uai2025 Rio de Janeiro, Brazil I 21-25 July, 2025

Distributionally & Adversarially Robust Logistic Regression

Intersecting Wasserstein Balls

Aras Selvi | Eleonora Kreacic | Mohsen Ghassemi | Vamsi Potluru | Tucker Balch | Manuela Veloso

IMPERIAL
BUSINESS SCHOOL

COG
computational optimisation group

JPMC

Sargent Centre
for Process Systems
Engineering

Aras Selvi | July 24th, 2025 | UAI 2025 @Rio

Empirical Risk Minimization

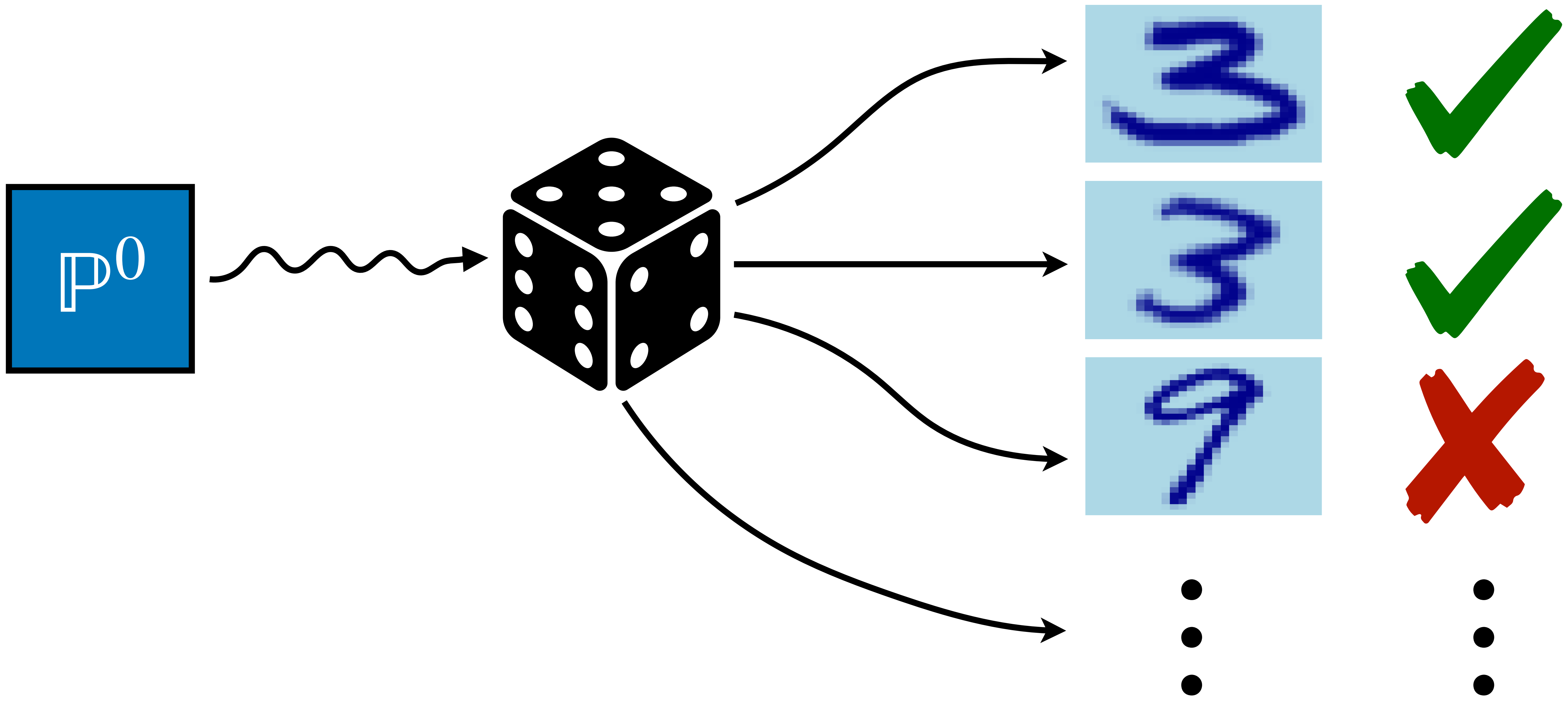
True Risk Minimization

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\text{minimize}} \; \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$$

Empirical Risk Minimization

True Risk Minimization

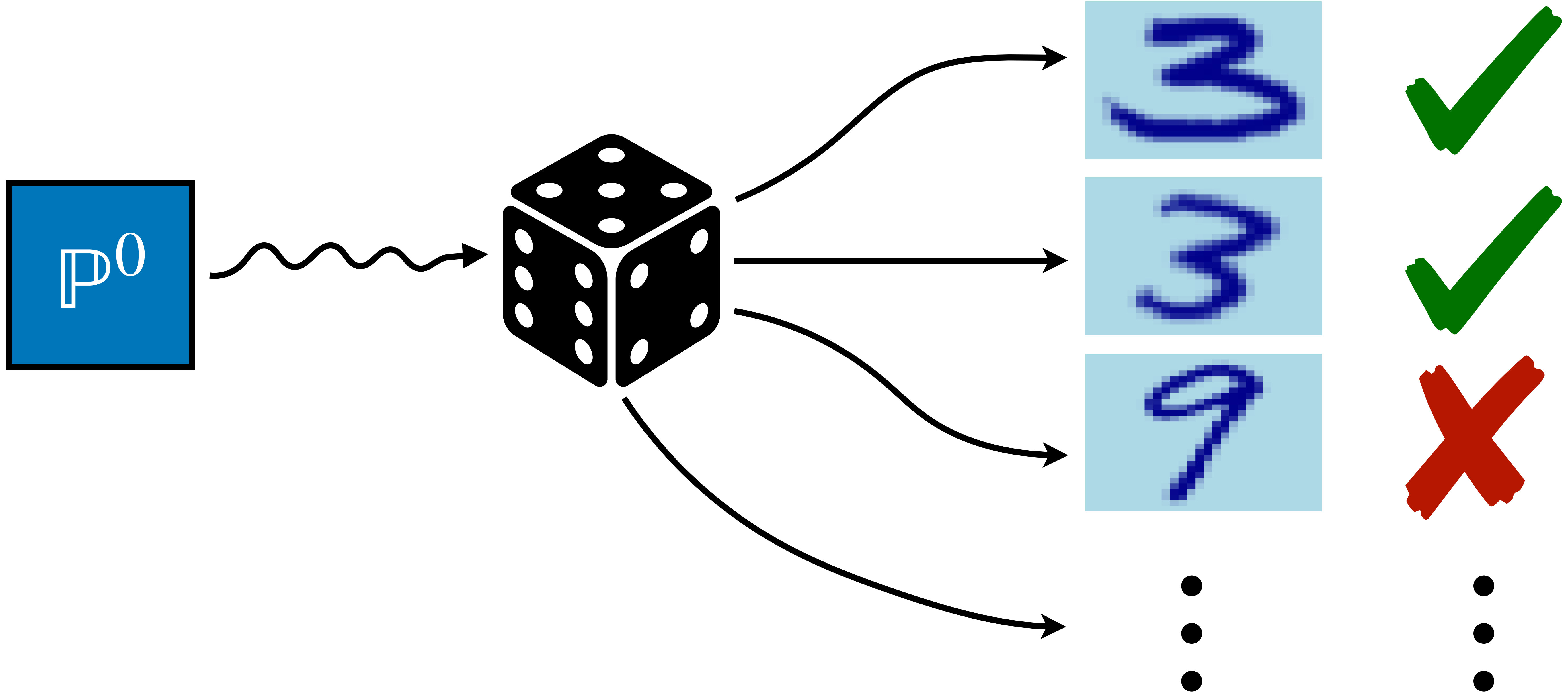
$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \; \mathbb{E}_{(x,y)\sim \mathbb{P}^0}[\ell_\beta(x, y)]$$



Empirical Risk Minimization

True Risk Minimization

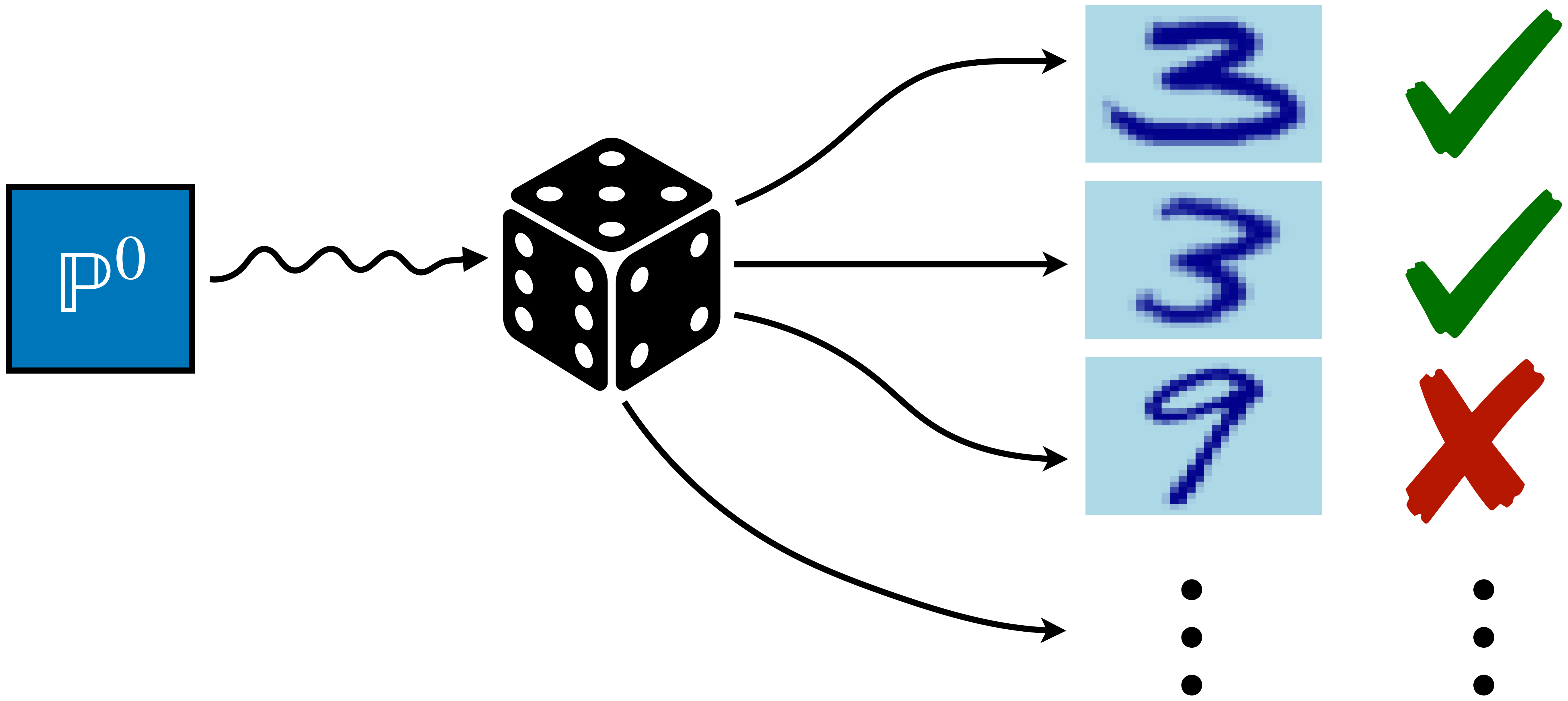
$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \ \mathbb{E}_{(x,y)\sim\mathbb{P}^0}[\ell_\beta(x, y)]$$



Empirical Risk Minimization

True Risk Minimization

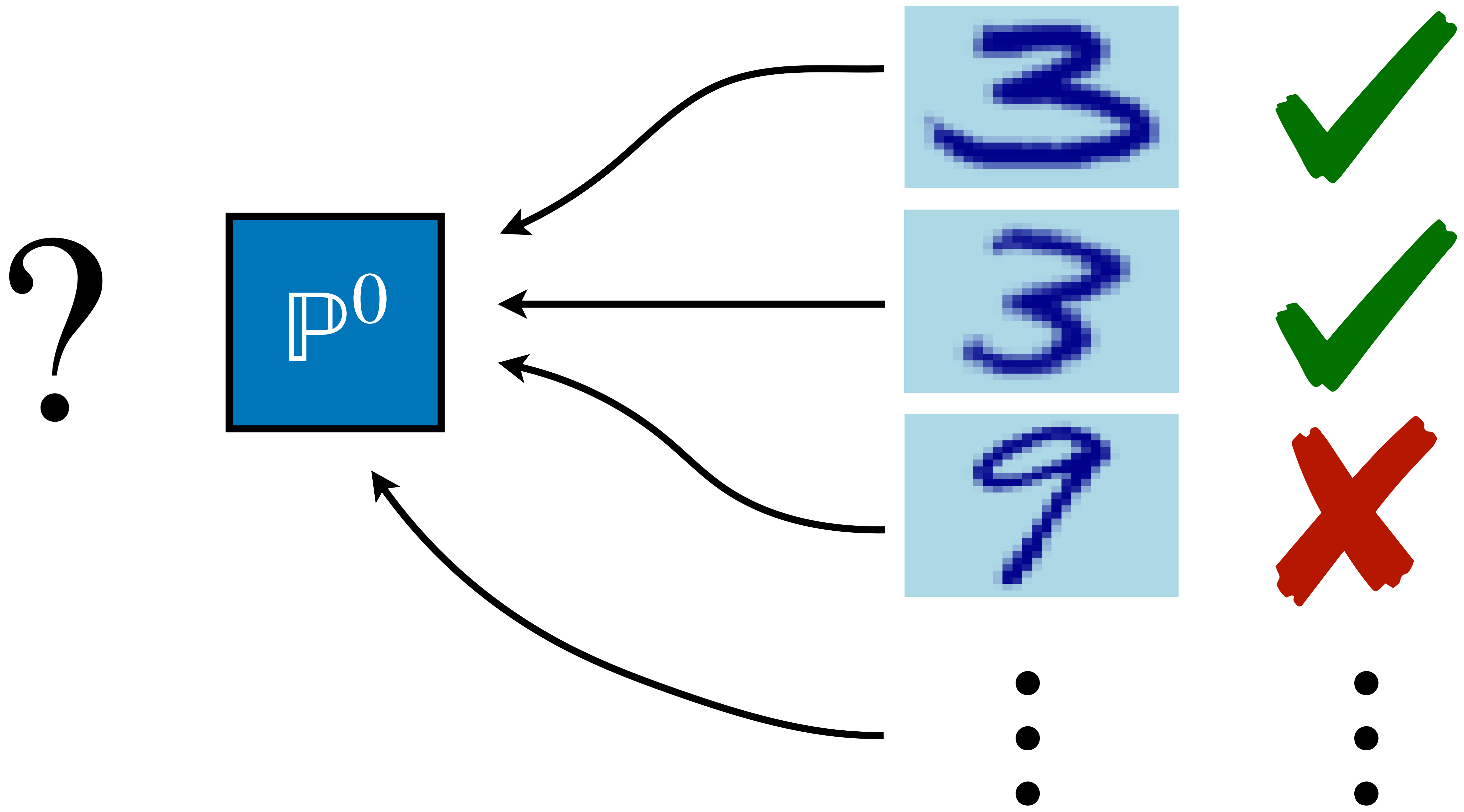
$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \ \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$$



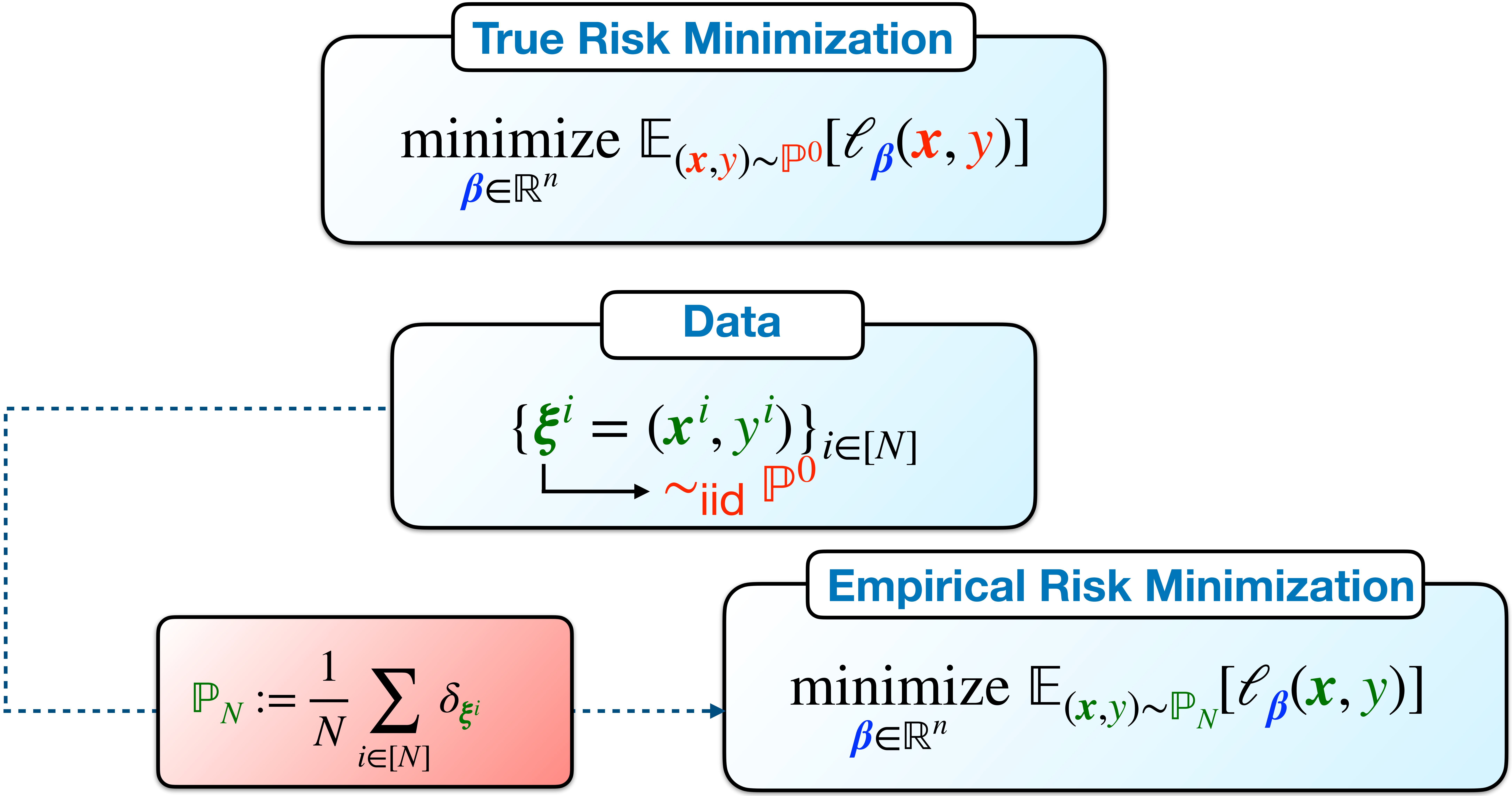
Empirical Risk Minimization

True Risk Minimization

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \; \mathbb{E}_{(x,y)\sim \mathbb{P}^0}[\ell_\beta(x,y)]$$



Empirical Risk Minimization



True Risk Minimization

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{minimize}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$$

Data

$$\{\boldsymbol{\xi}^i = (\boldsymbol{x}^i, y^i)\}_{i\in[N]}$$

$$\hookrightarrow \sim_{\text{iid}} \mathbb{P}^0$$

$$\mathbb{P}_N := \frac{1}{N}\sum_{i\in[N]} \delta_{\boldsymbol{\xi}^i}$$

Empirical Risk Minimization

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{minimize}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$$

Empirical Risk Minimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$

Empirical Risk Minimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$

Overfitting

We tend to underestimate the true risk with $\boldsymbol{\beta}^{\mathrm{ERM}}$

$$\mathbb{E}_{\mathbb{P}^0}[\ell_{\boldsymbol{\beta}^{\mathrm{ERM}}}(\boldsymbol{x},y)] - \mathbb{E}_{\mathbb{P}_N}[\ell_{\boldsymbol{\beta}^{\mathrm{ERM}}}(\boldsymbol{x},y)]$$

Empirical Risk Minimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$

Overfitting

We tend to underestimate the true risk with $\boldsymbol{\beta}^{\mathrm{ERM}}$

$$\mathbb{E}_{\mathbb{P}^0}[\ell_{\boldsymbol{\beta}^{\mathrm{ERM}}}(\boldsymbol{x},y)] - \mathbb{E}_{\mathbb{P}_N}[\ell_{\boldsymbol{\beta}^{\mathrm{ERM}}}(\boldsymbol{x},y)]$$

DRO philosophy: statistical error of estimating $\mathbb{P}^0$ via $\mathbb{P}_N$

Empirical Risk Minimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$

Overfitting

We tend to underestimate the true risk with $\boldsymbol{\beta}^{\mathrm{ERM}}$

$$\mathbb{E}_{\mathbb{P}^0}[\ell_{\boldsymbol{\beta}^{\mathrm{ERM}}}(\boldsymbol{x},y)] - \mathbb{E}_{\mathbb{P}_N}[\ell_{\boldsymbol{\beta}^{\mathrm{ERM}}}(\boldsymbol{x},y)]$$

DRO philosophy: statistical error of estimating $\mathbb{P}^0$ via $\mathbb{P}_N$

$$\mathrm{W}(\mathbb{P}_N, \mathbb{P}^0) > 0$$

The Wasserstein Distance

Metric on the feature-label space Ξ

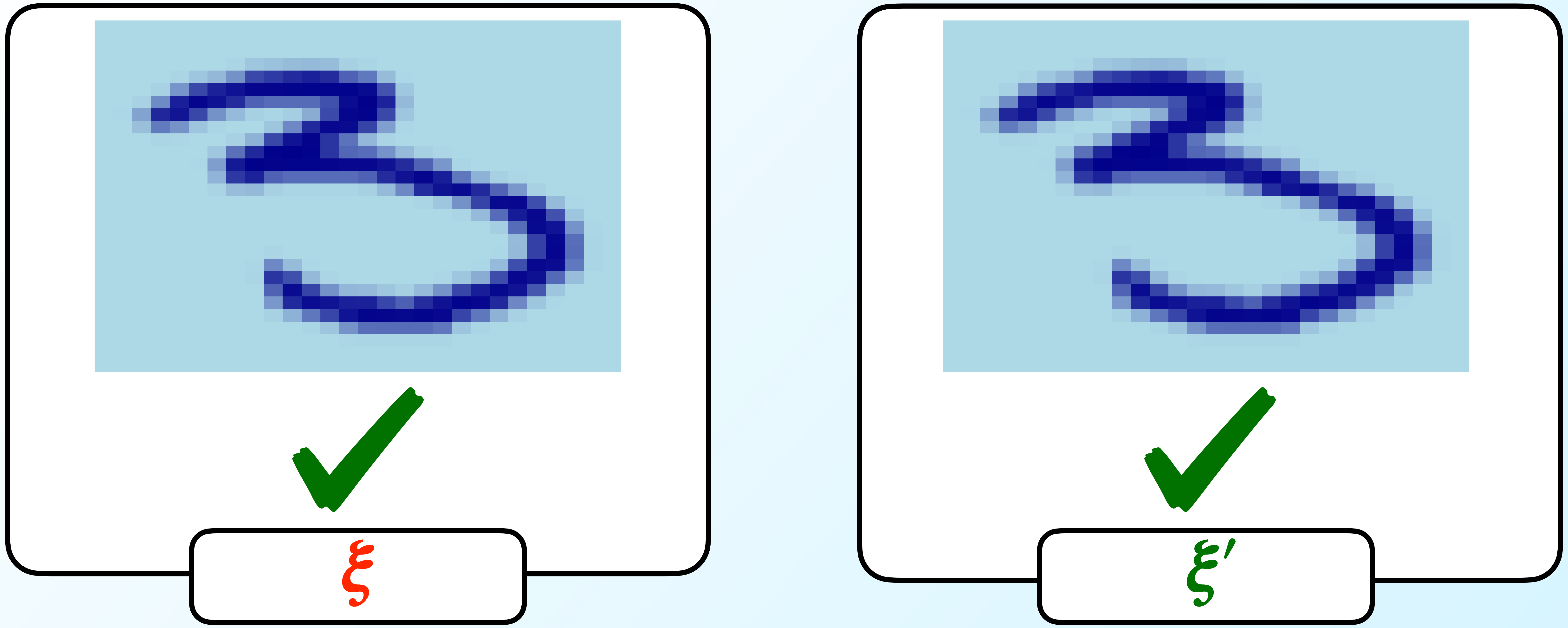
Distance between $\boldsymbol{\xi} = (\boldsymbol{x}, y) \in \Xi$ and $\boldsymbol{\xi}' = (\boldsymbol{x}', y') \in \Xi$ is

$$d(\boldsymbol{\xi}, \boldsymbol{\xi}') = \|\boldsymbol{x} - \boldsymbol{x}'\|_q + \kappa \cdot \mathbf{1}[y \neq y']$$

The Wasserstein Distance

Metric on the feature-label space Ξ

Example $\kappa = 28,\ q = 2$

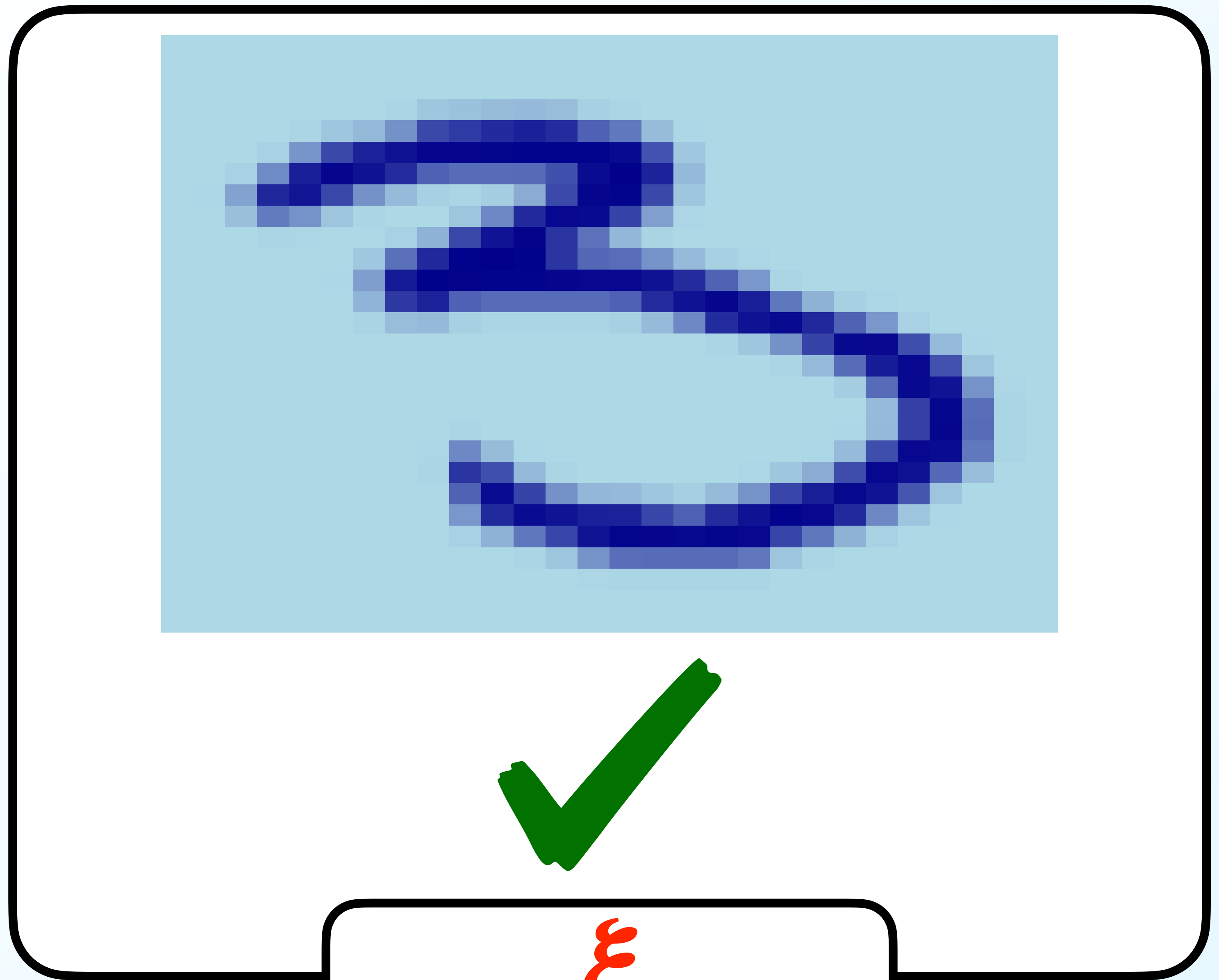


$$d(\xi, \xi') = 0$$

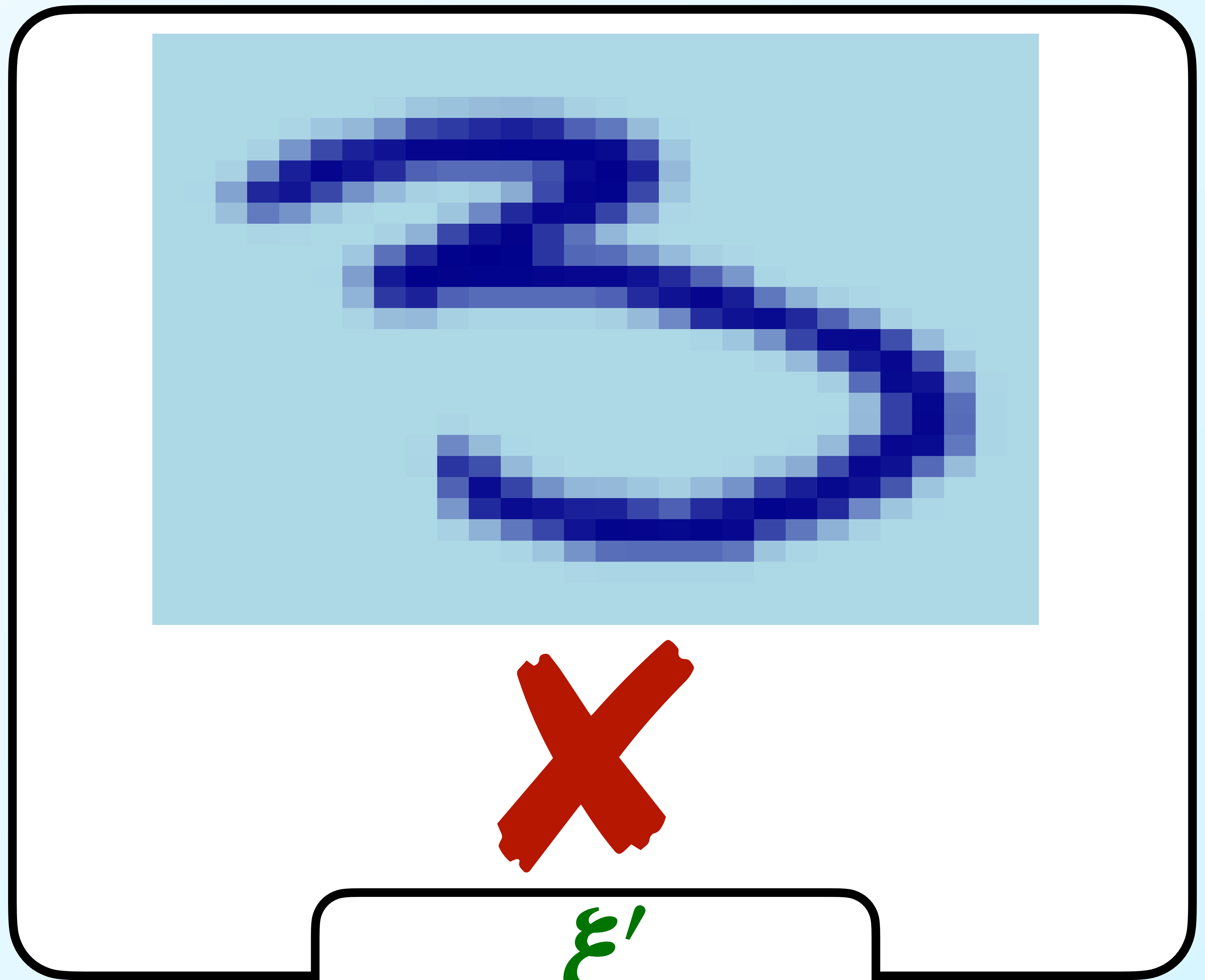
The Wasserstein Distance

Metric on the feature-label space Ξ

Example $\kappa = 28,\ q = 2$



ξ



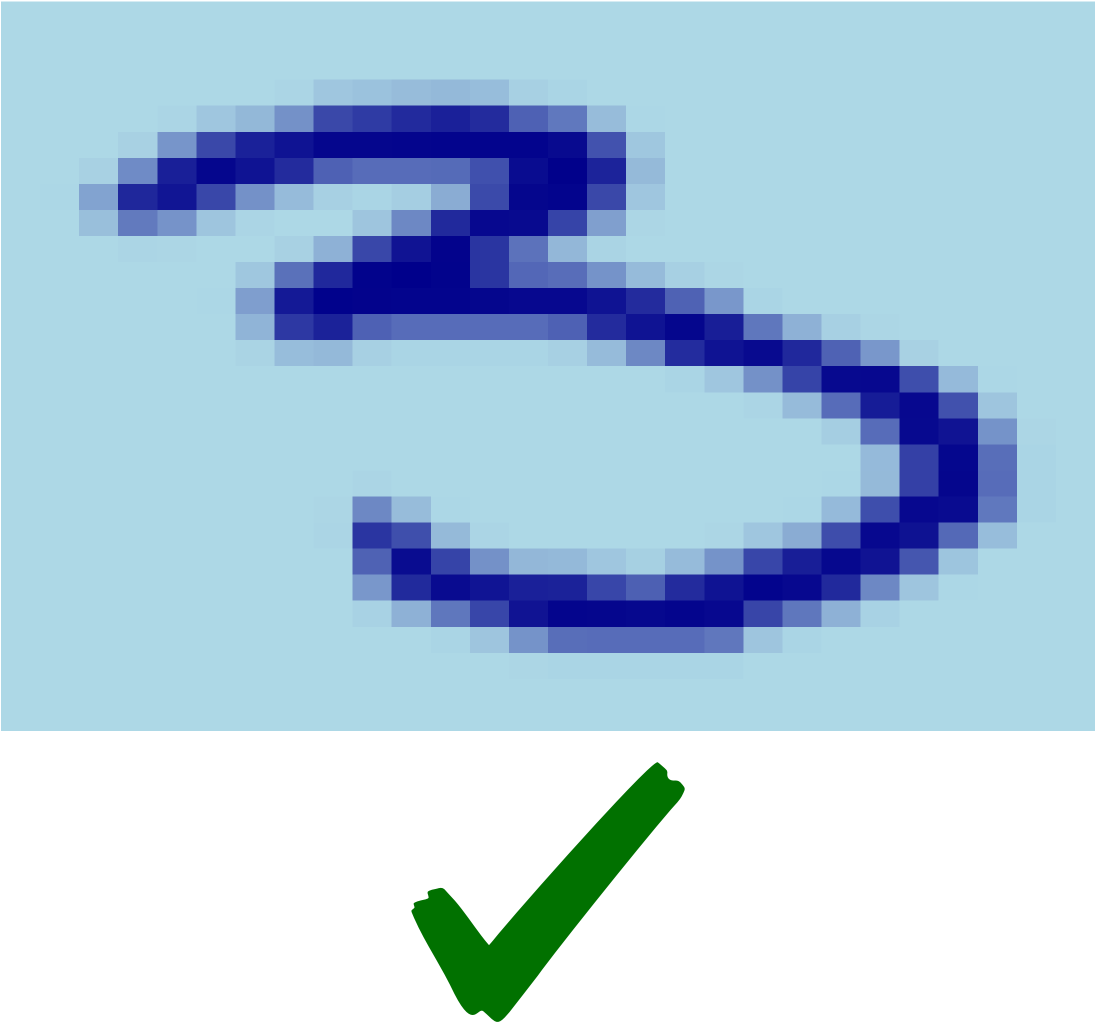
ξ'

$$d(\xi, \xi') = 0.5$$

The Wasserstein Distance

Metric on the feature-label space Ξ

Example $\kappa = 28,\ q = 2$



ξ

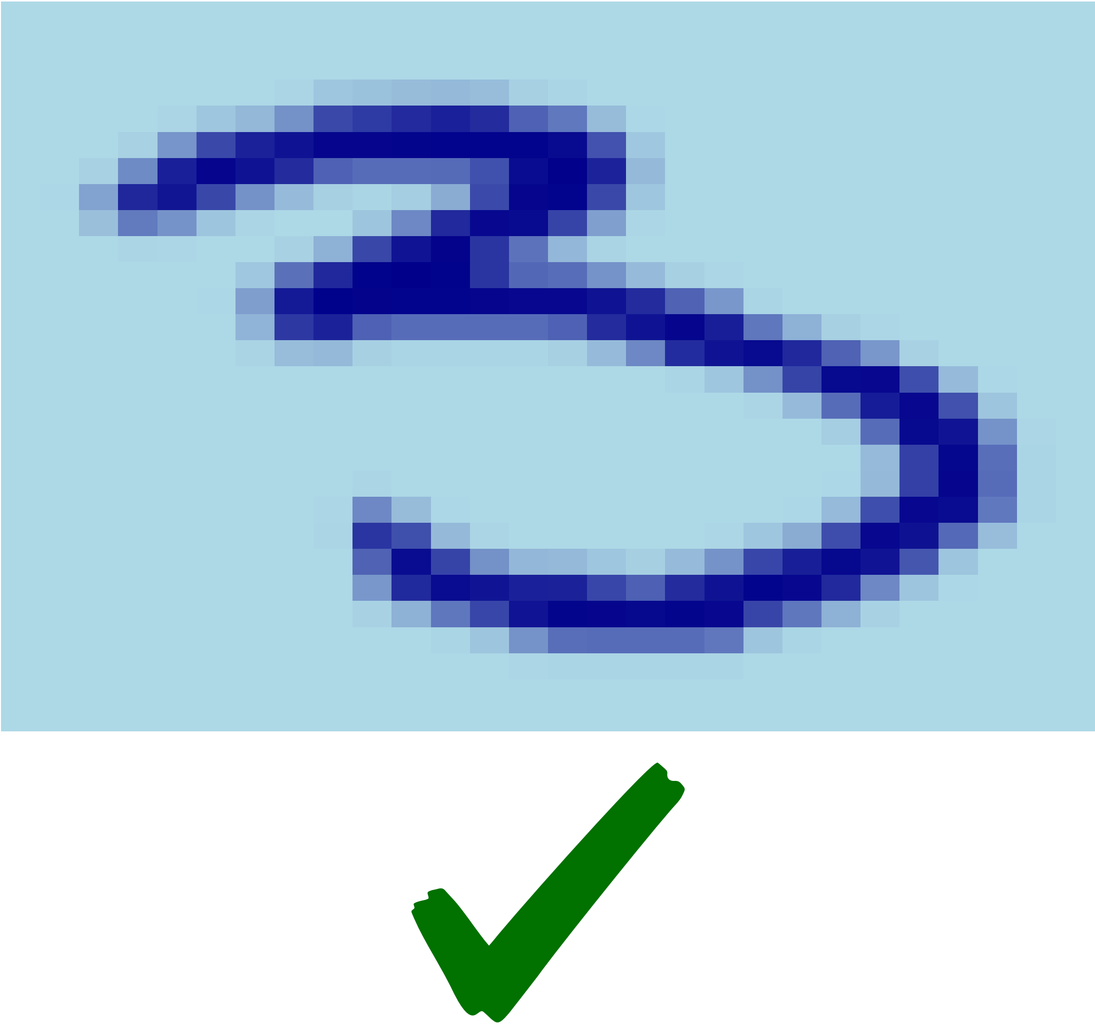
ξ'

$$d(\xi, \xi') = 0.21$$

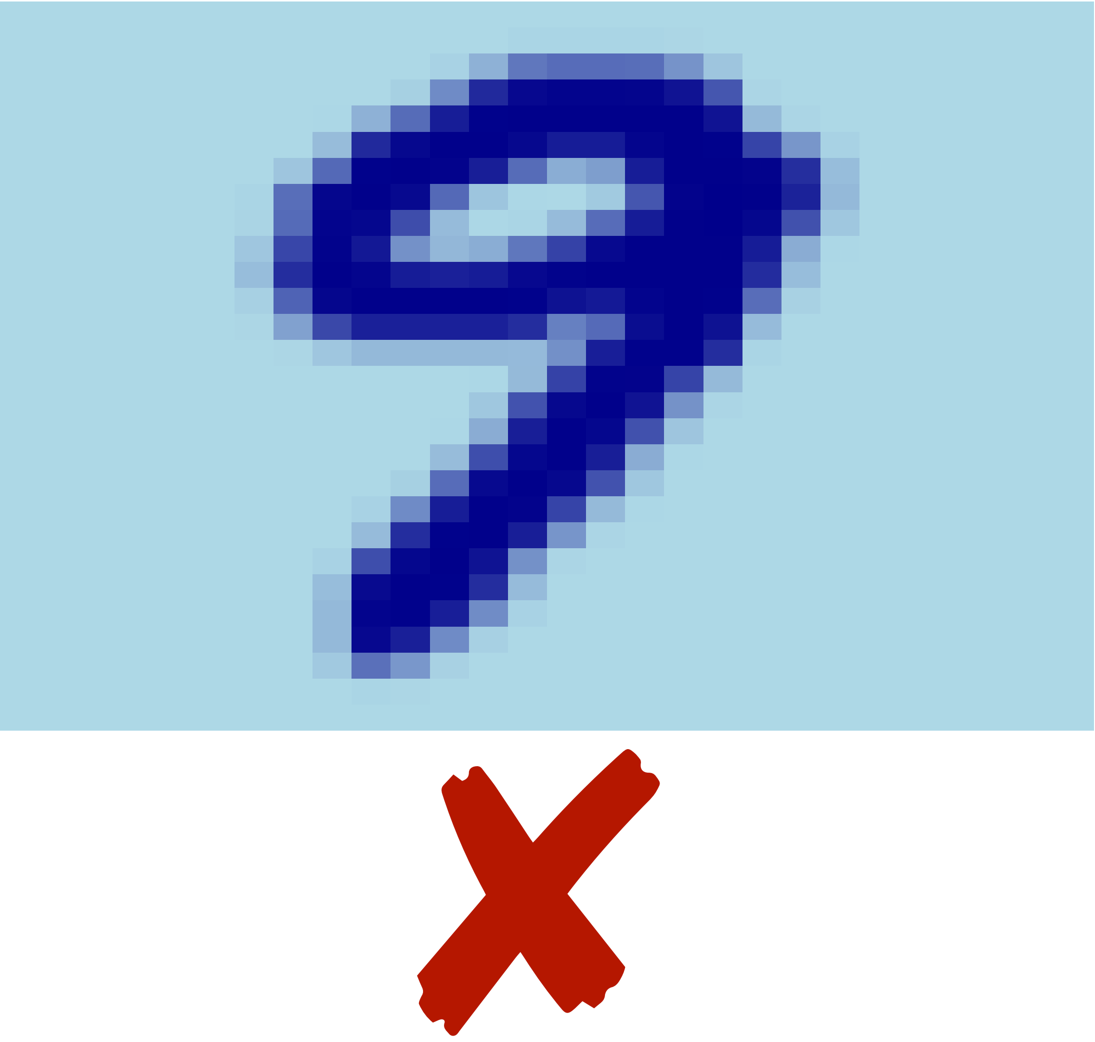
The Wasserstein Distance

Metric on the feature-label space Ξ

Example $\kappa = 28,\ q = 2$



ξ



ξ'

$$d(\xi, \xi') = 0.70$$

The Wasserstein Distance

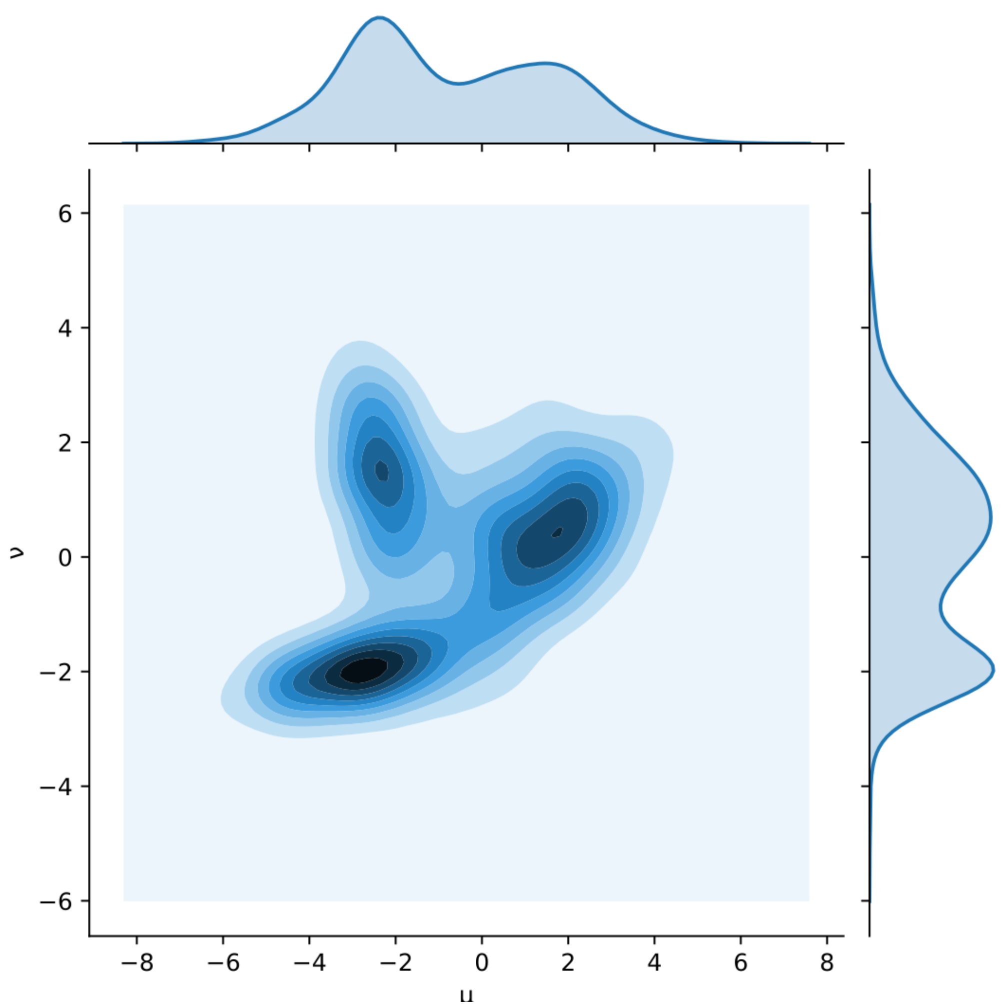
Metric on the feature-label space Ξ

Distance between $\xi = (x, y) \in \Xi$ and $\xi' = (x', y') \in \Xi$ is

$$d(\xi, \xi') = \|x - x'\|_q + \kappa \cdot \mathbf{1}[y \neq y']$$

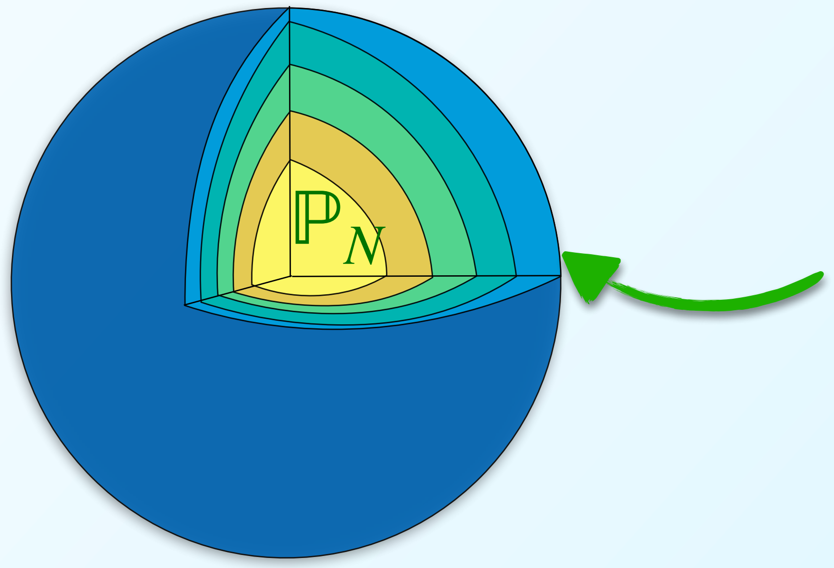
Wasserstein distance

$$\mathrm{W}(\mathbb{Q}, \mathbb{Q}') = \inf_{\Pi \in \mathscr{C}(\mathbb{Q}, \mathbb{Q}')} \mathbb{E}_{\Pi}[d(\xi, \xi')]$$



Wasserstein ball

$$\mathfrak{B}_{\varepsilon}(\mathbb{P}_N) = \{\mathbb{Q} : \ \mathrm{W}(\mathbb{P}_N, \mathbb{Q}) \leq \varepsilon\}$$



$$\mathbb{P}_N = \frac{1}{N} \sum_{i \in [N]} \delta_{\xi^i}$$

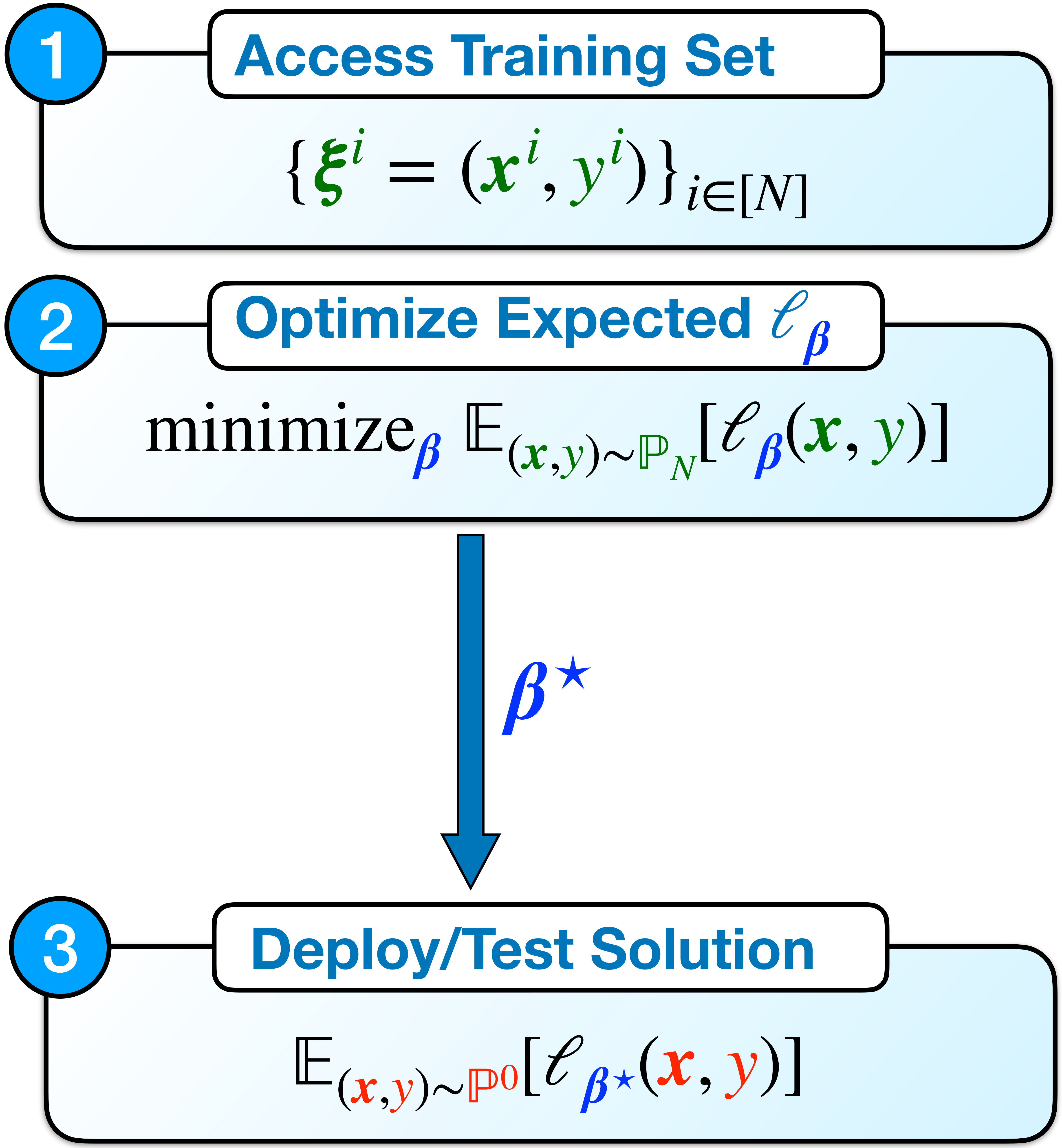
Distributionally Robust Optimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$

Distributionally Robust Optimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$

Adversarial Attacks



1 **Access Training Set**

$$\{\boldsymbol{\xi}^i = (\boldsymbol{x}^i, y^i)\}_{i\in[N]}$$

2 **Optimize Expected $\ell_{\boldsymbol{\beta}}$**

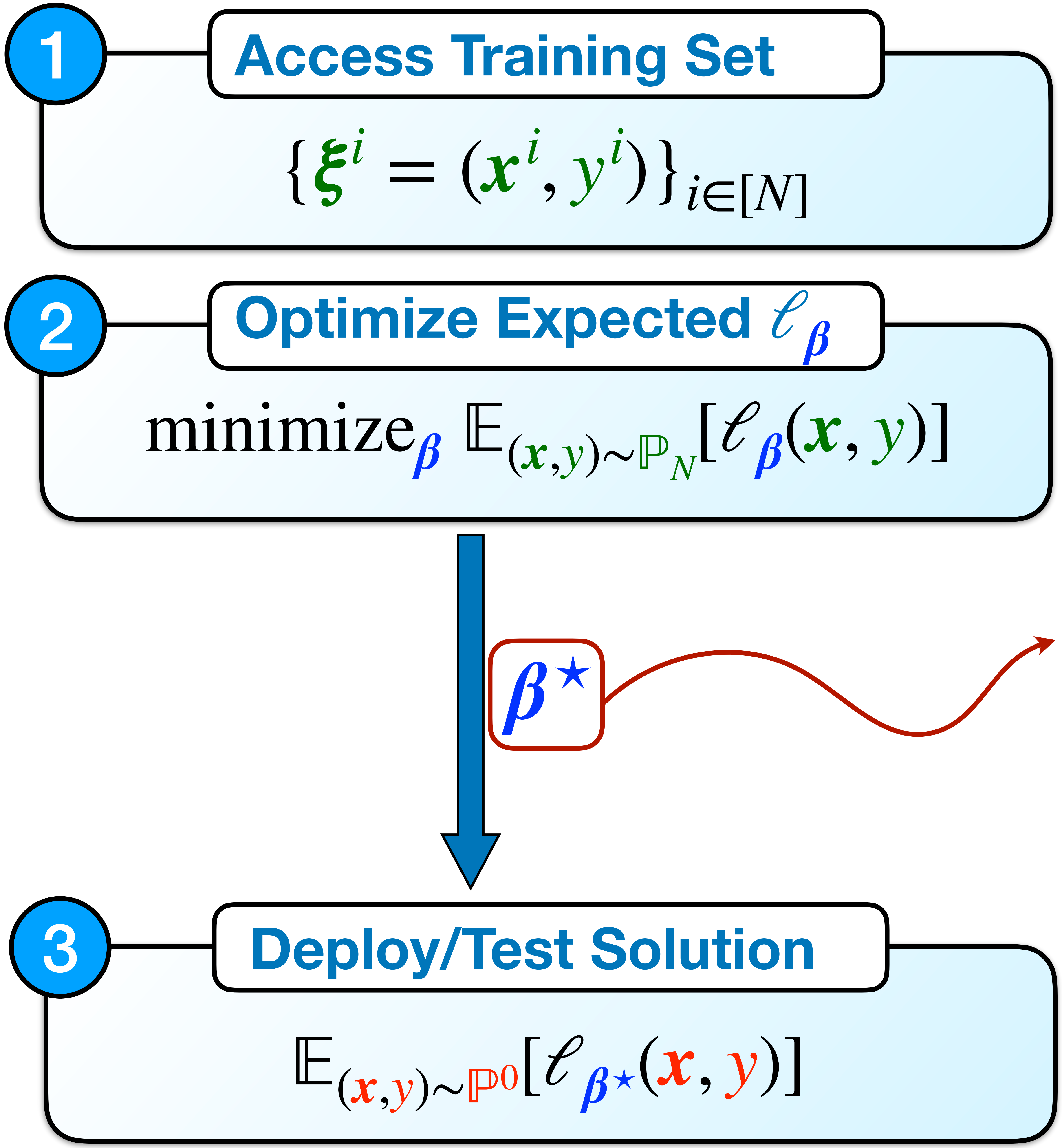
$$\text{minimize}_{\boldsymbol{\beta}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$$

$\boldsymbol{\beta}^\star$

3 **Deploy/Test Solution**

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}^\star}(\boldsymbol{x},y)]$$

Adversarial Attacks



1 **Access Training Set**

$$\{\boldsymbol{\xi}^i = (\boldsymbol{x}^i, y^i)\}_{i \in [N]}$$

2 **Optimize Expected $\ell_{\boldsymbol{\beta}}$**

$$\text{minimize}_{\boldsymbol{\beta}} \ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$$

$\boldsymbol{\beta}^\star$

3 **Deploy/Test Solution**

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}^\star}(\boldsymbol{x}, y)]$$

Adversarial Attacks

1 Access Training Set

$$\{\boldsymbol{\xi}^i = (\boldsymbol{x}^i, y^i)\}_{i \in [N]}$$

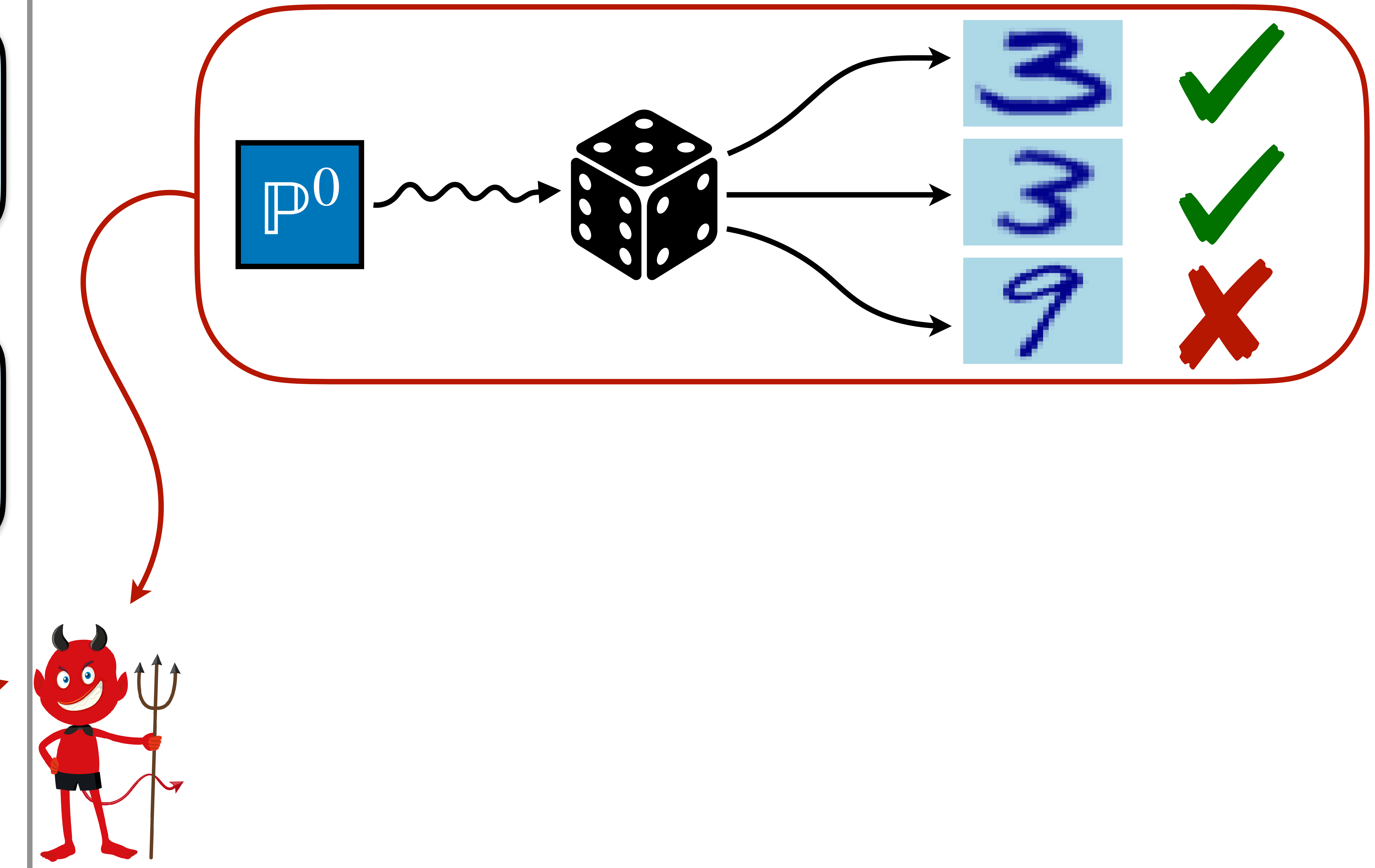
2 Optimize Expected $\ell_{\boldsymbol{\beta}}$

$$\text{minimize}_{\boldsymbol{\beta}} \ \mathbb{E}_{(\boldsymbol{x},y)\sim \mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$$

$\boldsymbol{\beta}^\star$

3 Deploy/Test Solution

$$\mathbb{E}_{(\boldsymbol{x},y)\sim \mathbb{P}^0}[\ell_{\boldsymbol{\beta}^\star}(\boldsymbol{x}, y)]$$



Adversarial Attacks

1 **Access Training Set**

$$\{\boldsymbol{\xi}^i = (\boldsymbol{x}^i, y^i)\}_{i\in[N]}$$

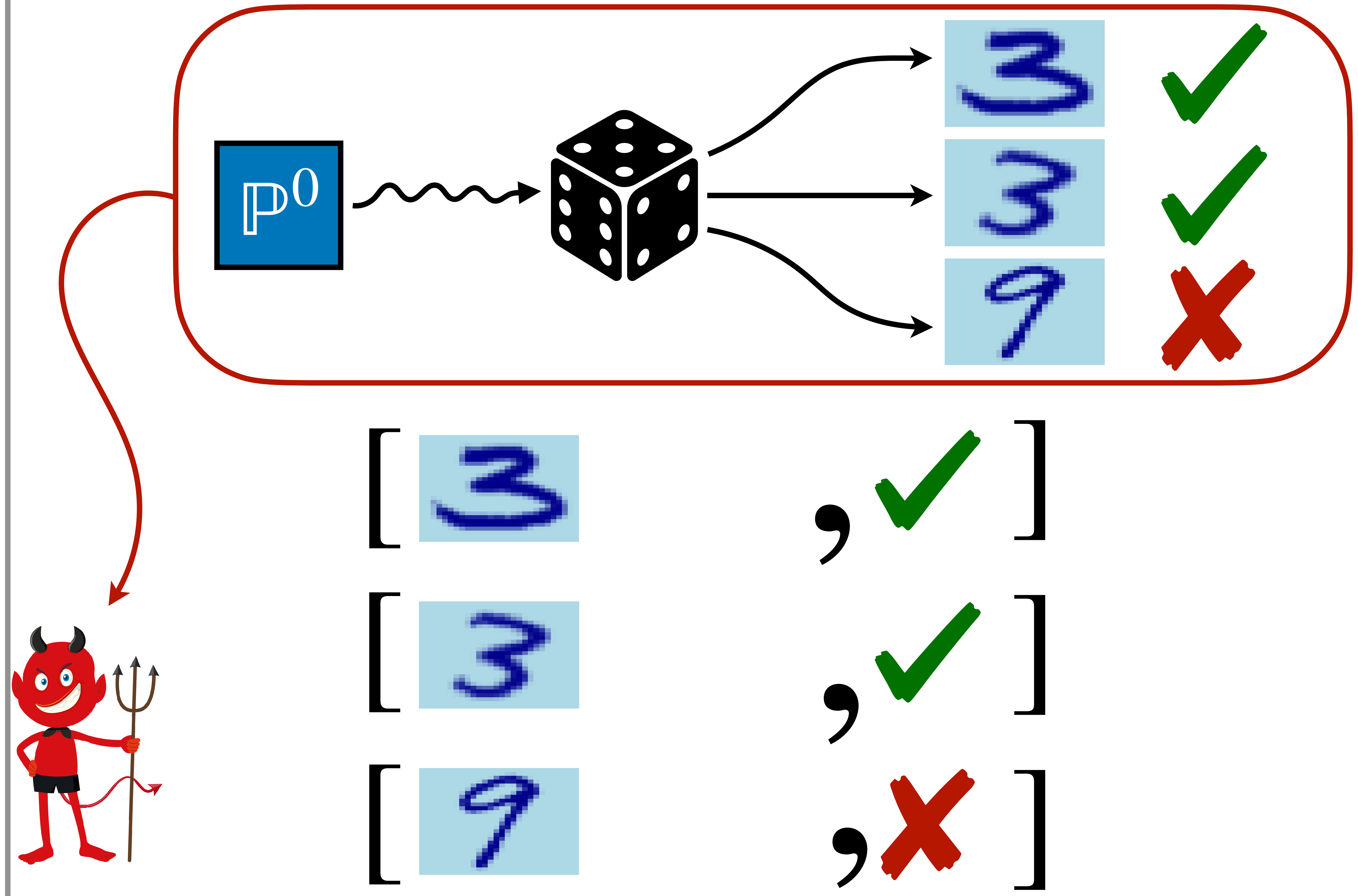
2 **Optimize Expected ℓ_β**

$$\text{minimize}_\beta \ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_\beta(\boldsymbol{x}, y)]$$

$\boldsymbol{\beta}^\star$

3 **Deploy/Test Solution**

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}^\star}(\boldsymbol{x}, y)]$$



Adversarial Attacks

1 **Access Training Set**

$$\{\boldsymbol{\xi}^i = (\boldsymbol{x}^i, y^i)\}_{i\in[N]}$$

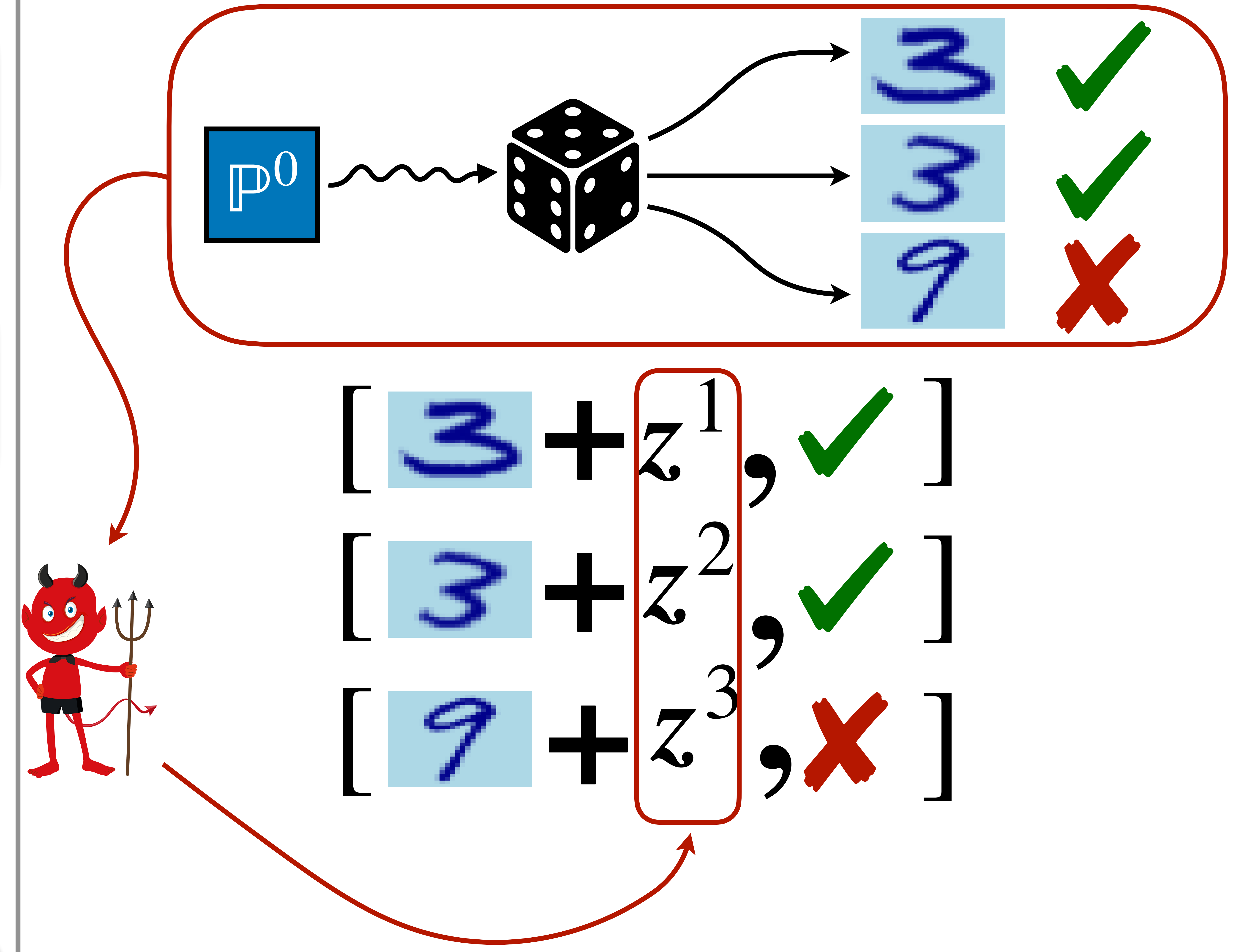
2 **Optimize Expected $\ell_{\boldsymbol{\beta}}$**

$$\text{minimize}_{\boldsymbol{\beta}} \; \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$$

$\boldsymbol{\beta}^\star$

3 **Deploy/Test Solution**

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}^\star}(\boldsymbol{x}, y)]$$



Adversarial Attacks

1 **Access Training Set**

$$\{\boldsymbol{\xi}^i = (\boldsymbol{x}^i, y^i)\}_{i\in[N]}$$

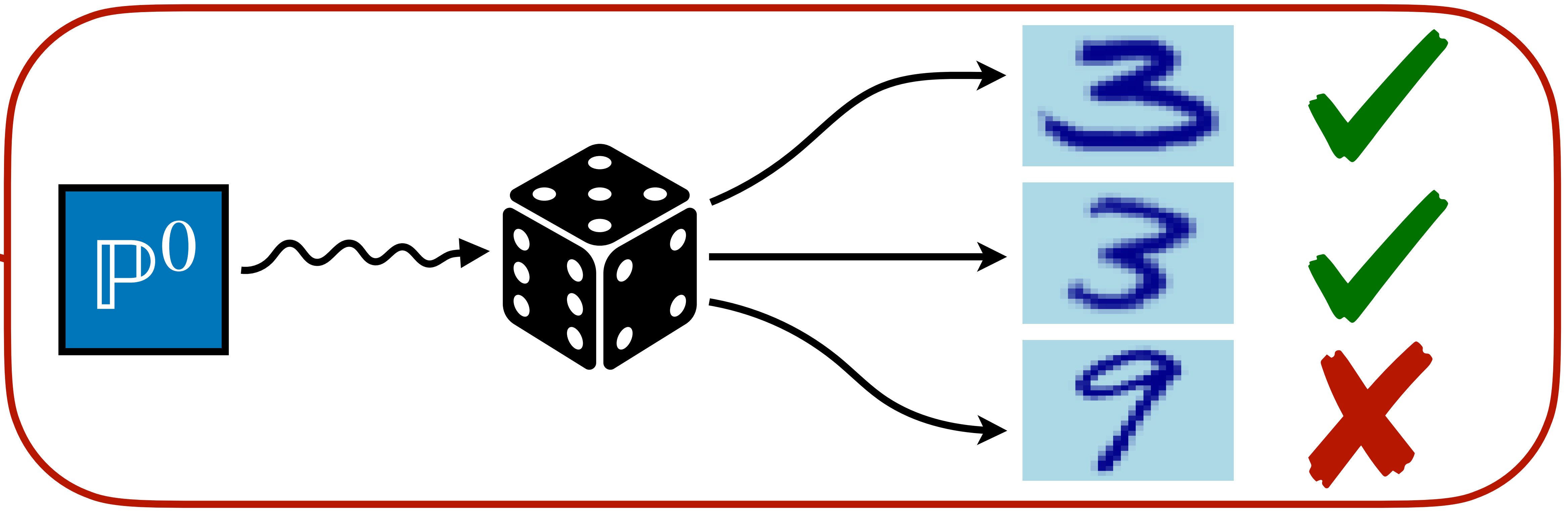
2 **Optimize Expected $\ell_{\boldsymbol{\beta}}$**

$$\text{minimize}_{\boldsymbol{\beta}} \ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$$

$\boldsymbol{\beta}^\star$

3 **Deploy/Test Solution**

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}^\star}(\boldsymbol{x}, y)]$$



3 **Adversarial Attacks**

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}\Big[\sup_{\|z\|_p\le\alpha} \ell_{\boldsymbol{\beta}^\star}(\boldsymbol{x}+z, y)\Big]$$

Adversarial Attacks

ℓ_2-attack for $\boldsymbol{\beta}^{\mathrm{ERM}}$

$\boldsymbol{\beta}^{\mathrm{ERM}^\top}\boldsymbol{x} = 34$

$\boldsymbol{\beta}^{\mathrm{ERM}^\top}(\boldsymbol{x}+\boldsymbol{z}) = 1$

$\boldsymbol{\beta}^{\mathrm{ERM}^\top}(\boldsymbol{x}+\boldsymbol{z}) = -76$

$\boldsymbol{\beta}^{\mathrm{ERM}^\top}(\boldsymbol{x}+\boldsymbol{z}) = -408$

Stronger attack radius α

Distributionally Robust Optimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Adversarially RO		$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\sup_{\Vert z\Vert_p\leq\alpha} \ell_{\boldsymbol{\beta}}(\boldsymbol{x}+z,y)]$

Distributionally Robust Optimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Adversarially RO	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\sup_{\lVert z\rVert_p\le\alpha} \ell_{\boldsymbol{\beta}}(\boldsymbol{x}+z,y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\sup_{\lVert z\rVert_p\le\alpha} \ell_{\boldsymbol{\beta}}(\boldsymbol{x}+z,y)]$

ℓ_2-attack for $\boldsymbol{\beta}^{\text{ARO}}$

$\boldsymbol{\beta}^{\text{ARO}^\top}\boldsymbol{x} = 1.90$

$\boldsymbol{\beta}^{\text{ARO}^\top}(\boldsymbol{x}+\boldsymbol{z}) = 1.80$

$\boldsymbol{\beta}^{\text{ARO}^\top}(\boldsymbol{x}+\boldsymbol{z}) = 1.56$

$\boldsymbol{\beta}^{\text{ARO}^\top}(\boldsymbol{x}+\boldsymbol{z}) = 0.53$

Stronger attack radius α

Distributionally Robust Optimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Adversarially RO	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\sup_{\lVert z\rVert_p\leq\alpha} \ell_{\boldsymbol{\beta}}(\boldsymbol{x}+z,y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\sup_{\lVert z\rVert_p\leq\alpha} \ell_{\boldsymbol{\beta}}(\boldsymbol{x}+z,y)]$

Distributionally Robust Optimization

Robust Overfitting

- ARO models overfit despite being “robust”
- Even more severe than overfitting of ERM
- We want DRO and ARO simultaneously

Paradigm		e Risk
Empirical Risk Min		$_0[\ell_\beta(x, y)]$
Distributionally RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(x,y)\sim\mathbb{Q}}[\ell_\beta(x, y)]$	$\mathbb{E}_{(x,y)\sim\mathbb{P}^0}[\ell_\beta(x, y)]$
Adversarially RO	$\mathbb{E}_{(x,y)\sim\mathbb{P}_N}[\sup_{\|z\|_p\leq\alpha} \ell_\beta(x + z, y)]$	$\mathbb{E}_{(x,y)\sim\mathbb{P}^0}[\sup_{\|z\|_p\leq\alpha} \ell_\beta(x + z, y)]$

Adversarial Attacks

$$\sup_{\|z\|_p \leq \alpha} \ell_{\beta}(x + z, y) = \sup_{\|z\|_p \leq \alpha} \log(1 + \exp(-y \cdot \beta^\top (x + z)))$$

Standard RO techniques: Bertsimas et al. (2019)

$$\sup_{\|z\|_p \le \alpha} \ell_{\beta}(x + z, y) = \sup_{\|z\|_p \le \alpha} \log(1 + \exp(-y \cdot \beta^\top (x + z)))$$

$$= \log(1 + \exp(\sup_{\|z\|_p \le \alpha} \{-y \cdot \beta^\top (x + z)\}))$$

Monotonicity: supremum goes inside

Adversarial Attacks

$$\sup_{\|z\|_p\leq\alpha} \ell_{\boldsymbol{\beta}}(\boldsymbol{x}+z, y) = \sup_{\|z\|_p\leq\alpha} \log(1+\exp(-y\cdot\boldsymbol{\beta}^\top(\boldsymbol{x}+z)))$$

$$= \log(1+\exp(\sup_{\|z\|_p\leq\alpha}\{-y\cdot\boldsymbol{\beta}^\top(\boldsymbol{x}+z)\}))$$

$$= \log(1+\exp(-y\cdot\boldsymbol{\beta}^\top\boldsymbol{x}+\sup_{\|z\|_p\leq\alpha}\{-y\cdot\boldsymbol{\beta}^\top z\}))$$

Rewrite: take constants out

Adversarial Attacks

$$\sup_{\|z\|_p \leq \alpha} \ell_{\beta}(x + z, y) = \sup_{\|z\|_p \leq \alpha} \log(1 + \exp(-y \cdot \beta^\top (x + z)))$$

$$= \log(1 + \exp(\sup_{\|z\|_p \leq \alpha} \{-y \cdot \beta^\top (x + z)\}))$$

$$= \log(1 + \exp(-y \cdot \beta^\top x + \sup_{\|z\|_p \leq \alpha} \{-y \cdot \beta^\top z\}))$$

$$= \log(1 + \exp(-y \cdot \beta^\top x + \alpha \cdot \|-y \cdot \beta\|_{p^\star}))$$

Dual norm: use definition

Adversarial Attacks

$$\sup_{\|z\|_p \le \alpha} \ell_{\boldsymbol{\beta}}(\boldsymbol{x} + z, y) = \sup_{\|z\|_p \le \alpha} \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top(\boldsymbol{x} + z)))$$

$$\Big\downarrow$$

$$= \log(1 + \exp(\sup_{\|z\|_p \le \alpha} \{-y \cdot \boldsymbol{\beta}^\top(\boldsymbol{x} + z)\}))$$

$$\Big\downarrow$$

$$= \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \sup_{\|z\|_p \le \alpha} \{-y \cdot \boldsymbol{\beta}^\top z\}))$$

$$\Big\downarrow$$

$$= \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star}))$$

Adversarial Attacks

$$\sup_{\|z\|_p \leq \alpha} \ell_{\boldsymbol{\beta}}(\boldsymbol{x} + z, y) = \sup_{\|z\|_p \leq \alpha} \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top(\boldsymbol{x} + z)))$$

$$\Big\downarrow$$

$$= \log(1 + \exp(\sup_{\|z\|_p \leq \alpha} \{-y \cdot \boldsymbol{\beta}^\top(\boldsymbol{x} + z)\}))$$

$$\Big\downarrow$$

$$= \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \sup_{\|z\|_p \leq \alpha} \{-y \cdot \boldsymbol{\beta}^\top z\}))$$

$$\Big\downarrow$$

$$= \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) =: \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, y)$$

Adversarial Attacks

$$\sup_{\|z\|_p \le \alpha} \ell_{\boldsymbol{\beta}}(\boldsymbol{x} + z, y) = \sup_{\|z\|_p \le \alpha} \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top(\boldsymbol{x} + z)))$$

$= \log(1 + \mathrm{ex}$ $z)\}))$

Adversarial loss

- Can be interpreted as a new loss function
- Convex and Lipschitz
- Existing Lipschitz Wasserstein DRO theory is applicable

$$= \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) =: \boxed{\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, y)}$$

Adversarially Robust Optimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally RO	$\sup\limits_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Adversarially RO	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\sup\limits_{\Vert z\Vert_p\leq\alpha} \ell_{\boldsymbol{\beta}}(\boldsymbol{x}+z,y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\sup\limits_{\Vert z\Vert_p\leq\alpha} \ell_{\boldsymbol{\beta}}(\boldsymbol{x}+z,y)]$

Adversarially Robust Optimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally RO	$\sup\limits_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Adversarially RO	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$

Adversarially Robust Optimization

Paradigm	**Training Risk**	**True Risk**
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Adversarially RO	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\boxed{\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]}$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\boxed{\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]}$

ARO calibrates the loss

Adversarially Robust Optimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Adversarially RO	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\boxed{\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]}$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\boxed{\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]}$

Statistical error
$\mathrm{W}(\mathbb{P}_N, \mathbb{P}^0)$ **stays**

ARO calibrates the loss

Distributionally & Adversarially Robust Optimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Adversarially RO	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally & Adversarially RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$

Exact Convex Reformulation

Distributionally and Adversarially Robust Optimization Problem **(DR-ARO)**

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \quad \sup_{\mathbb{Q} \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)} \quad \mathbb{E}_{(x,y)\sim\mathbb{Q}}\Big[\sup_{\|z\|_p \leq \alpha} \{\ell_\beta(x+z,y)\}\Big]$$

Exact Convex Reformulation

Distributionally and Adversarially Robust Optimization Problem **(DR-ARO)**

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{minimize}} \quad \sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\sup_{\|z\|_p\leq\alpha} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+z, y)\}]$$

$$\begin{array}{lll} \underset{\boldsymbol{\beta},\lambda,\boldsymbol{s}}{\text{minimize}} & \varepsilon\lambda+\frac{1}{N}\sum_{i=1}^{N} s_i & \\ \text{subject to} & \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}^i, y^i) \leq s_i & \forall i \in [N] \\ & \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}^i,-y^i)-\lambda\kappa \leq s_i & \forall i \in [N] \\ & \|\boldsymbol{\beta}\|_{q^\star} \leq \lambda & \\ & \boldsymbol{\beta} \in \mathbb{R}^n, \lambda \geq 0, \boldsymbol{s} \in \mathbb{R}^N_+ . & \end{array}$$

Exact Convex Reformulation

Distributionally and Adversarially Robust Optimization Problem **(DR-ARO)**

$$\underset{\beta\in\mathbb{R}^n}{\text{minimize}} \quad \sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \quad \mathbb{E}_{(x,y)\sim\mathbb{Q}}\Big[\sup_{\|z\|_p\leq\alpha}\{\ell_\beta(x+z,y)\}\Big]$$

$$\begin{array}{lll}
\underset{\beta,\lambda,s}{\text{minimize}} & \boxed{\varepsilon\lambda+\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}s_i} \xrightarrow{\quad} \textbf{Linear obj.} & \\
\text{subject to} & \ell^\alpha_\beta(x^i,y^i)\leq s_i & \forall i\in[N] \\
& \ell^\alpha_\beta(x^i,-y^i)-\lambda\kappa\leq s_i & \forall i\in[N] \\
& \|\beta\|_{q^\star}\leq\lambda & \\
& \beta\in\mathbb{R}^n,\ \lambda\geq 0,\ s\in\mathbb{R}^N_+. &
\end{array}$$

DR-ARO

Exact Convex Reformulation

Distributionally and Adversarially Robust Optimization Problem **(DR-ARO)**

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{minimize}} \quad \sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}\Big[\sup_{\|z\|_p\le\alpha}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+z,y)\}\Big]$$

$$\begin{array}{lll}
\underset{\boldsymbol{\beta},\lambda,\boldsymbol{s}}{\text{minimize}} & \varepsilon\lambda+\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i & \\
\text{subject to} & \ell_{\boldsymbol{\beta}}^{\alpha}(\boldsymbol{x}^i, y^i) \le s_i & \forall i\in[N] \\
 & \ell_{\boldsymbol{\beta}}^{\alpha}(\boldsymbol{x}^i, -y^i) - \lambda\kappa \le s_i & \forall i\in[N] \\
 & \|\boldsymbol{\beta}\|_{q^\star} \le \lambda & \\
 & \boldsymbol{\beta}\in\mathbb{R}^n,\ \lambda\ge 0,\ \boldsymbol{s}\in\mathbb{R}^N_+ . & \longrightarrow \mathcal{O}(N)\ \textbf{many}
\end{array}$$

Exact Convex Reformulation

Distributionally and Adversarially Robust Optimization Problem **(DR-ARO)**

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \quad \sup_{\mathbb{Q} \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)} \quad \mathbb{E}_{(x,y)\sim\mathbb{Q}}\left[\sup_{\|z\|_p \le \alpha} \{\ell_\beta(x + z, y)\}\right]$$

$$\begin{array}{lll} \underset{\beta,\lambda,s}{\text{minimize}} & \varepsilon\lambda + \frac{1}{N}\sum_{i=1}^{N} s_i & \\ \text{subject to} & \ell_\beta^\alpha(x^i, y^i) \le s_i & \forall i \in [N] \\ & \ell_\beta^\alpha(x^i, -y^i) - \lambda\kappa \le s_i & \forall i \in [N] \\ & \|\beta\|_{q^\star} \le \lambda & \\ & \beta \in \mathbb{R}^n, \lambda \ge 0, s \in \mathbb{R}_+^N. & \end{array}$$

Convex for
$q \in \{1, 2, \infty\}$

Exact Convex Reformulation

Distributionally and Adversarially Robust Optimization Problem **(DR-ARO)**

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{minimize}} \quad \sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}\Big[\sup_{\|z\|_p\leq\alpha}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+z,y)\}\Big]$$

$$\begin{array}{lll} \underset{\boldsymbol{\beta},\lambda,\boldsymbol{s}}{\text{minimize}} & \varepsilon\lambda+\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i & \\ \text{subject to} & \ell_{\boldsymbol{\beta}}^{\alpha}(\boldsymbol{x}^i, y^i) \leq s_i & \forall i \in [N] \\ & \ell_{\boldsymbol{\beta}}^{\alpha}(\boldsymbol{x}^i, -y^i) - \lambda\kappa \leq s_i & \forall i \in [N] \\ & \|\boldsymbol{\beta}\|_{q^\star} \leq \lambda & \\ & \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \boldsymbol{s} \in \mathbb{R}^N_+. & \end{array}$$

Exponential cone repr.

Exact Convex Reformulation

Distributionally and Adversarially Robust Optimization Problem **(DR-ARO)**

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \quad \sup_{\mathbb{Q} \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)} \quad \mathbb{E}_{(x,y)\sim\mathbb{Q}}[\ \sup_{\|z\|_p \leq \alpha} \{\ell_\beta(x+z, y)\}]$$

$$\begin{array}{ll} \underset{\beta,\lambda,s}{\text{minimize}} & \varepsilon\lambda + \frac{1}{N}\sum_{i=1}^{N} s_i \\ \text{subject to} & \ell_\beta^\alpha(x^i, y^i) \leq s_i \\ & \ell_\beta^\alpha(x^i, -y^i) - \lambda\kappa \leq s_i \\ & \|\beta\|_{q^\star} \leq \lambda \\ & \beta \in \mathbb{R}^n,\ \lambda \geq 0,\ s \in \mathbb{R}_+^N. \end{array}$$

- Adversarial loss being ℓ_β^α
- ℓ_β^α being convex & Lipschitz
- Shafieezadeh-Abadeh (2019)

Distributionally & Adversarially Robust Optimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Adversarially RO	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally & Adversarially RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$

How we address Robust Overfitting

Distributionally & Adversarially Robust Optimization

Paradigm	Training Risk	True Risk
Empirical Risk Min.	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Adversarially RO	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$
Distributionally & Adversarially RO	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}^0}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$

What other approaches?

Literature (e.g., Xing et al., 2022)

Data

$$\{\boldsymbol{\xi}^i = (\boldsymbol{x}^i, y^i)\}_{i\in[N]}$$

$\sim$iid $\mathbb{P}^0$

$$\mathbb{P}_N := \frac{1}{N}\sum_{i\in[N]} \delta_{\boldsymbol{\xi}^i}$$

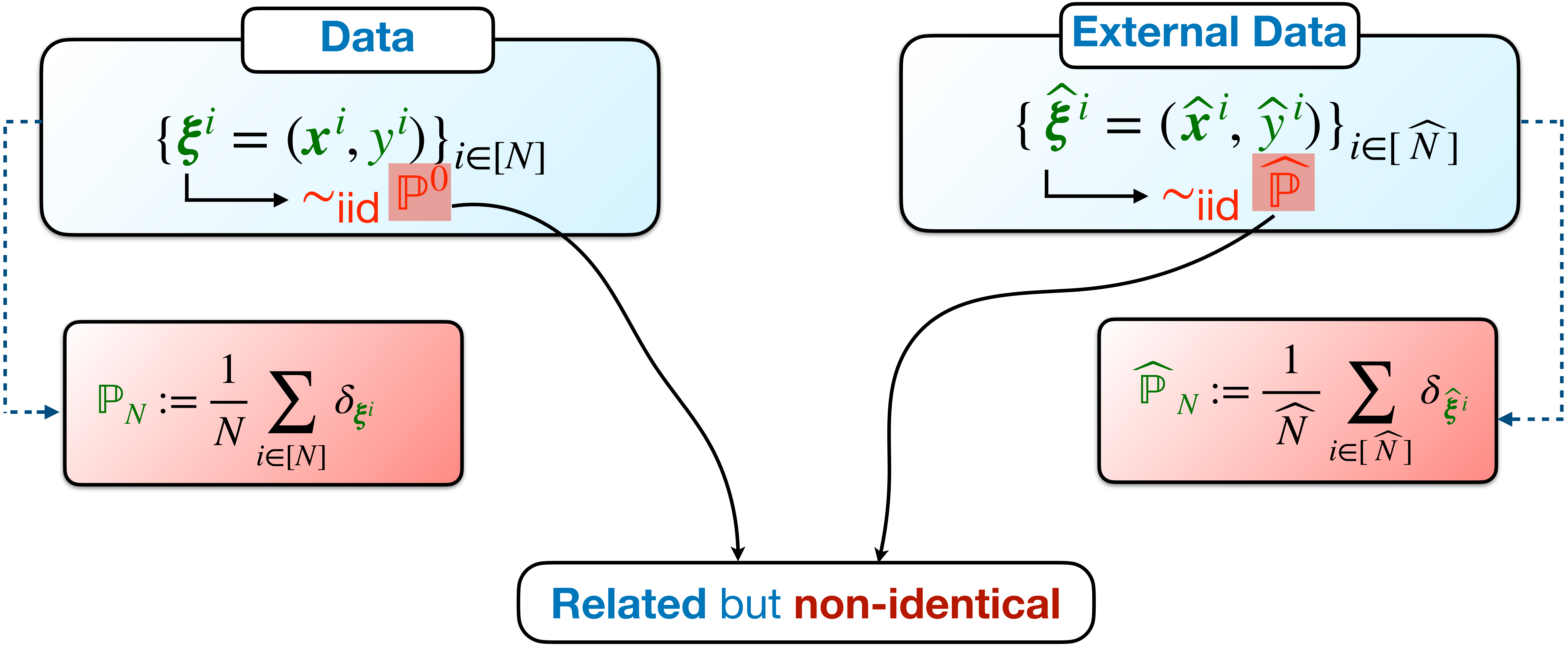
External Data

$$\{\widehat{\boldsymbol{\xi}}^i = (\widehat{\boldsymbol{x}}^i, \widehat{y}^i)\}_{i\in[\widehat{N}]}$$

$\sim$iid $\widehat{\mathbb{P}}$

$$\widehat{\mathbb{P}}_N := \frac{1}{\widehat{N}}\sum_{i\in[\widehat{N}]} \delta_{\widehat{\boldsymbol{\xi}}^i}$$

Literature (e.g., Xing et al., 2022)



Data

$$\{\boldsymbol{\xi}^i = (\boldsymbol{x}^i, y^i)\}_{i\in[N]}$$

$\sim$iid $\mathbb{P}^0$

$$\mathbb{P}_N := \frac{1}{N}\sum_{i\in[N]} \delta_{\boldsymbol{\xi}^i}$$

External Data

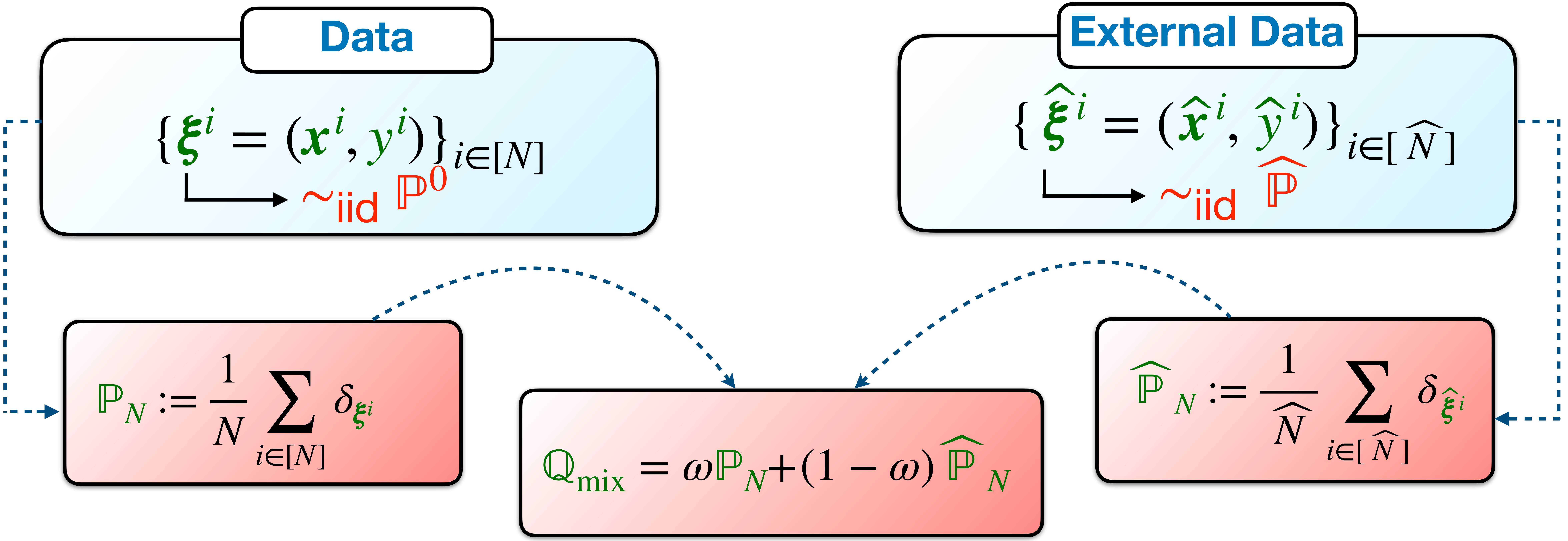
$$\{\widehat{\boldsymbol{\xi}}^i = (\widehat{\boldsymbol{x}}^i, \widehat{y}^i)\}_{i\in[\widehat{N}]}$$

$\sim$iid $\widehat{\mathbb{P}}$

$$\widehat{\mathbb{P}}_N := \frac{1}{\widehat{N}}\sum_{i\in[\widehat{N}]} \delta_{\widehat{\boldsymbol{\xi}}^i}$$

Related but **non-identical**

Literature (e.g., Xing et al., 2022)



Data

$$\{\boldsymbol{\xi}^i = (\boldsymbol{x}^i, y^i)\}_{i\in[N]}$$

$\sim_{\text{iid}}$ $\mathbb{P}^0$

External Data

$$\{\widehat{\boldsymbol{\xi}}^i = (\widehat{\boldsymbol{x}}^i, \widehat{y}^i)\}_{i\in[\widehat{N}]}$$

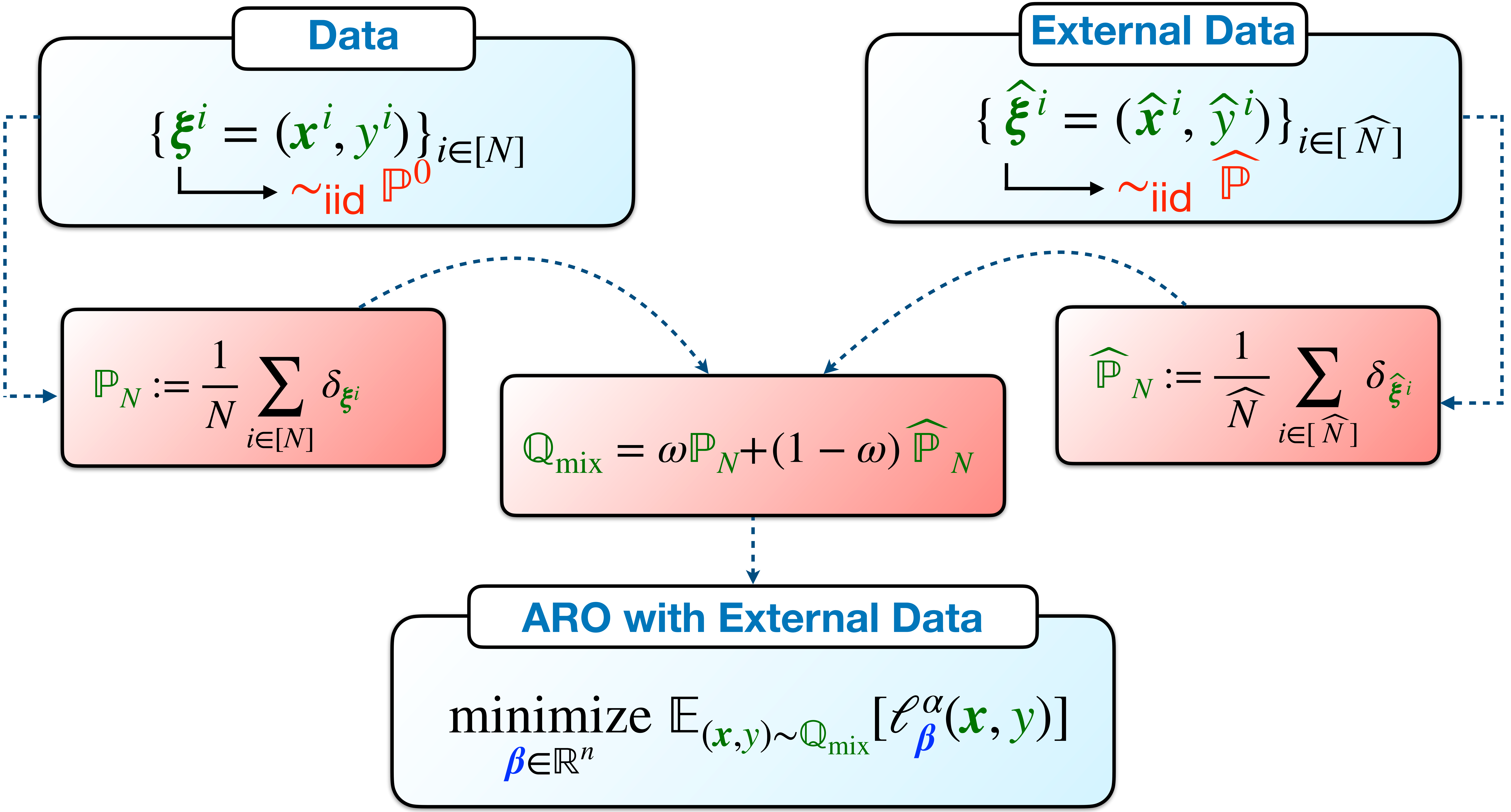
$\sim_{\text{iid}}$ $\widehat{\mathbb{P}}$

$$\mathbb{P}_N := \frac{1}{N}\sum_{i\in[N]} \delta_{\boldsymbol{\xi}^i}$$

$$\mathbb{Q}_{\text{mix}} = \omega\mathbb{P}_N + (1-\omega)\widehat{\mathbb{P}}_N$$

$$\widehat{\mathbb{P}}_N := \frac{1}{\widehat{N}}\sum_{i\in[\widehat{N}]} \delta_{\widehat{\boldsymbol{\xi}}^i}$$

Literature (e.g., Xing et al., 2022)



Data

$$\{\boldsymbol{\xi}^i = (\boldsymbol{x}^i, y^i)\}_{i\in[N]} \quad \sim_{\text{iid}} \mathbb{P}^0$$

External Data

$$\{\widehat{\boldsymbol{\xi}}^i = (\widehat{\boldsymbol{x}}^i, \widehat{y}^i)\}_{i\in[\widehat{N}]} \quad \sim_{\text{iid}} \widehat{\mathbb{P}}$$

$$\mathbb{P}_N := \frac{1}{N}\sum_{i\in[N]} \delta_{\boldsymbol{\xi}^i}$$

$$\widehat{\mathbb{P}}_N := \frac{1}{\widehat{N}}\sum_{i\in[\widehat{N}]} \delta_{\widehat{\boldsymbol{\xi}}^i}$$

$$\mathbb{Q}_{\text{mix}} = \omega\mathbb{P}_N + (1-\omega)\widehat{\mathbb{P}}_N$$

ARO with External Data

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{minimize}} \ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}_{\text{mix}}}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$$

Literature (e.g., Xing et al., 2022)

Data

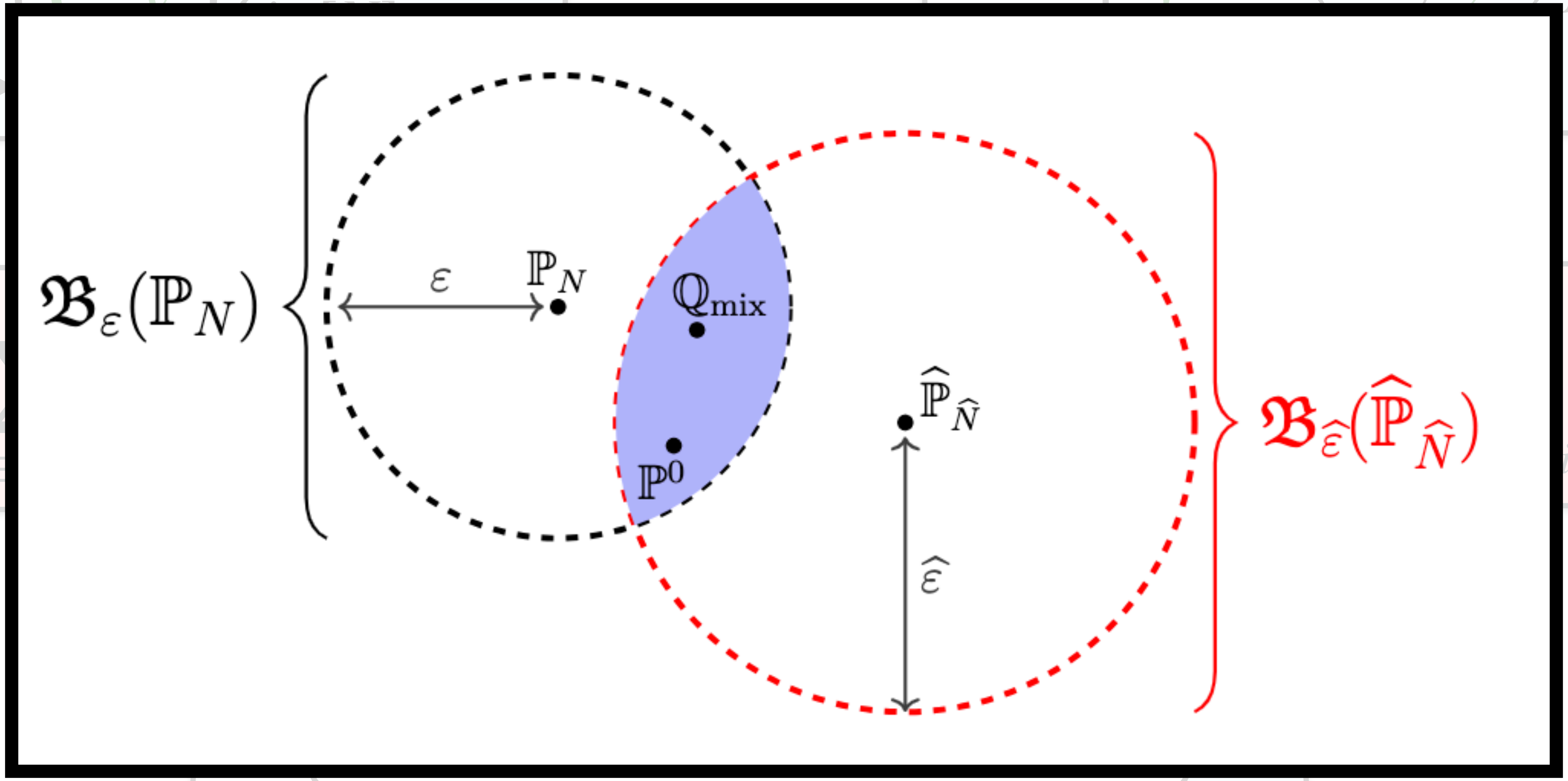
$\{\xi^i = (x^i, y^i)\}$

External Data

$\{\widehat{\xi}^i = (\widehat{x}^i, \widehat{y}^i)\}_{i\in[\widehat{N}]}$

$\mathbb{P}_N := \frac{1}{N}$

$\sum_{i\in[\widehat{N}]} \delta_{\widehat{\xi}^i}$



$$\underset{\beta\in\mathbb{R}^n}{\text{minimize}}\ \mathbb{E}_{(x,y)\sim\mathbb{Q}_{\text{mix}}}[\ell^{\alpha}_{\beta}(x,y)]$$

Inter-ARO

Exact Reformulation

ARO over intersection of Wasserstein balls **(Inter-ARO)**:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{minimize}} \quad \sup_{\mathbb{Q}\in\mathfrak{B}_{\varepsilon}(\mathbb{P}_N)\cap\mathfrak{B}_{\hat{\varepsilon}}(\widehat{\mathbb{P}}_N)} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[\sup_{\|z\|_p\leq\alpha}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+z,y)\}]$$

Exact Reformulation

ARO over intersection of Wasserstein balls **(Inter-ARO)**:

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \quad \sup_{\mathbb{Q} \in \mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\hat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})} \quad \mathbb{E}_{(x,y) \sim \mathbb{Q}}[\sup_{\|z\|_p \le \alpha} \{\ell_{\beta}(x + z, y)\}]$$

1

NP-hard even if $N = 1$ and $\widehat{N} = 1$.

Exact Reformulation

ARO over intersection of Wasserstein balls **(Inter-ARO)**:

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \quad \sup_{\mathbb{Q} \in \mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\hat{\varepsilon}}(\widehat{\mathbb{P}}_{N})} \quad \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathbb{Q}} \Big[\sup_{\|z\|_p \leq \alpha} \{ \ell_{\beta}(\boldsymbol{x} + z, y) \} \Big]$$

1

NP-hard even if $N = 1$ and $\widehat{N} = 1$.

2

Would admit an exact tractable reformulation if

- Squared-loss function (regression)
- Wasserstein ball around first and second moments
- No attack ($\alpha = 0$)

Taskesen et al. (2021)

Tractable Relaxation of Inter-ARO

Static Relaxation Technique

We consult to the **adjustable RO** literature for a **relaxation** that is:

- Convex with $\mathcal{O}(N \cdot \widehat{N})$ exponential conic constraints

Tractable Relaxation of Inter-ARO

Static Relaxation Technique

We consult to the **adjustable RO** literature for a **relaxation** that is:

- Convex with $\mathcal{O}(N \cdot \widehat{N})$ exponential conic constraints
- Coincides with the **exact** formulation when $\widehat{\varepsilon} \to \infty$

Tractable Relaxation of Inter-ARO

Static Relaxation Technique

We consult to the **adjustable RO** literature for a **relaxation** that is:

- Convex with $\mathcal{O}(N \cdot \widehat{N})$ exponential conic constraints
- Coincides with the **exact** formulation when $\widehat{\varepsilon} \to \infty$
- "**Not learning** from the external distribution" is **always feasible**

Tractable Relaxation of Inter-ARO

Static Relaxation Technique

We consult to the **adjustable RO** literature for a **relaxation** that is:

- Convex with $\mathcal{O}(N \cdot \widehat{N})$ exponential conic constraints
- Coincides with the **exact** formulation when $\widehat{\varepsilon} \to \infty$
- "**Not learning** from the external distribution" is **always feasible**
- Only uses the **external data** if it improves the objective

Tractable Relaxation of Inter-ARO

Static Relaxation Technique

We consult to the **adjustable RO** literature for a **relaxation** that is:

- Convex with $\mathcal{O}(N \cdot \widehat{N})$ exponential conic constraints
- Coincides with the **exact** formulation when $\widehat{\varepsilon} \to \infty$
- "**Not learning** from the external distribution" is **always feasible**
- Only uses the **external data** if it improves the objective
- Recovers the presented model from the literature as a **special case**

3 Sets of Numerical Experiments

Artificial experiments

- External data is **artificially** generated
- Direct control over distributions

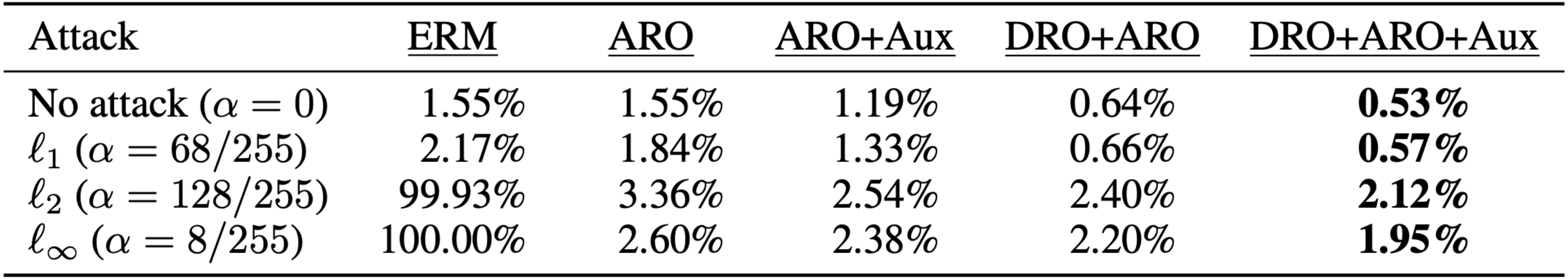
UCI experiments

- Most popular UCI classification datasets
- External data via **synthetic** data generation

MNIST experiments

- Digit recognition (e.g., 3 vs 9)
- External data is digits of high school students

MNIST Experiments



Attack	ERM	ARO	ARO+Aux	DRO+ARO	DRO+ARO+Aux
No attack ($\alpha = 0$)	1.55%	1.55%	1.19%	0.64%	**0.53%**
ℓ_1 ($\alpha = 68/255$)	2.17%	1.84%	1.33%	0.66%	**0.57%**
ℓ_2 ($\alpha = 128/255$)	99.93%	3.36%	2.54%	2.40%	**2.12%**
ℓ_∞ ($\alpha = 8/255$)	100.00%	2.60%	2.38%	2.20%	**1.95%**

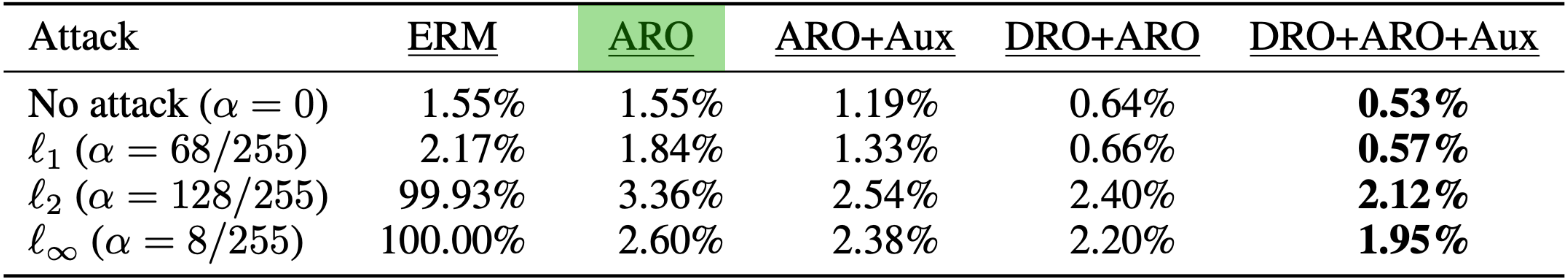
MNIST Experiments

Our DRO models

Attack	ERM	ARO	ARO+Aux	DRO+ARO	DRO+ARO+Aux
No attack ($\alpha = 0$)	1.55%	1.55%	1.19%	0.64%	**0.53%**
ℓ_1 ($\alpha = 68/255$)	2.17%	1.84%	1.33%	0.66%	**0.57%**
ℓ_2 ($\alpha = 128/255$)	99.93%	3.36%	2.54%	2.40%	**2.12%**
ℓ_∞ ($\alpha = 8/255$)	100.00%	2.60%	2.38%	2.20%	**1.95%**

MNIST Experiments

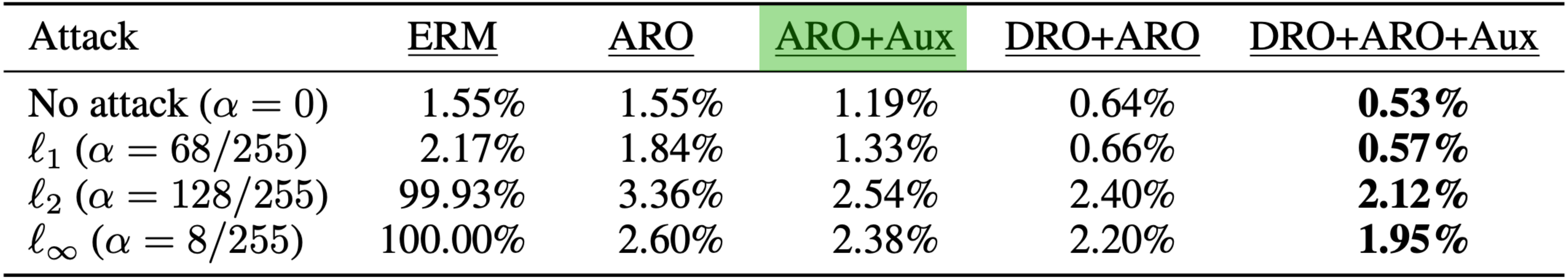
Adversarially robust optimization



Attack	ERM	ARO	ARO+Aux	DRO+ARO	DRO+ARO+Aux
No attack ($\alpha = 0$)	1.55%	1.55%	1.19%	0.64%	**0.53%**
ℓ_1 ($\alpha = 68/255$)	2.17%	1.84%	1.33%	0.66%	**0.57%**
ℓ_2 ($\alpha = 128/255$)	99.93%	3.36%	2.54%	2.40%	**2.12%**
ℓ_∞ ($\alpha = 8/255$)	100.00%	2.60%	2.38%	2.20%	**1.95%**

MNIST Experiments

Adversarially robust optimization (over its mixture with external data)



Attack	ERM	ARO	ARO+Aux	DRO+ARO	DRO+ARO+Aux
No attack ($\alpha = 0$)	1.55%	1.55%	1.19%	0.64%	**0.53%**
ℓ_1 ($\alpha = 68/255$)	2.17%	1.84%	1.33%	0.66%	**0.57%**
ℓ_2 ($\alpha = 128/255$)	99.93%	3.36%	2.54%	2.40%	**2.12%**
ℓ_∞ ($\alpha = 8/255$)	100.00%	2.60%	2.38%	2.20%	**1.95%**

Thank you for listening!

Personal Webpage

IMPERIAL

↓

PRINCETON UNIVERSITY

↓

UCL

Future Work

- Different loss functions
- Intersection of more balls
- Comparison of different relaxation techniques
- Specialised algorithms for $\ell_1, \ell_2, \ell_\infty$ norms in the feature-label metric
- Derive your own algorithm instead of using MOSEK
- Ball around $\mathbb{Q}_{\text{mix}}$ directly

Key Lemma for Tractability

$L(z)$ is convex with $\mathrm{lip}(L) = 1$, $\boldsymbol{\omega}, \boldsymbol{a} \in \mathbb{R}^n$ and $\lambda > 0$. Then:

$$\sup_{\boldsymbol{x} \in \mathbb{R}^n} \{L(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda \|\boldsymbol{a} - \boldsymbol{x}\|_q\}$$

Key Lemma for Tractability

$L(z)$ is convex with $\mathrm{lip}(L) = 1$, $\boldsymbol{\omega}, \boldsymbol{a} \in \mathbb{R}^n$ and $\lambda > 0$. Then:

$$\sup_{\boldsymbol{x} \in \mathbb{R}^n} \{L(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda \|\boldsymbol{a} - \boldsymbol{x}\|_q\}$$

$$= \begin{cases} L(\boldsymbol{a}^\top \boldsymbol{\omega}) & \text{if } \|\boldsymbol{\omega}\|_{q^\star} \leq \lambda \\ +\infty & \text{otherwise}. \end{cases}$$

Key Lemma for Tractability

$L(z)$ is convex with $\operatorname{lip}(L) = 1$, $\boldsymbol{\omega}, \boldsymbol{a} \in \mathbb{R}^n$ and $\lambda > 0$. Then:

$$\sup_{\boldsymbol{x} \in \mathbb{R}^n} \left\{ L(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda \|\boldsymbol{a} - \boldsymbol{x}\|_q \right\}$$

DC Maximization

$$= \begin{cases} L(\boldsymbol{a}^\top \boldsymbol{\omega}) & \text{if } \|\boldsymbol{\omega}\|_{q^\star} \leq \lambda \\ +\infty & \text{otherwise}\,. \end{cases}$$

Key Lemma for Tractability

$L(z)$ is convex with $\operatorname{lip}(L) = 1$, $\omega, a \in \mathbb{R}^n$ and $\lambda > 0$. Then:

$$\sup_{x\in\mathbb{R}^n}\left\{L(\omega^\top x)-\lambda\|a-x\|_q\right\} = \begin{cases} L(a^\top\omega) & \text{if } \|\omega\|_{q^\star} \le \lambda \\ +\infty & \text{otherwise}. \end{cases}$$

Convex fn. in ω

Convex constraint on ω

Non-tractability of Inter-ARO

Key Reason for non-tractability

$L(z)$ is convex with $\mathrm{lip}(L) = 1$, $\boldsymbol{\omega}, \boldsymbol{a}, \widehat{\boldsymbol{a}} \in \mathbb{R}^n$ and λ, $\widehat{\lambda} > 0$. Then:

$$\sup_{\boldsymbol{x} \in \mathbb{R}^n} \{L(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda \|\boldsymbol{a} - \boldsymbol{x}\|_q - \widehat{\lambda} \|\widehat{\boldsymbol{a}} - \boldsymbol{x}\|_q\}$$

Non-tractability of Inter-ARO

Key Reason for non-tractability

$L(z)$ is convex with $\mathrm{lip}(L) = 1, \boldsymbol{\omega}, \boldsymbol{a}, \widehat{\boldsymbol{a}} \in \mathbb{R}^n$ and $\lambda, \widehat{\lambda} > 0$. Then:

$$\sup_{\boldsymbol{x}\in\mathbb{R}^n} \{L(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda \|\boldsymbol{a} - \boldsymbol{x}\|_q - \widehat{\lambda} \|\widehat{\boldsymbol{a}} - \boldsymbol{x}\|_q\}$$

$$=: g(\boldsymbol{\omega})$$

Non-tractability of Inter-ARO

Key Reason for non-tractability

$L(z)$ is convex with $\mathrm{lip}(L) = 1$, $\boldsymbol{\omega}, \boldsymbol{a}, \hat{\boldsymbol{a}} \in \mathbb{R}^n$ and $\lambda, \hat{\lambda} > 0$. Then:

$$\sup_{\boldsymbol{x}\in\mathbb{R}^n} \{L(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda \|\boldsymbol{a} - \boldsymbol{x}\|_q - \hat{\lambda} \|\hat{\boldsymbol{a}} - \boldsymbol{x}\|_q\}$$

$$=: g(\boldsymbol{\omega})$$

We have $\mathcal{O}(N \cdot \widehat{N})$ constraints of type $g(\boldsymbol{\omega}) \leq$ constant

Non-tractability of Inter-ARO

Key Reason for non-tractability

$L(z)$ is convex with $\mathrm{lip}(L) = 1$, $\boldsymbol{\omega}, \boldsymbol{a}, \hat{\boldsymbol{a}} \in \mathbb{R}^n$ and $\lambda, \hat{\lambda} > 0$. Then:

$$\sup_{\boldsymbol{x} \in \mathbb{R}^n} \{L(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda \|\boldsymbol{a} - \boldsymbol{x}\|_q - \hat{\lambda} \|\hat{\boldsymbol{a}} - \boldsymbol{x}\|_q\}$$

DC Maximization

Non-tractability of Inter-ARO

Key Reason for non-tractability

$L(z)$ is convex with $\mathrm{lip}(L) = 1, \boldsymbol{\omega}, \boldsymbol{a}, \widehat{\boldsymbol{a}} \in \mathbb{R}^n$ and $\lambda, \widehat{\lambda} > 0$. Then:

$$\sup_{\boldsymbol{x}\in\mathbb{R}^n} \{L(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda\|\boldsymbol{a} - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{a}} - \boldsymbol{x}\|_q\}$$

$$= \sup_{\theta\in\mathrm{dom}(L^\star)} -L^\star(\theta) + \theta\cdot\boldsymbol{\omega}^\top\boldsymbol{a} + \inf_{z\in\mathbb{R}^n} \{\theta\cdot z^\top(\widehat{\boldsymbol{a}} - \boldsymbol{a}) \ : \ |\theta|\cdot\|\boldsymbol{\omega} - z\|_{q^\star} \le \lambda, \ |\theta|\cdot\|z\|_{q^\star} \le \widehat{\lambda}\}$$

Non-tractability of Inter-ARO

Key Reason for non-tractability

$L(z)$ is convex with $\mathrm{lip}(L) = 1$, $\boldsymbol{\omega}, \boldsymbol{a}, \hat{\boldsymbol{a}} \in \mathbb{R}^n$ and λ, $\hat{\lambda} > 0$. Then:

$$\sup_{\boldsymbol{x}\in\mathbb{R}^n} \{L(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda \|\boldsymbol{a} - \boldsymbol{x}\|_q - \hat{\lambda} \|\hat{\boldsymbol{a}} - \boldsymbol{x}\|_q\}$$

$$= \sup_{\theta\in\mathrm{dom}(L^\star)} - L^\star(\theta) + \theta \cdot \boldsymbol{\omega}^\top \boldsymbol{a} + \inf_{z\in\mathbb{R}^n} \{\theta \cdot z^\top (\hat{\boldsymbol{a}} - \boldsymbol{a}) \ : \ |\theta| \cdot \|\boldsymbol{\omega} - z\|_{q^\star} \leq \lambda, \ |\theta| \cdot \|z\|_{q^\star} \leq \hat{\lambda}\}$$

Minimax theorem not applicable

Non-tractability of Inter-ARO

Key Reason for non-tractability

$L(z)$ is convex with $\mathrm{lip}(L) = 1$, $\boldsymbol{\omega}, \boldsymbol{a}, \hat{\boldsymbol{a}} \in \mathbb{R}^n$ and $\lambda, \hat{\lambda} > 0$. Then:

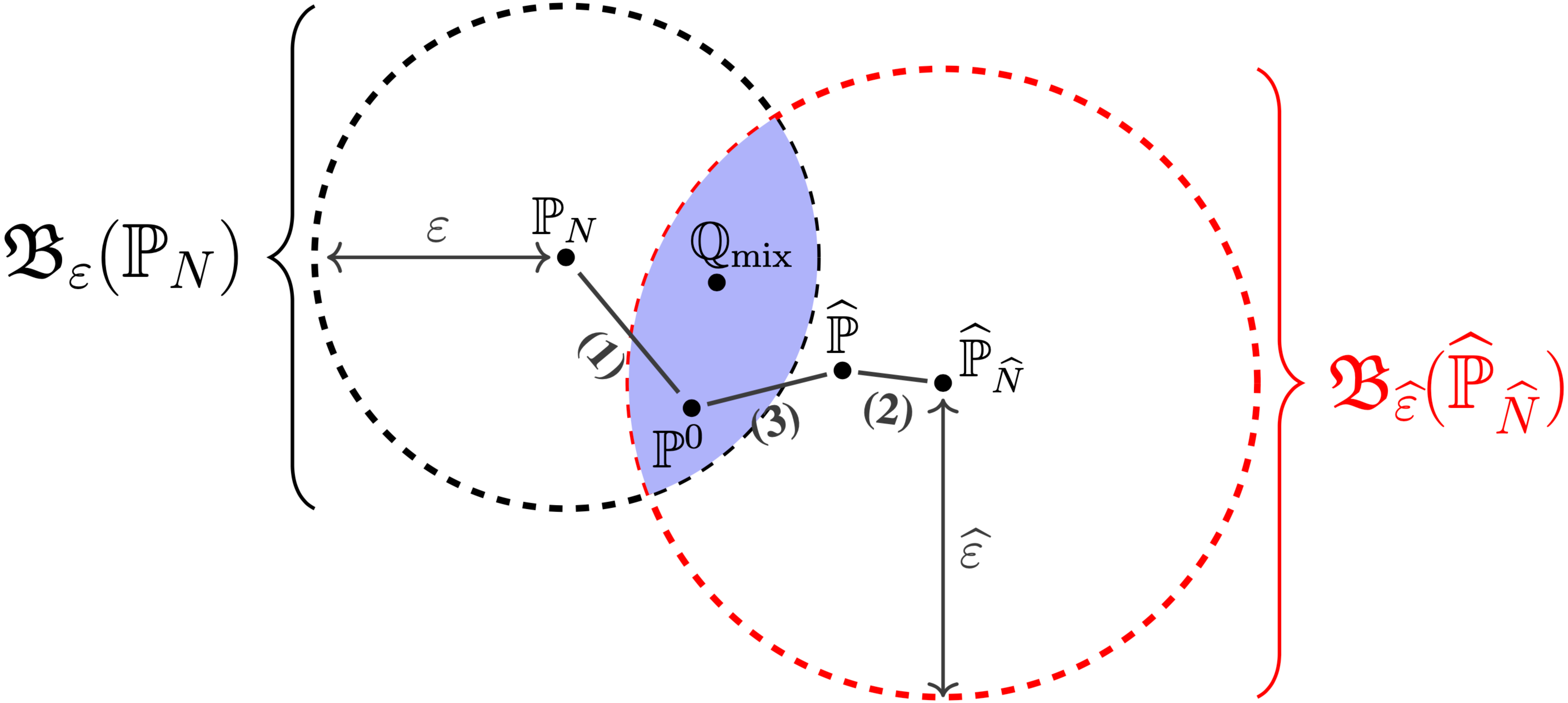
$$\sup_{\boldsymbol{x}\in\mathbb{R}^n} \{L(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda \|\boldsymbol{a} - \boldsymbol{x}\|_q - \hat{\lambda} \|\hat{\boldsymbol{a}} - \boldsymbol{x}\|_q\}$$

$$= \sup_{\theta\in\mathrm{dom}(L^\star)} - L^\star(\theta) + \theta \cdot \boldsymbol{\omega}^\top \boldsymbol{a} + \inf_{z\in\mathbb{R}^n} \{\theta \cdot z^\top (\hat{\boldsymbol{a}} - \boldsymbol{a}) \ : \ |\theta| \cdot \|\boldsymbol{\omega} - z\|_{q^\star} \le \lambda, \ |\theta| \cdot \|z\|_{q^\star} \le \hat{\lambda}\}$$

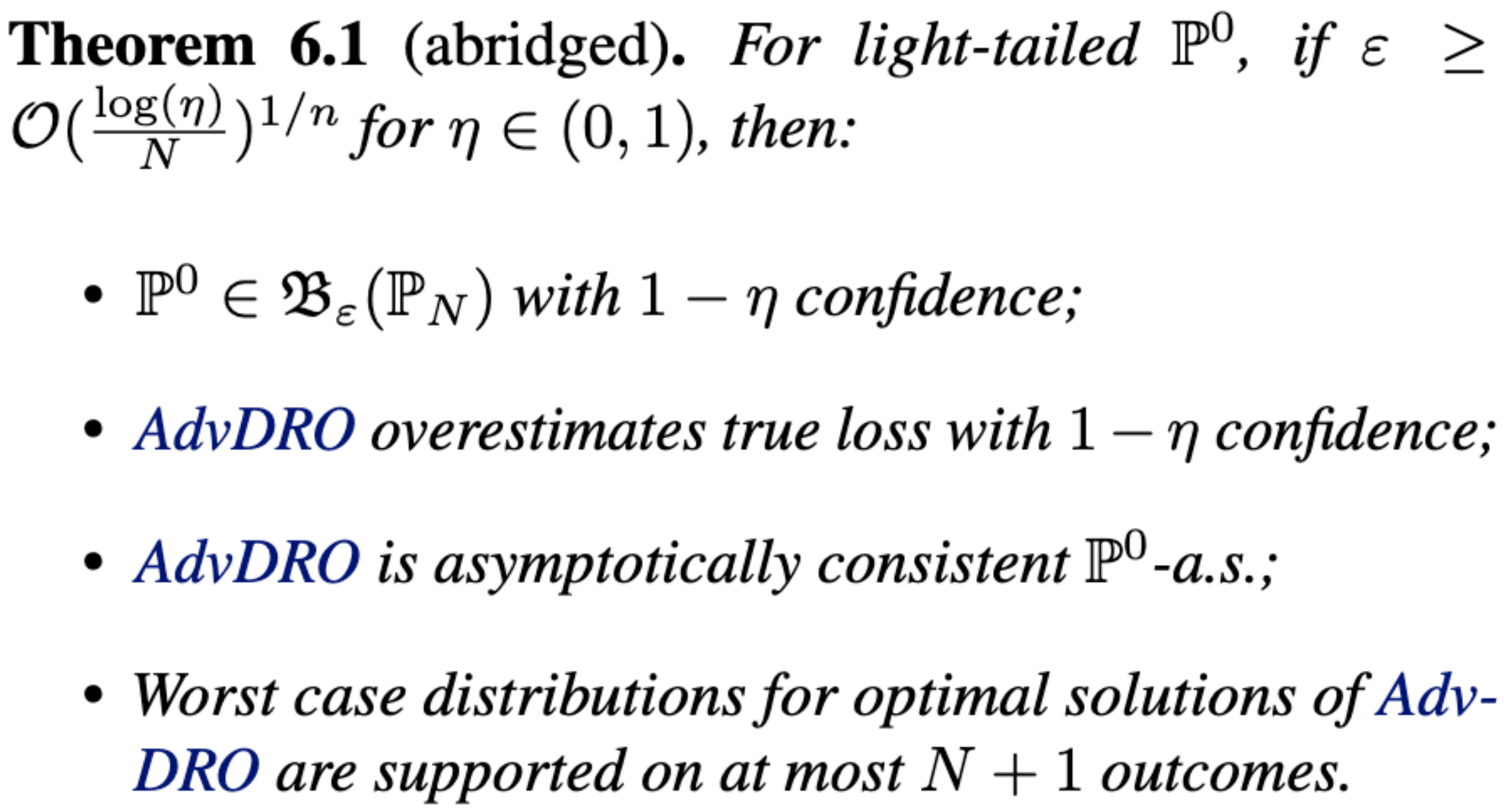
$\forall \theta$

$\exists z$

Can be viewed as an adjustable RO constraint


$\mathfrak{B}_{\varepsilon}(\mathbb{P}_N)$
ε
$\mathbb{P}_N$
$\mathbb{Q}_{\text{mix}}$
(1)
$\widehat{\mathbb{P}}$
$\widehat{\mathbb{P}}_{\widehat{N}}$
$\mathbb{P}^0$
(3)
(2)
$\widehat{\varepsilon}$
$\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$

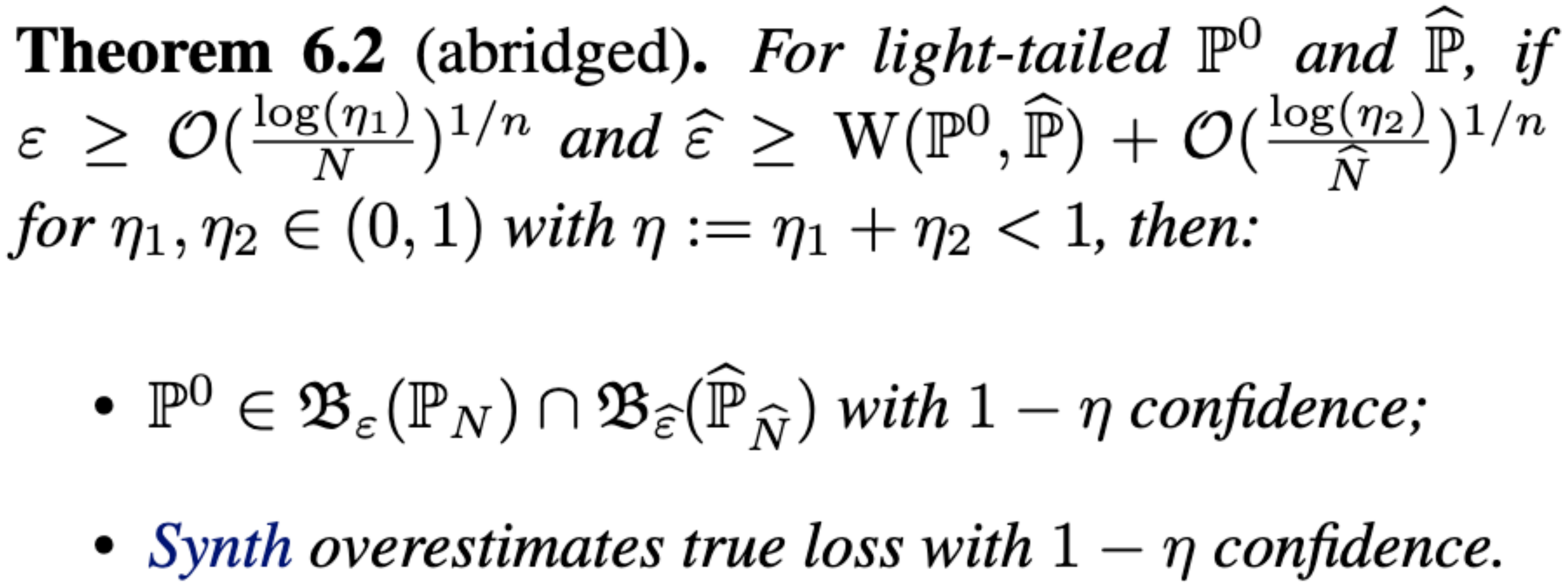
Case 1: Fix $\widehat{\varepsilon} \to \infty$



Theorem 6.1 (abridged)**.** *For light-tailed* $\mathbb{P}^0$, *if* $\varepsilon \geq \mathcal{O}(\frac{\log(\eta)}{N})^{1/n}$ *for* $\eta \in (0,1)$, *then:*

- $\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$ *with* $1-\eta$ *confidence;*
- *AdvDRO overestimates true loss with* $1-\eta$ *confidence;*
- *AdvDRO is asymptotically consistent* $\mathbb{P}^0$*-a.s.;*
- *Worst case distributions for optimal solutions of AdvDRO are supported on at most* $N+1$ *outcomes.*

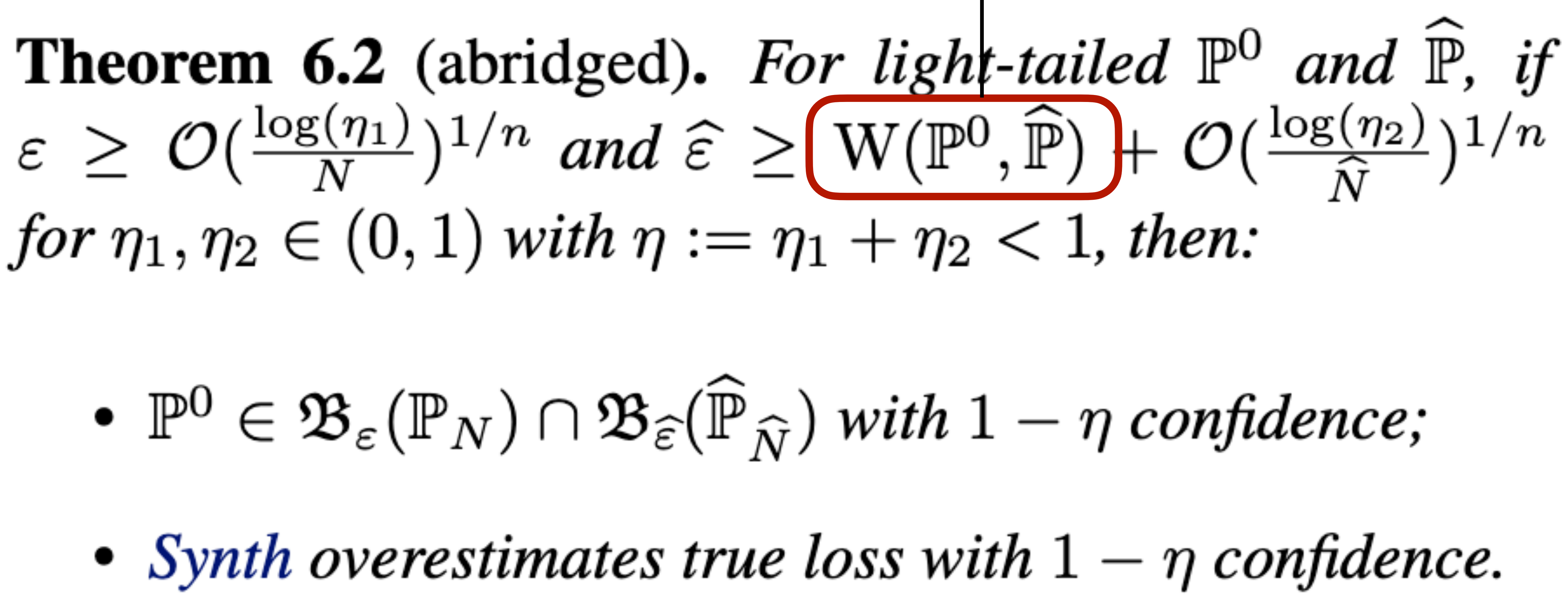
Case 2: Simultaneous



Theorem 6.2 (abridged)**.** *For light-tailed* $\mathbb{P}^0$ *and* $\widehat{\mathbb{P}}$*, if* $\varepsilon \geq \mathcal{O}(\frac{\log(\eta_1)}{N})^{1/n}$ *and* $\widehat{\varepsilon} \geq \mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}}) + \mathcal{O}(\frac{\log(\eta_2)}{\widehat{N}})^{1/n}$ *for* $\eta_1, \eta_2 \in (0,1)$ *with* $\eta := \eta_1 + \eta_2 < 1$*, then:*

- $\mathbb{P}^0 \in \mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ *with* $1 - \eta$ *confidence;*
- *Synth overestimates true loss with* $1 - \eta$ *confidence.*

Case 2: Simultaneous



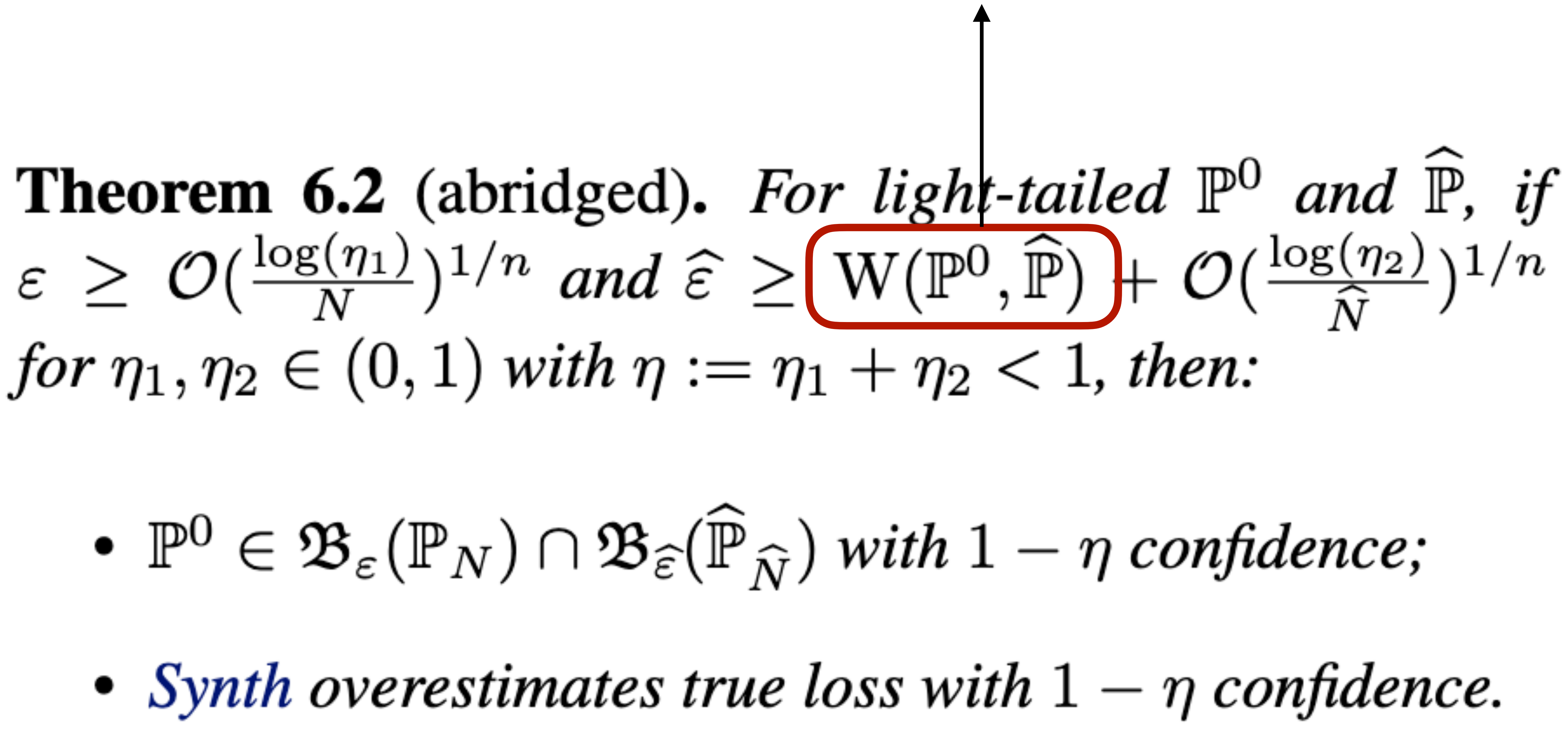
Big assumption!

Theorem 6.2 (abridged). *For light-tailed* $\mathbb{P}^0$ *and* $\widehat{\mathbb{P}}$, *if* $\varepsilon \geq \mathcal{O}(\frac{\log(\eta_1)}{N})^{1/n}$ *and* $\widehat{\varepsilon} \geq \boxed{\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})} + \mathcal{O}(\frac{\log(\eta_2)}{\widehat{N}})^{1/n}$ *for* $\eta_1, \eta_2 \in (0, 1)$ *with* $\eta := \eta_1 + \eta_2 < 1$, *then:*

- $\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ *with* $1 - \eta$ *confidence;*
- *Synth overestimates true loss with* $1 - \eta$ *confidence.*

Case 2: Simultaneous

Big assumption!



Theorem 6.2 (abridged). *For light-tailed* $\mathbb{P}^0$ *and* $\widehat{\mathbb{P}}$*, if* $\varepsilon \geq \mathcal{O}(\frac{\log(\eta_1)}{N})^{1/n}$ *and* $\widehat{\varepsilon} \geq \boxed{\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})} + \mathcal{O}(\frac{\log(\eta_2)}{\widehat{N}})^{1/n}$ *for* $\eta_1, \eta_2 \in (0,1)$ *with* $\eta := \eta_1 + \eta_2 < 1$*, then:*

- $\mathbb{P}^0 \in \mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ *with* $1 - \eta$ *confidence;*
- *Synth overestimates true loss with* $1 - \eta$ *confidence.*

What can be done beyond cross-validation?

1. **Uber vs Lyft (Taskesen et al, 2021)**
2. **Opt-out data with differential privacy (Ullman and Vadhan, 2020)**
3. **Wasserstein GANs comes with guarantees on** $\mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}})$