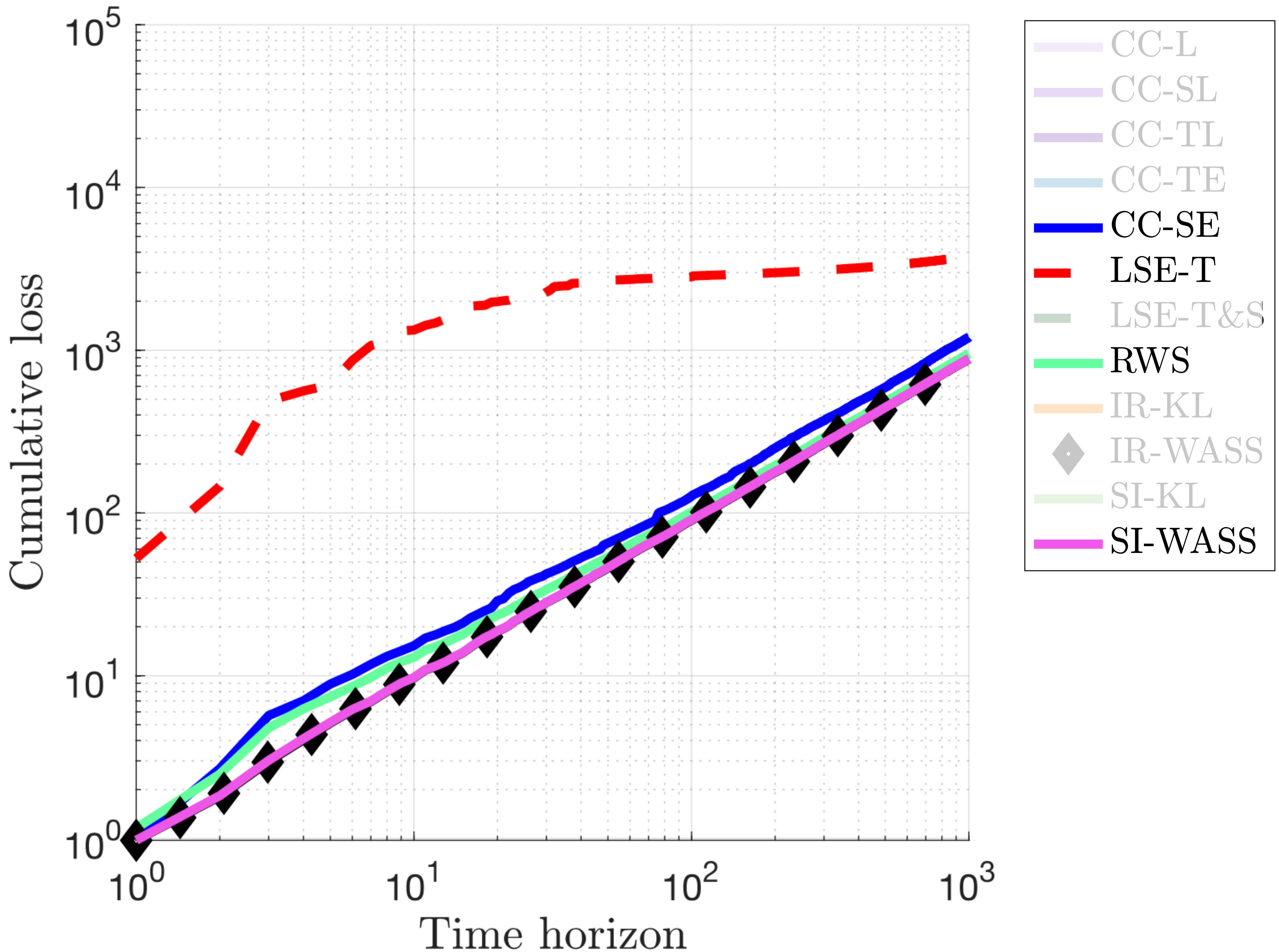
Numerical Experiments



Numerical Experiments



Numerical Experiments



Concluding Remarks

- KL-type divergence based methods are **less** numerically stable
- Asymmetry of the KL-type divergence
- Extrapolating schemes
- Favourable properties of different divergences
- Theoretical guarantees for the out-of-sample performance

References

[1] Taskesen B., Yue M.C., Blanchet J., Kuhn D, Nguyen V.A.. **Sequential Domain Adaptation by Synthesizing Distributionally Robust Experts**, *ICML* 2021

[2] Garcke, J. and Vanck, T. Importance weighted inductive transfer learning for regression. *In Joint European conference on machine learning and knowledge discovery in databases*, pp. 466–481, 2014

[3] Wintenberger, O. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106(1):119–141, 2017.

Thanks a lot for your attention!