Regularization for Adversarial Robust Learning

Jie Wang
Georgia Institute of Technology

2024 PURDUE OPERATIONS CONFERENCE
Student Session 02

Collaborators

Rui Gao

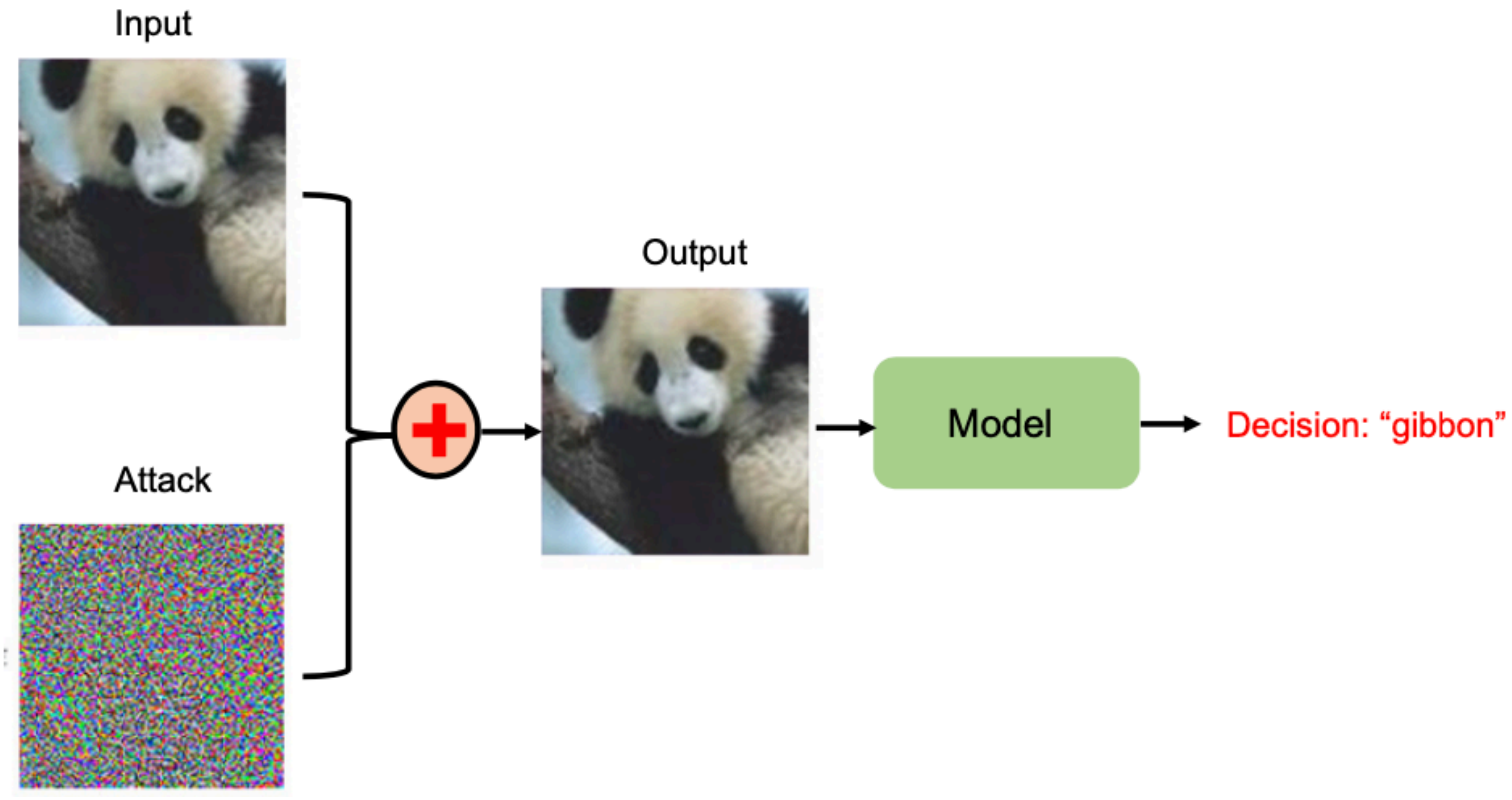
The University of Texas at Austin

Yao Xie

Georgia Institute of Technology

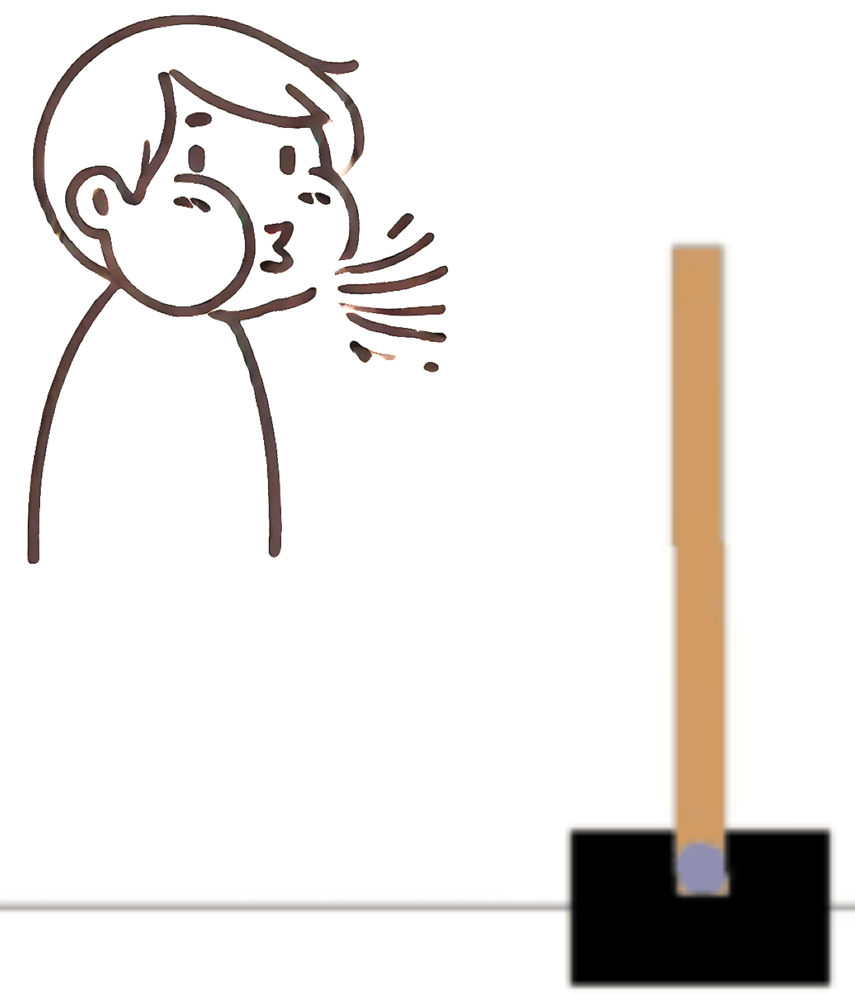
1. Introduction

Motivation [Goodfellow et al. 2015]

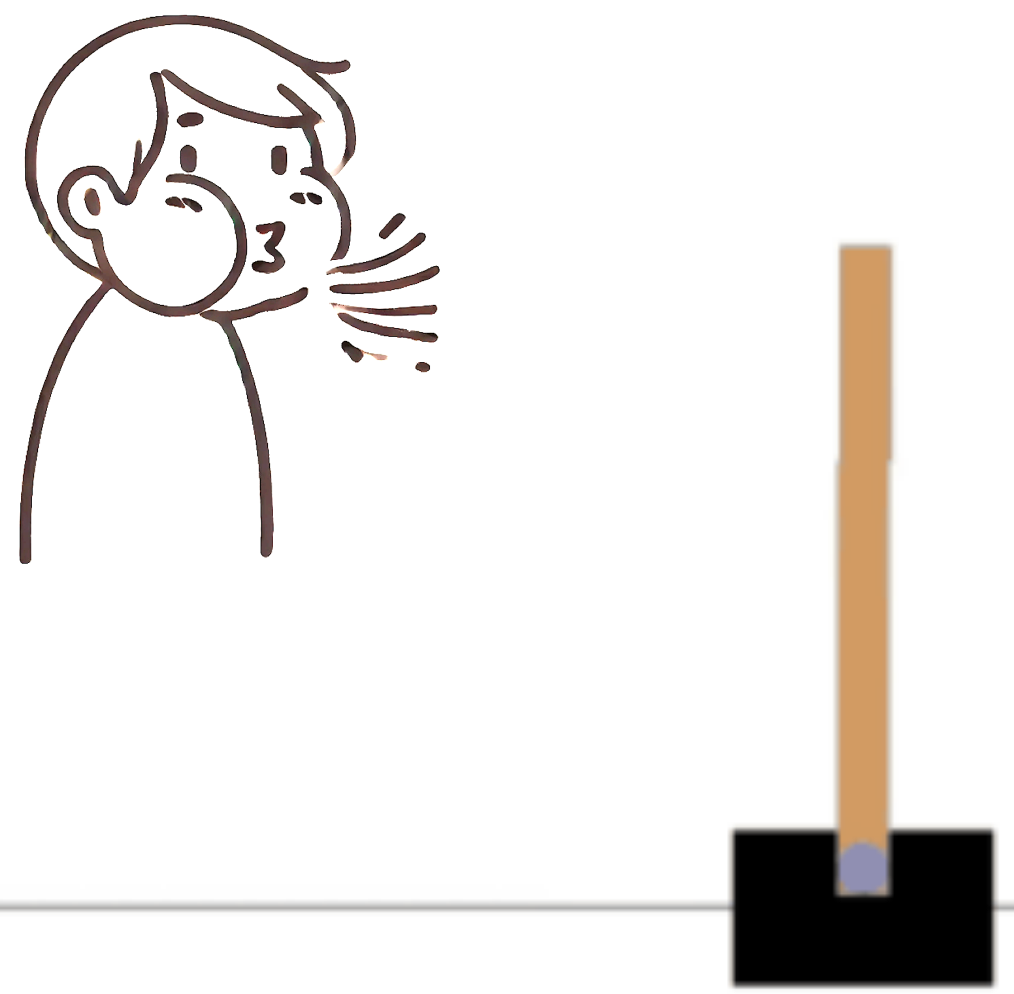


Motivation [Sutton et al. 2018]

Non-robust Algorithm

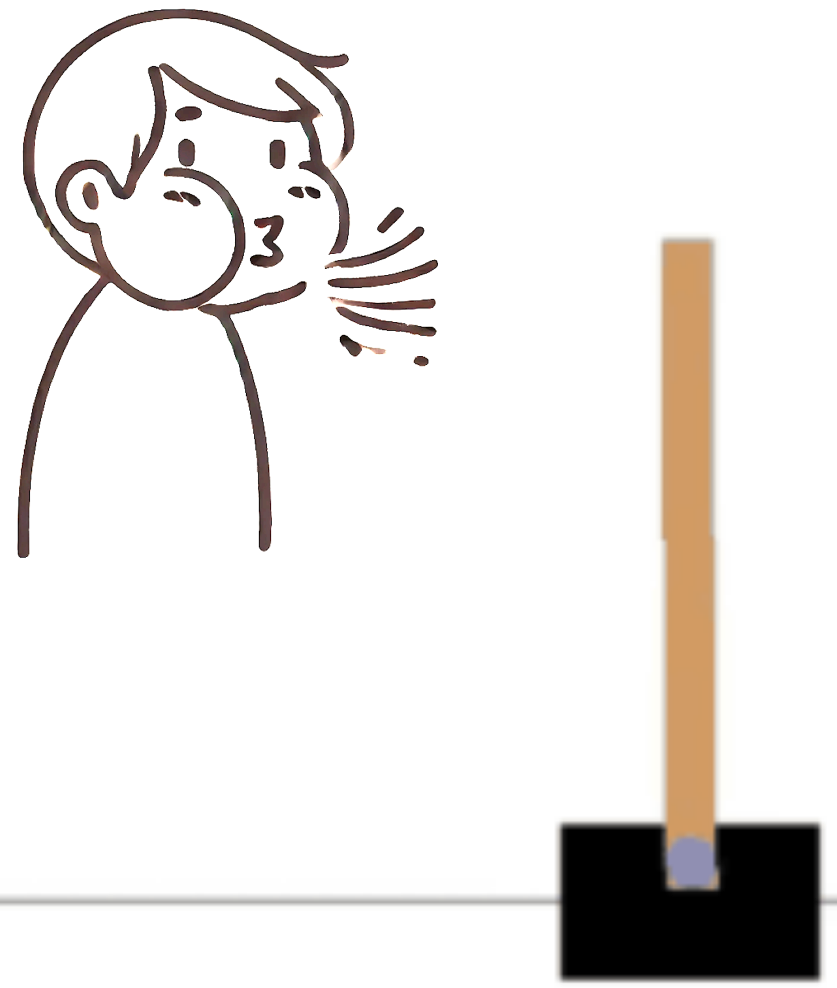


Robust Algorithm

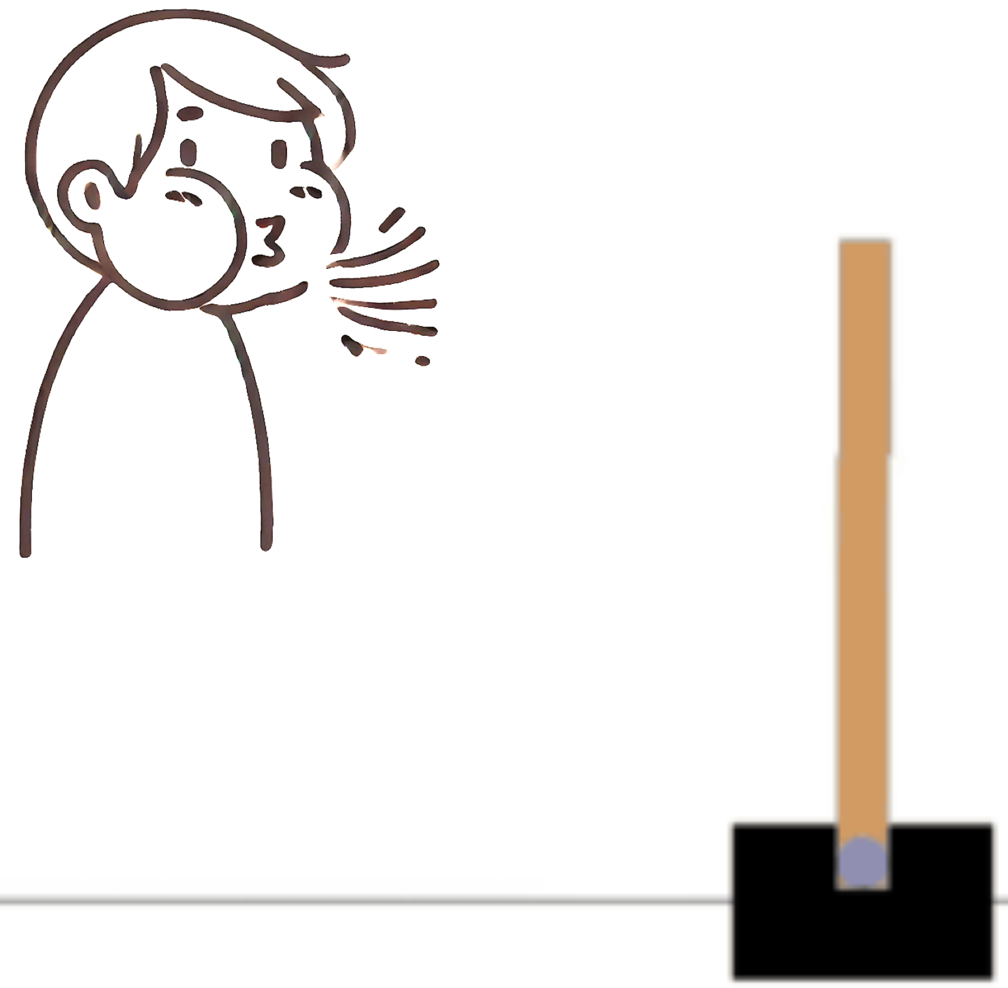


Motivation [Sutton et al. 2018]

Non-robust Algorithm



Robust Algorithm



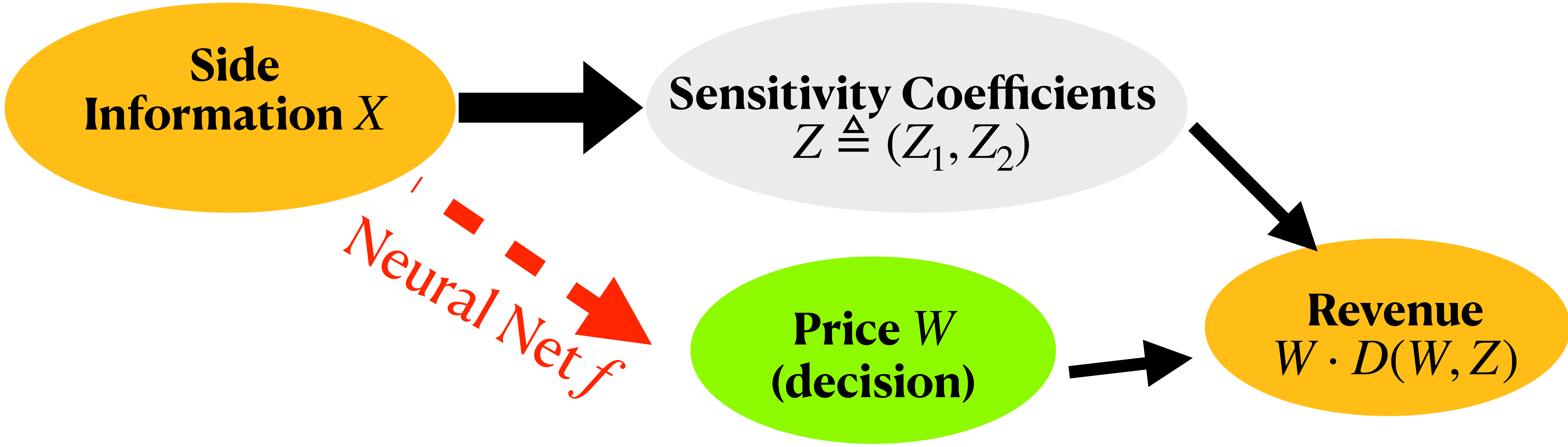
Motivation on Data-Driven Personalized Pricing

- Linear demand model:

$$D(w, z) = z_1 w + z_2$$

Price

Sensitivity coefficients



$$\inf_{f} \mathbb{E}_{(X,Z)\sim\mathbb{P}^0} \left[-f(X) \cdot D(f(X), Z) \right] \qquad \text{(Nominal Problem)}$$

Cons: Distribution Shift on side information!

Adversarial Risk Minimization

Loss Function

$$\min_{\theta \in \Theta} \left\{ \mathbb{E}_{z \sim \mathbb{P}_n} \left[\sup_{\mathbf{d}(z, z') \leq \rho} \ell(z'; \theta) \right] \right\}$$

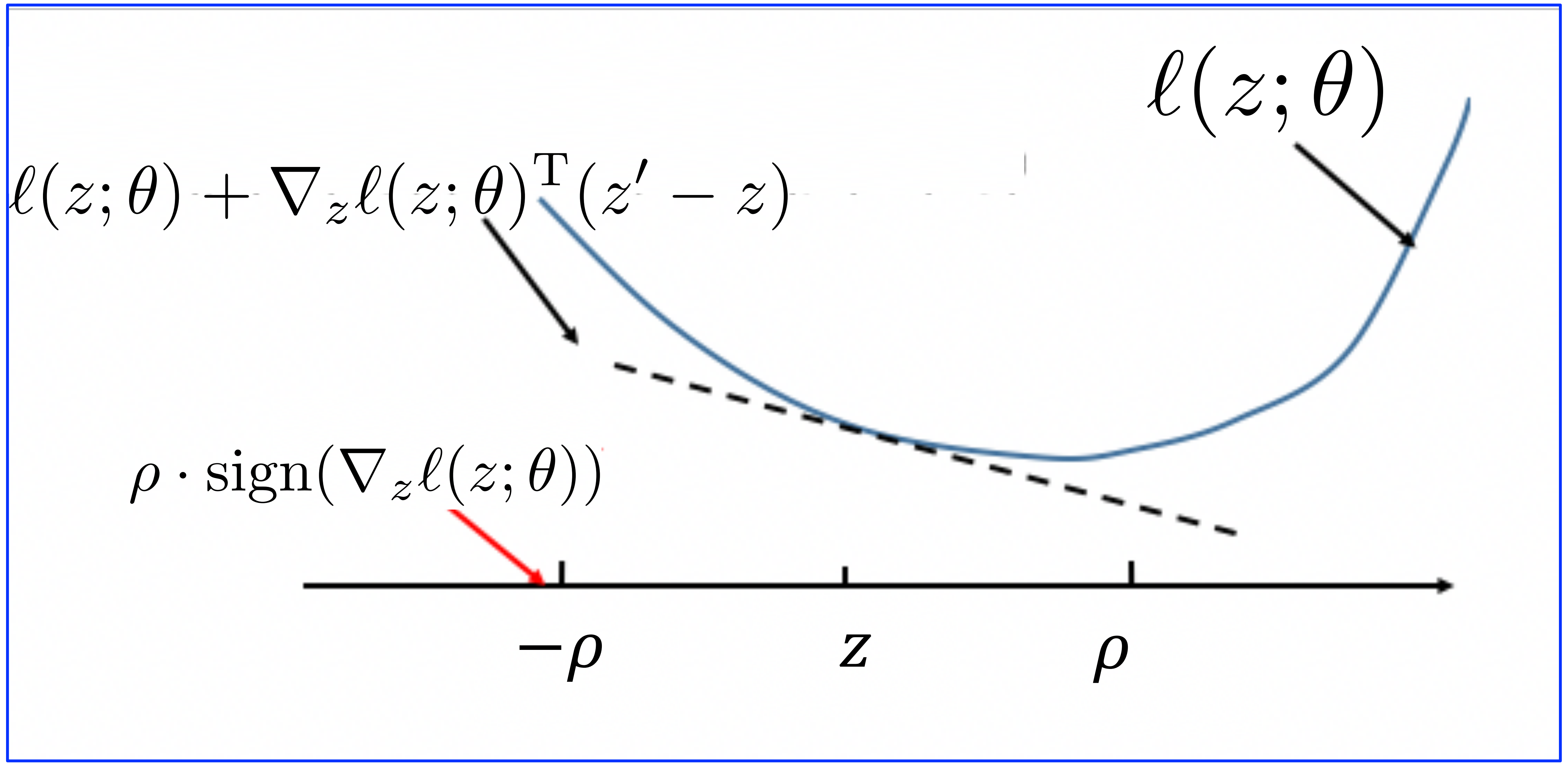
Data (e.g., feature-label pair) following empirical distribution

Perturbation Constriants

Baseline Approach: Linearizing Objective Function

$$\min_{\theta\in\Theta} \left\{ \mathbb{E}_{z\sim\mathbb{P}_n} \left[\sup_{\mathbf{d}(z,z')\leq\rho} \ell(z';\theta) \right] \right\}$$

- **Fast Gradient Method (FGM)** [Goodfellow et al. 2015]



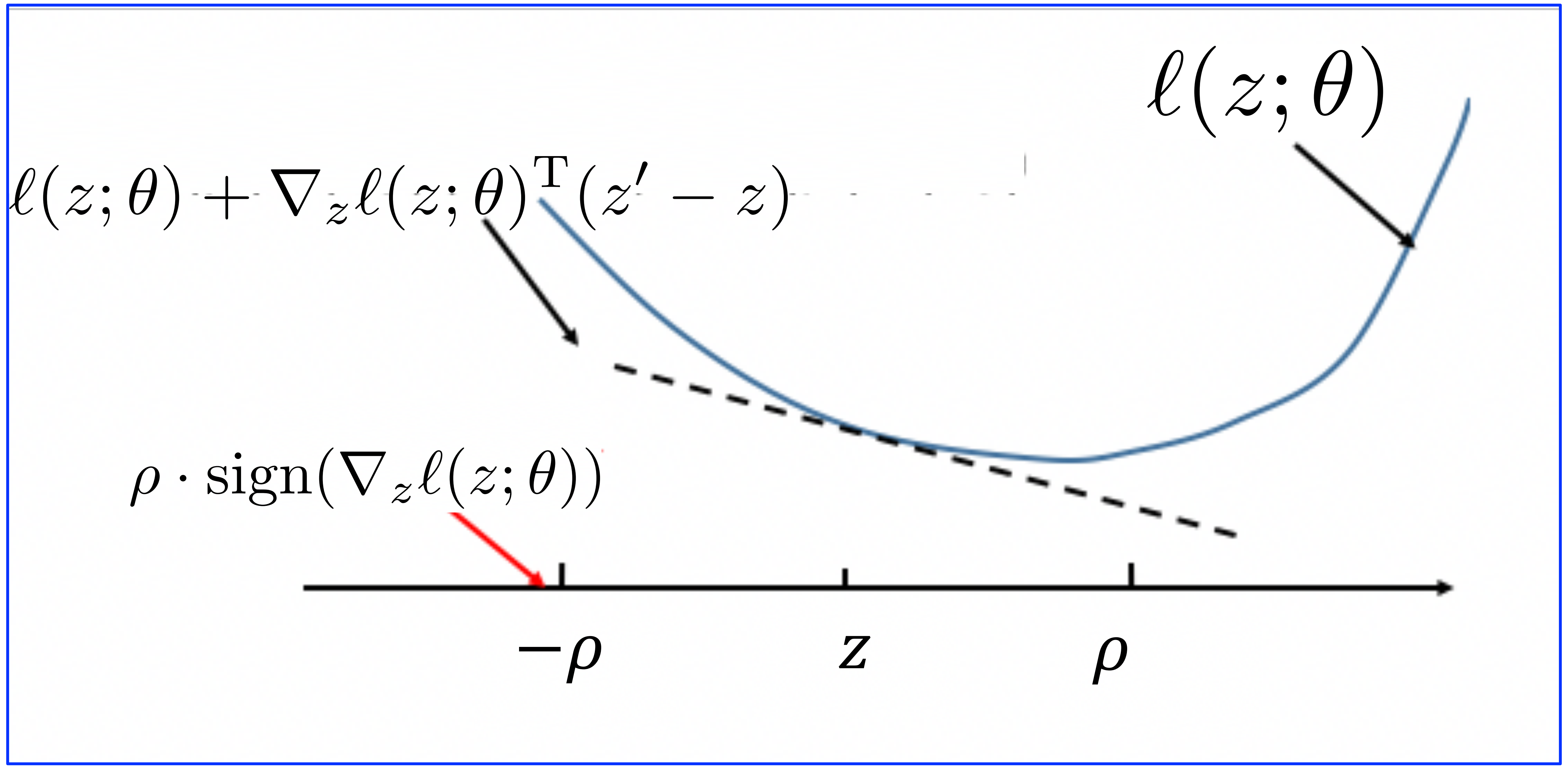
- $z' \approx \underset{\|z-z'\|_\infty\leq\rho}{\arg\max} \left[\ell(z;\theta) + \nabla_z \ell(z;\theta)^{\mathrm{T}}(z'-z) \right]$

  $= z + \rho \cdot \mathrm{sign}(\nabla_z \ell(z;\theta))$

Baseline Approach: Linearizing Objective Function

$$\min_{\theta\in\Theta} \left\{ \mathbb{E}_{z\sim\mathbb{P}_n} \left[\sup_{\mathbf{d}(z,z')\leq\rho} \ell(z';\theta) \right] \right\}$$

- **Iterative Fast Gradient Method (FGM)** [Goodfellow et al. 2015]



- $z^0 = z$

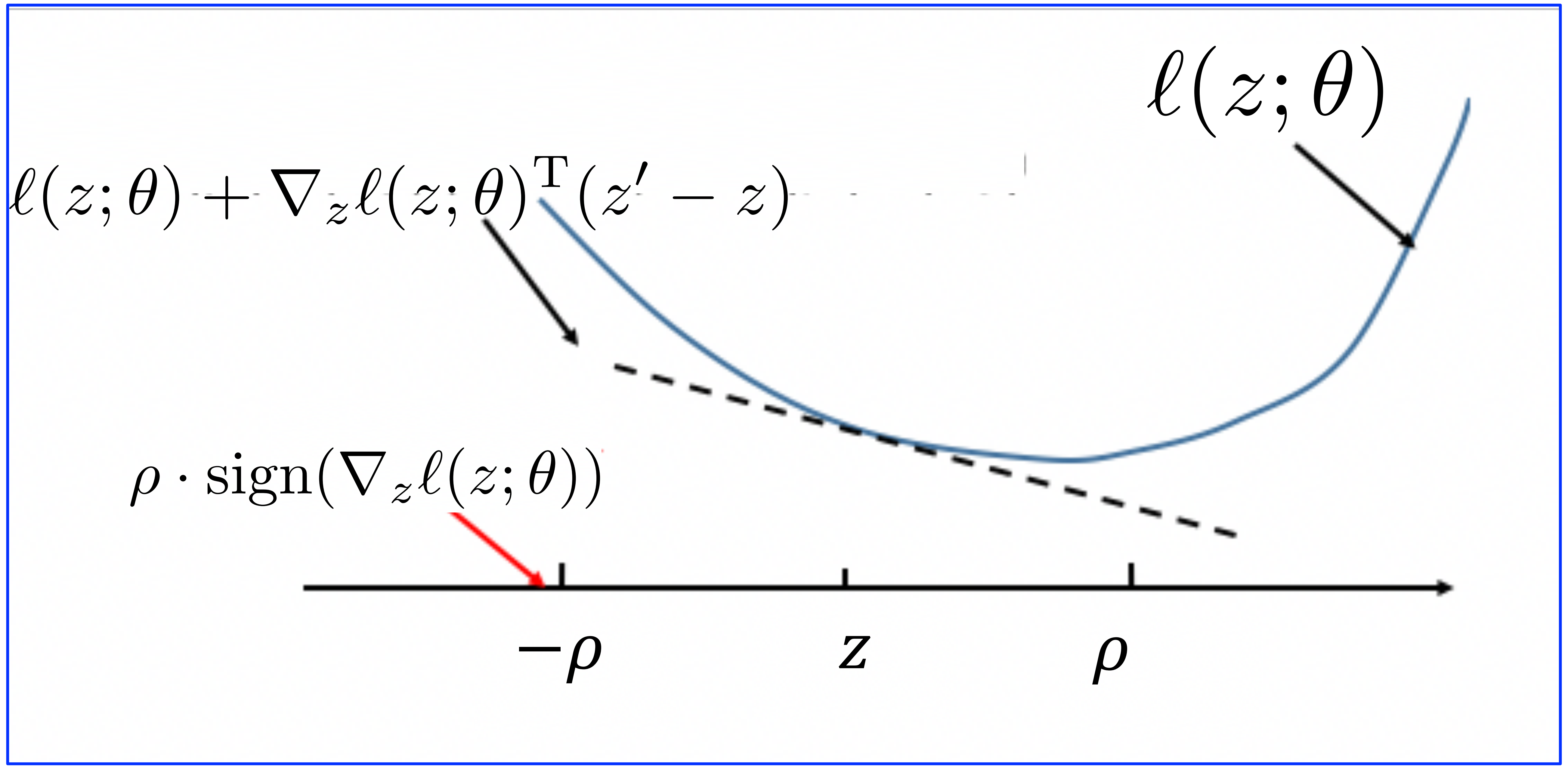
  $z^k = z^{k-1} + \alpha\cdot\mathrm{sign}(\nabla_z\ell(z^{k-1};\theta)),$

  $k = 1,\ldots,T-1, \alpha = \dfrac{\rho}{T}$

Baseline Approach: Linearizing Objective Function

$$\min_{\theta\in\Theta} \left\{ \mathbb{E}_{z\sim\mathbb{P}_n} \left[\sup_{\mathbf{d}(z,z')\leq\rho} \ell(z';\theta) \right] \right\}$$

- **Iterative Fast Gradient Method (FGM)** [Goodfellow et al. 2015]



- $z^0 = z$

$$z^k = z^{k-1} + \alpha\cdot\mathrm{sign}(\nabla_z\ell(z^{k-1};\theta)),$$
$$k = 1,\ldots,T-1, \alpha = \frac{\rho}{T}$$

Cons: Optimization error is non-negligible for large ρ!

Connections with Wasserstein Robust Optimization

$$\min_{\theta \in \Theta} \left\{ \mathbb{E}_{z \sim \mathbb{P}_n} \left[\sup_{\mathbf{d}(z, z') \leq \rho} \ell(z'; \theta) \right] \right\}$$

Connections with Wasserstein Robust Optimization

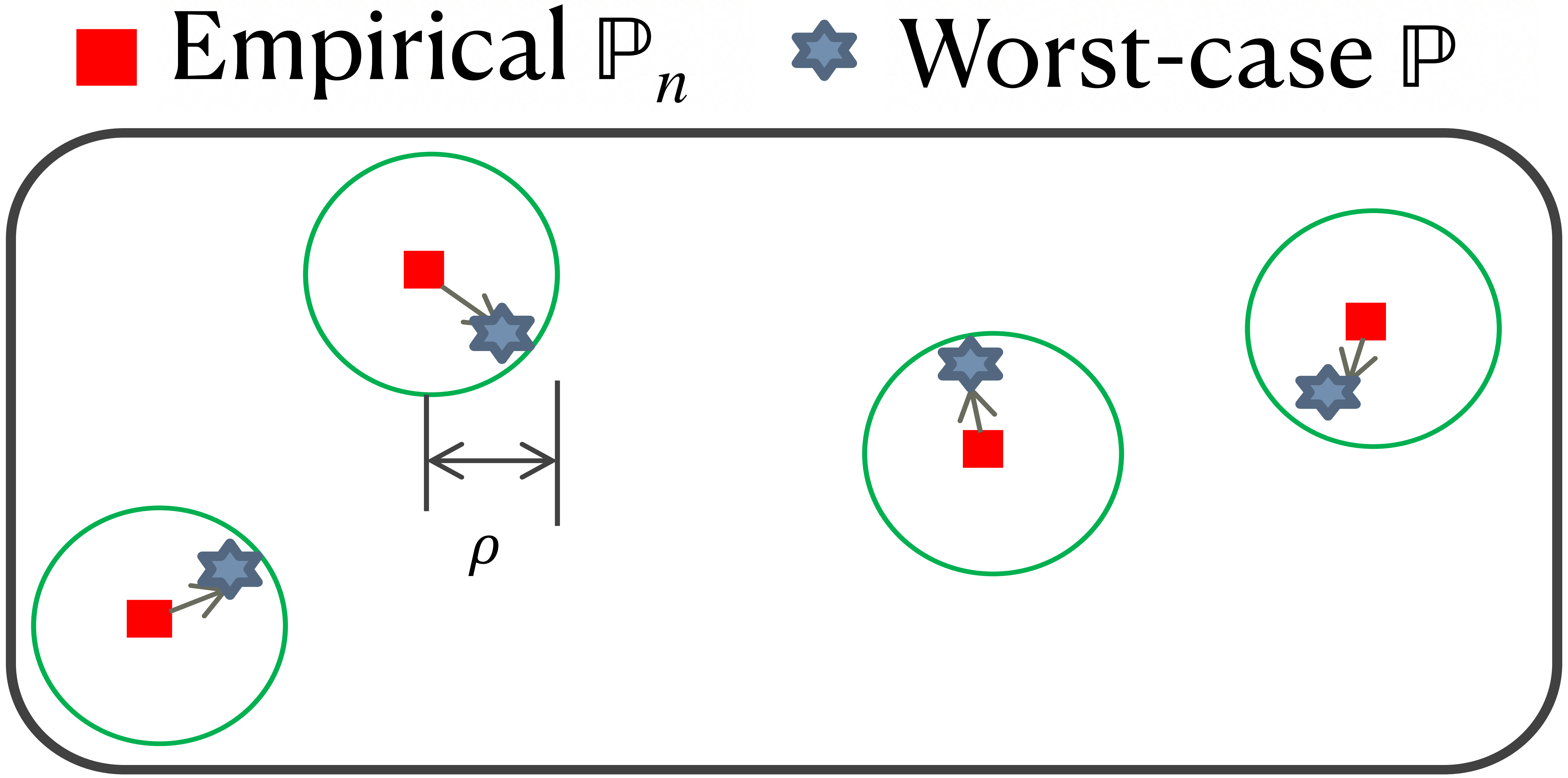
$$\min_{\theta\in\Theta} \left\{ \sup_{\mathbb{P}:\ \mathcal{W}_\infty(\mathbb{P},\mathbb{P}_n)\leq\rho} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] \right\}$$

$$\mathcal{W}_\infty(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left\{ \gamma\text{-esssup } \mathbf{d}(x,y) \right\}$$

Connections with Wasserstein Robust Optimization

$$\min_{\theta\in\Theta}\left\{\sup_{\mathbb{P}:\ \mathcal{W}_\infty(\mathbb{P},\mathbb{P}_n)\le\rho}\mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]\right\}$$

$$\mathcal{W}_\infty(\mathbb{P},\mathbb{Q})=\inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})}\left\{\gamma\text{-esssup }\mathbf{d}(x,y)\right\}$$



Literature Review

$$\min_{\theta\in\Theta}\left\{\sup_{\mathbb{P}:\ \mathcal{W}_\infty(\mathbb{P},\mathbb{P}_n)\le\rho}\mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]\right\}$$

p-Wasserstein DRO Approximation

[Sinha, Namkoong, Volpi, Duchi, 2020]

$$\min_{\theta\in\Theta}\left\{\sup_{\mathbb{P}:\ \mathcal{W}_p(\mathbb{P},\mathbb{P}_n)\le\rho}\mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]\right\}$$

$$\mathcal{W}_p(\mathbb{P},\mathbb{Q})=\inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})}\left\{\left(\mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|^p]\right)^{1/p}\right\}$$

Literature Review

$$\min_{\theta\in\Theta} \left\{ \sup_{\mathbb{P}:\ \mathcal{W}_\infty(\mathbb{P},\mathbb{P}_n)\leq\rho} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] \right\}$$

p-Wasserstein DRO Approximation

[Sinha, Namkoong, Volpi, Duchi, 2020]

$$\min_{\theta\in\Theta,\lambda\geq 0} \left\{ \lambda\rho^p + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\sup_z \left\{ \ell(z;\theta) - \lambda\|z-x\|^p \right\} \right] \right\}$$

- Easy to optimize for large choice of λ

Literature Review

$$\min_{\theta\in\Theta}\left\{\sup_{\mathbb{P}:\ \mathcal{W}_\infty(\mathbb{P},\mathbb{P}_n)\le\rho}\mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]\right\}$$

p-Wasserstein DRO

$$\min_{\theta\in\Theta,\lambda\ge 0}\left\{\lambda\rho^p+\mathbb{E}_{x\sim\mathbb{P}_n}\left[\sup_z\left\{\ell(z;\theta)-\lambda\|z-x\|^p\right\}\right]\right\}$$

Entropic Regularized p-Wasserstein DRO Approximation

[Wang, Gao, Xie, 2021]

$$\min_{\theta\in\Theta}\left\{\sup_{\mathbb{P}:\ \mathcal{S}_{p,\epsilon}(\mathbb{P},\mathbb{P}_n)\le\rho}\mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]\right\}$$

$$\mathcal{S}_{p,\epsilon}(\mathbb{P},\mathbb{P}_n)=\inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{P}_n)}\left\{\mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|^p]+\epsilon\mathbb{E}_{(x,y)\sim\gamma}\left[\log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}x\mathrm{d}\gamma(y)}\right)\right]\right\}$$

Literature Review

$$\min_{\theta\in\Theta} \left\{ \sup_{\mathbb{P}:\ \mathcal{W}_\infty(\mathbb{P},\mathbb{P}_n)\le\rho} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] \right\}$$

***p*-Wasserstein DRO**

$$\min_{\theta\in\Theta,\lambda\ge 0} \left\{ \lambda\rho^p + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\sup_z \left\{\ell(z;\theta) - \lambda\|z-x\|^p\right\}\right] \right\}$$

Entropic Regularized *p*-Wasserstein DRO Approximation

[Wang, Gao, Xie, 2021]

$$\min_{\theta\in\Theta,\lambda\ge 0} \left\{ \lambda\bar{\rho} + \mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon \log \mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\right] \right\}$$

$$\frac{\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)}{\mathrm{d}z} \propto e^{-\|z-x\|^p/\epsilon}$$

1. Entropic regularization brings **computational benefits**
2. Entropic regularization introduces **absolutely continuous worst-case distributions**

Literature Review

$$\min_{\theta\in\Theta}\left\{\sup_{\mathbb{P}:\ \mathcal{W}_\infty(\mathbb{P},\mathbb{P}_n)\le\rho}\mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]\right\}$$

How about adding regularization directly?

***p*-Wasserstein DRO**

$$\min_{\theta\in\Theta,\lambda\ge 0}\left\{\lambda\rho^p+\mathbb{E}_{x\sim\mathbb{P}_n}\left[\sup_z\left\{\ell(z;\theta)-\lambda\|z-x\|^p\right\}\right]\right\}$$

Entropic Regularized *p*-Wasserstein DRO Approximation

[Wang, Gao, Xie, 2021]

$$\min_{\theta\in\Theta,\lambda\ge 0}\left\{\lambda\bar{\rho}+\mathbb{E}_{x\sim\mathbb{P}_n}\left[\lambda\epsilon\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}\left[e^{f(z)/(\lambda\epsilon)}\right]\right]\right\}$$

$$\frac{\mathrm{d}\mathbb{Q}_{x,\epsilon}(z)}{\mathrm{d}z}\propto e^{-\|z-x\|^p/\epsilon}$$

1. Entropic regularization brings **computational benefits**
2. Entropic regularization introduces **absolutely continuous worst-case distributions**

Regularized Adversarial Robust Learning

$$\min_{\theta\in\Theta} \ \sup_{\mathbb{P},\gamma} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] \colon \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \mathbf{d}(x,y) \le \rho \end{array} \right\}$$

Regularized Adversarial Robust Learning

$$\min_{\theta\in\Theta} \sup_{\mathbb{P},\gamma} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] - \eta \mathbb{D}_f(\gamma,\gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \mathbf{d}(x,y) \le \rho \end{array} \right\}$$

- **f-divergence (e.g., KL or χ^2):**

$$\mathbb{D}_f(\gamma,\gamma_0) = \int f\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\gamma_0}\right) \mathrm{d}\gamma_0$$

- **Reference transport γ_0 is uniform:**

For each $z \in \mathrm{supp}(\mathbb{P}_n)$, $\gamma_0(\,\cdot \mid z)$ is uniform on $\mathbb{B}_\rho(z)$

$$\gamma_0(z,z') = \mathbb{P}_n(z) \cdot \mathbb{Q}_{z,\rho}(z')$$

2. Strong Duality

Strong Dual Reformulation

Under mild conditions, $V_{\text{Primal}}=V_{\text{dual}}$:

$$V_{\text{Primal}} = \sup_{\mathbb{P},\gamma} \left\{ \mathbb{E}_{\mathbb{P}}[\ell(z;\theta)] - \eta \mathbb{D}_f(\gamma, \gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \mathbf{d}(x,y) \le \rho \end{array} \right\}$$

By $\gamma_0(z,z') = \mathbb{P}_n(z) \cdot \mathbb{Q}_{z,\rho}(z')$, $\quad \gamma(z,z') = \mathbb{P}_n(z) \cdot \mathbb{P}_z(z')$,

$$V_{\text{Dual}} = \mathbb{E}_{z\sim\mathbb{P}_n}\left[\sup_{\mathbb{P}_z} \underbrace{\left\{ \mathbb{E}_{z'\sim\mathbb{P}_z}[\ell(z';\theta)] - \eta \mathbb{D}_f(\mathbb{P}_z, \mathbb{Q}_{z,\rho}) \right\}}_{\textbf{Penalized } f\textbf{-divergence DRO}} \right]$$

Strong Dual Reformulation

Under mild conditions, $V_{\text{Primal}} = V_{\text{dual}}$:

$$V_{\text{Primal}} = \sup_{\mathbb{P},\gamma} \left\{ \mathbb{E}_{\mathbb{P}}[\ell(z;\theta)] - \eta \mathbb{D}_f(\gamma, \gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup}\ \mathbf{d}(x, y) \le \rho \end{array} \right\}$$

$$V_{\text{Dual}} = \mathbb{E}_{z\sim\mathbb{P}_n} \left[\inf_{\mu\in\mathbb{R}} \left\{ \mu + \mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[(\eta f)^*(\ell(z';\theta) - \mu)] \right\} \right]$$

Divergence $\mathbb{D}_f(\cdot\,,\cdot)$	**Choice of** $f(x)$	V_{Dual}
KL-Divergence	$x \log x - x + 1$	$\mathbb{E}_{\mathbb{P}_n}\left[\eta \log \mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[e^{\ell(z';\theta)/\eta}]\right]$
χ²-Divergence	$\frac{1}{2}(x^2 - 1)$	$\mathbb{E}_{z\sim\mathbb{P}_n}\left[\inf_{\mu\in\mathbb{R}} \left\{\frac{1}{2\eta}\mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[\ell(z';\theta) - \mu]_+^2 + \frac{\eta}{2} + \mu\right\}\right]$

Strong Dual Reformulation

Divergence $\mathbb{D}_f(\cdot,\cdot)$	**Choice of** $f(x)$	V_{Dual}
KL-Divergence	$x \log x - x + 1$	$\mathbb{E}_{\mathbb{P}_n}\left[\eta \log \mathbb{E}_{z' \sim \mathbb{Q}_{z,\rho}}[e^{\ell(z';\theta)/\eta}]\right]$
χ²-Divergence	$\frac{1}{2}(x^2 - 1)$	$\mathbb{E}_{z \sim \mathbb{P}_n}\left[\inf_{\mu \in \mathbb{R}} \left\{\frac{1}{2\eta}\mathbb{E}_{z' \sim \mathbb{Q}_{z,\rho}}[\ell(z';\theta) - \mu]_+^2 + \frac{\eta}{2} + \mu\right\}\right]$

Strong Dual Reformulation

Divergence $\mathbb{D}_f(\cdot,\cdot)$	Choice of $f(x)$	V_{Dual}
KL-Divergence	$x \log x - x + 1$	$\mathbb{E}_{\mathbb{P}_n}\left[\eta \log \mathbb{E}_{z' \sim \mathbb{Q}_{z,\rho}}[e^{\ell(z';\theta)/\eta}]\right]$
χ²-Divergence	$\frac{1}{2}(x^2-1)$	$\mathbb{E}_{z\sim\mathbb{P}_n}\left[\inf_{\mu\in\mathbb{R}} \left\{\frac{1}{2\eta}\mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[\ell(z';\theta)-\mu]_+^2 + \frac{\eta}{2} + \mu\right\}\right]$

Strong dual for un-regularized case (η=0) [Gao et al., 2022]:

$$V_{\text{Primal}} = \sup_{\mathbb{P},\gamma} \left\{ \mathbb{E}_{\mathbb{P}}[\ell(z;\theta)] : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \mathbf{d}(x,y) \le \rho \end{array} \right\}$$

$$V_{\text{Dual}} = \mathbb{E}_{z\sim\mathbb{P}_n}\left[\sup_{z':\ \mathbf{d}(z,z')\le\rho} \ell(z';\theta) \right]$$

Strong Dual Reformulation

Divergence $\mathbb{D}_f(\cdot,\cdot)$	Choice of $f(x)$	V_{Dual}
KL-Divergence	$x\log x - x + 1$	$\mathbb{E}_{\mathbb{P}_n}\left[\eta \log \mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[e^{\ell(z';\theta)/\eta}]\right]$
χ²-Divergence	$\frac{1}{2}(x^2-1)$	$\mathbb{E}_{z\sim\mathbb{P}_n}\left[\inf_{\mu\in\mathbb{R}}\left\{\frac{1}{2\eta}\mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[\ell(z';\theta)-\mu]_+^2+\frac{\eta}{2}+\mu\right\}\right]$

Strong dual for un-regularized case (η=0) [Gao et al., 2022]:

$$V_{\text{Primal}} = \sup_{\mathbb{P},\gamma}\left\{\mathbb{E}_{\mathbb{P}}[\ell(z;\theta)] : \begin{array}{l}\gamma\in\Gamma(\mathbb{P}_n,\mathbb{P})\\ \gamma\text{-esssup}\ \mathbf{d}(x,y)\le\rho\end{array}\right\}$$

$$V_{\text{Dual}} = \mathbb{E}_{z\sim\mathbb{P}_n}\left[\sup_{z':\ \mathbf{d}(z,z')\le\rho}\ell(z';\theta)\right]$$

Laplace's Method

Extension of Laplace's Method

Divergence $\mathbb{D}_f(\cdot,\cdot)$	**Choice of $f(x)$**	V_{Dual}
KL-Divergence	$x\log x - x + 1$	$\mathbb{E}_{\mathbb{P}_n}\left[\eta\log\mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[e^{\ell(z';\theta)/\eta}]\right]$
χ²-Divergence	$\frac{1}{2}(x^2-1)$	$\mathbb{E}_{z\sim\mathbb{P}_n}\left[\inf_{\mu\in\mathbb{R}}\left\{\frac{1}{2\eta}\mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[\ell(z';\theta)-\mu]_+^2+\frac{\eta}{2}+\mu\right\}\right]$

Example: $f(x)=\mathbb{I}\{0\le x\le\alpha^{-1}\},\ V_{\text{Dual}}=\mathbb{E}_{z\sim\mathbb{P}_n}\left[\mathtt{AV@R}_{\alpha,\mathbb{Q}_{z,\rho}}(\ell(\cdot;\theta))\right]$

Extension of Laplace's Method

Divergence $\mathbb{D}_f(\cdot,\cdot)$	Choice of $f(x)$	V_{Dual}
KL-Divergence	$x\log x - x + 1$	$\mathbb{E}_{\mathbb{P}_n}\left[\eta \log \mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[e^{\ell(z';\theta)/\eta}]\right]$
χ²-Divergence	$\frac{1}{2}(x^2-1)$	$\mathbb{E}_{z\sim\mathbb{P}_n}\left[\inf_{\mu\in\mathbb{R}}\left\{\frac{1}{2\eta}\mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[\ell(z';\theta)-\mu]_+^2+\frac{\eta}{2}+\mu\right\}\right]$

Consistency property (regularized adversarial loss converges to non-regularized one) holds if $\mathrm{dom}(f)=\mathbb{R}_+$

Example: $f(x)=\mathbb{I}\{0\le x\le \alpha^{-1}\}$, $V_{\text{Dual}}=\mathbb{E}_{z\sim\mathbb{P}_n}\left[\mathtt{AV@R}_{\alpha,\mathbb{Q}_{z,\rho}}(\ell(\cdot\,;\theta))\right]$

Recovery of Worst-case Distribution

$$(\mathbb{P}^*, \gamma^*) = \underset{\mathbb{P},\gamma}{\operatorname{argmax}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] - \mathbb{D}_f(\gamma, \gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \mathbf{d}(x, y) \le \rho \end{array} \right\}$$

$$\frac{\mathrm{d}\mathbb{P}^*(\omega)}{\mathrm{d}\omega} = \mathbb{E}_{z\sim\mathbb{P}_n}\left[\alpha_z \cdot \mathbf{1}\{\mathbf{d}(\omega, z) \le \rho\} \cdot (\eta f)^{*\prime}(\ell(\omega;\theta) - \mu_z^*)\right]$$

Normalizing Constant

Support Constraint

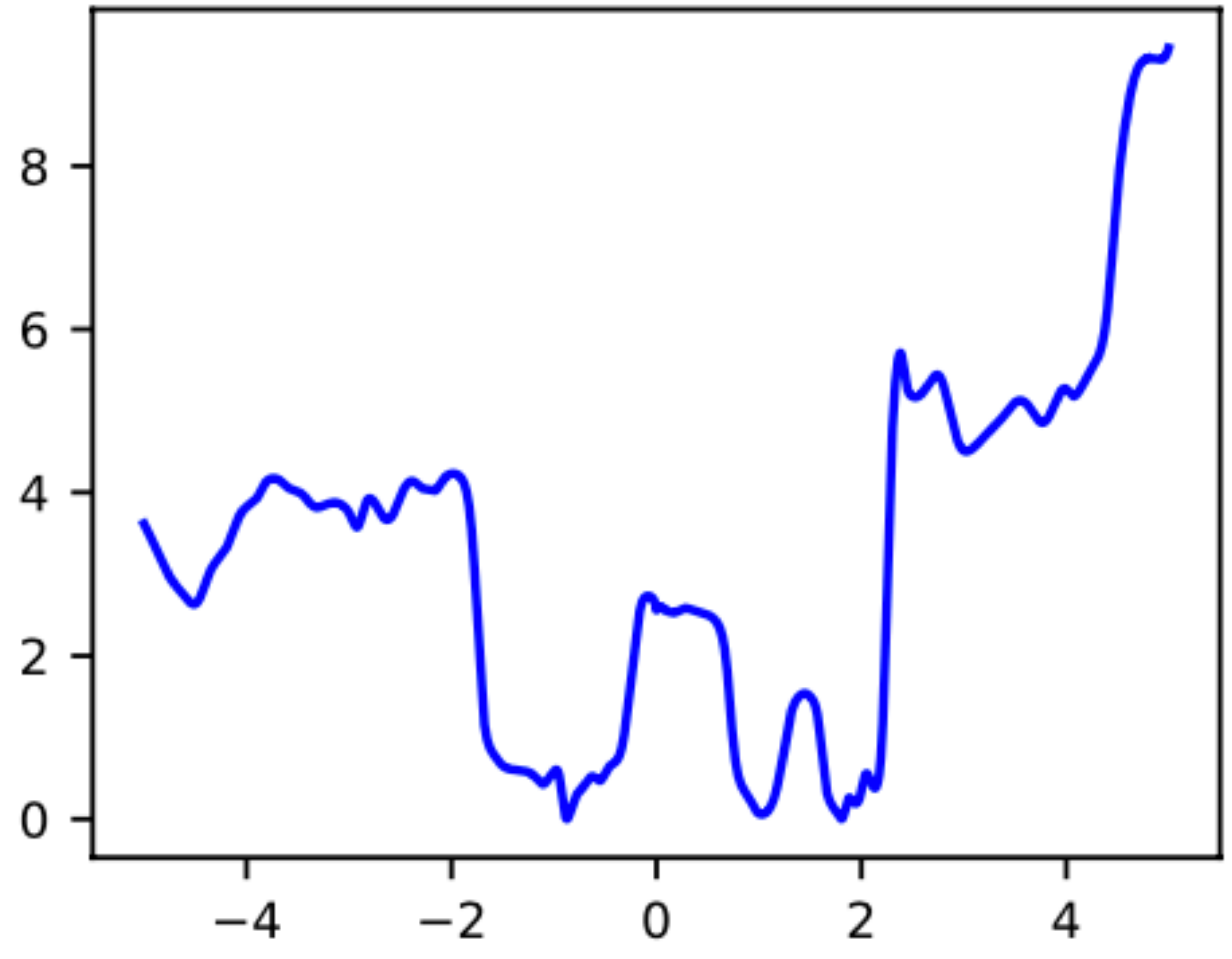
Density contributed by z

Recovery of Worst-case Distribution

$$(\mathbb{P}^*, \gamma^*) = \underset{\mathbb{P},\gamma}{\operatorname{argmax}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] - \mathbb{D}_f(\gamma, \gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \mathbf{d}(x, y) \leq \rho \end{array} \right\}$$

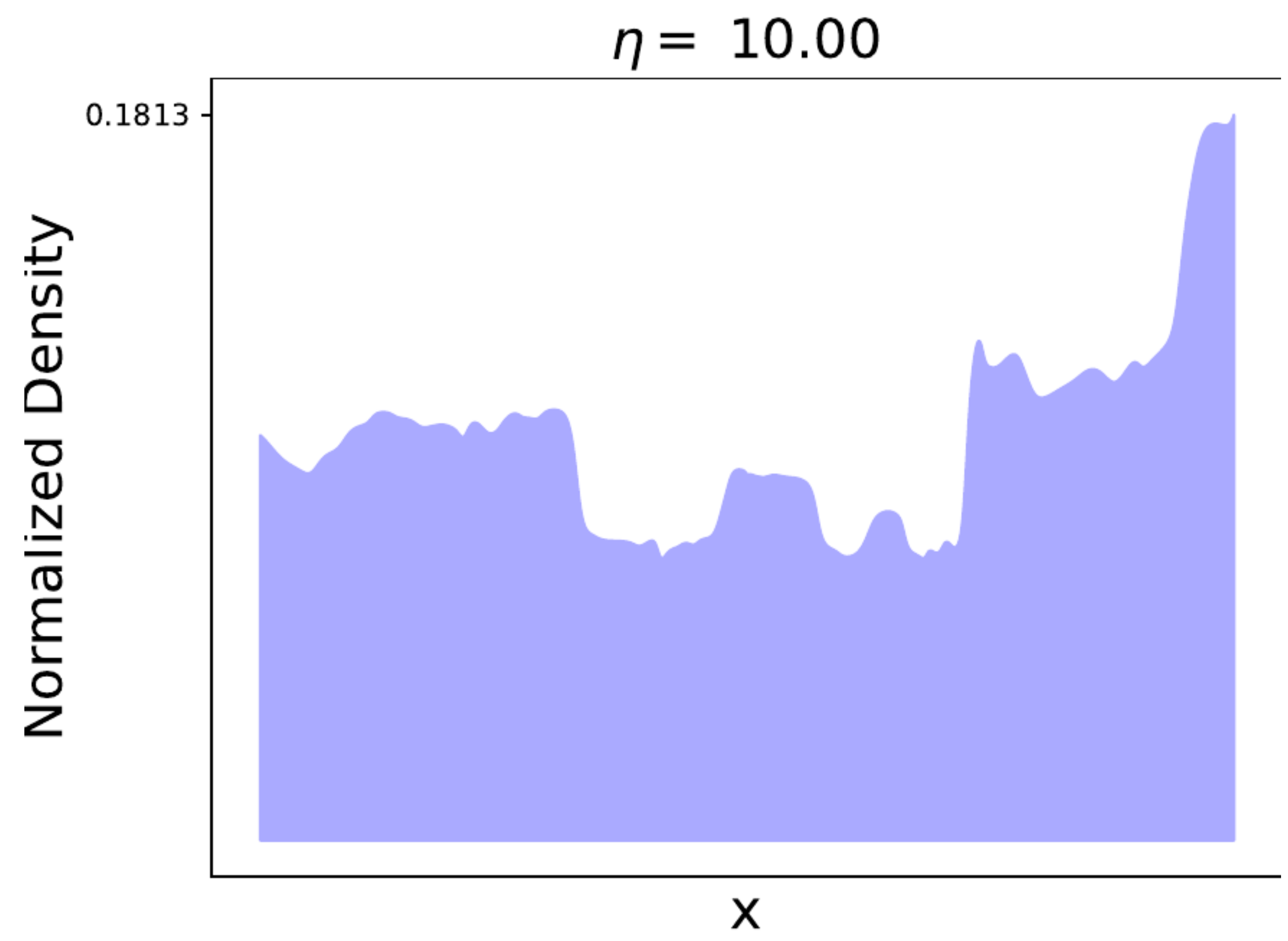
Recovery of Worst-case Distribution

$$(\mathbb{P}^*, \gamma^*) = \underset{\mathbb{P},\gamma}{\operatorname{argmax}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] - \mathbb{D}_f(\gamma, \gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \mathbf{d}(x, y) \leq \rho \end{array} \right\}$$



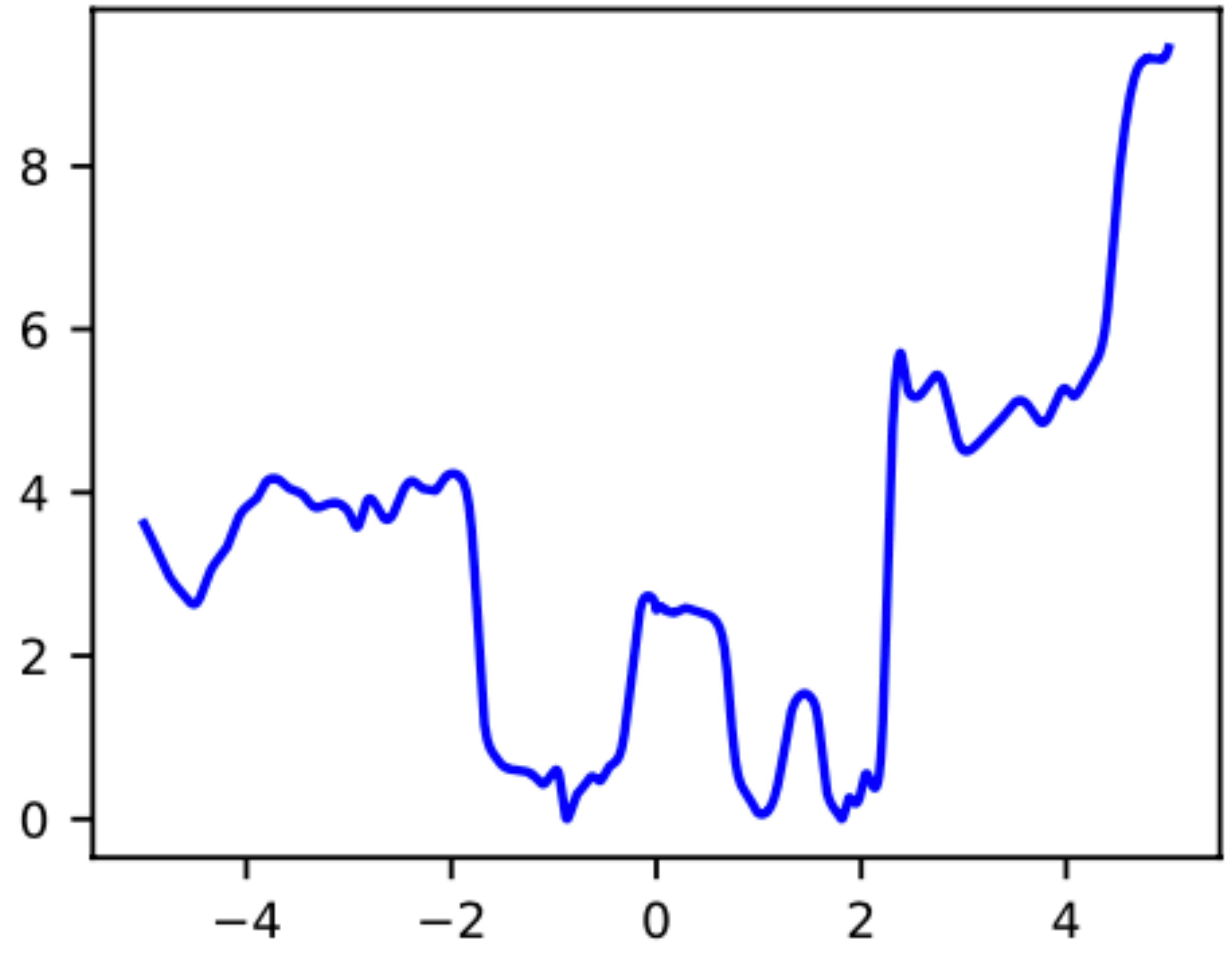
Landscape of 3-layer neural network

Entropic Regularization:



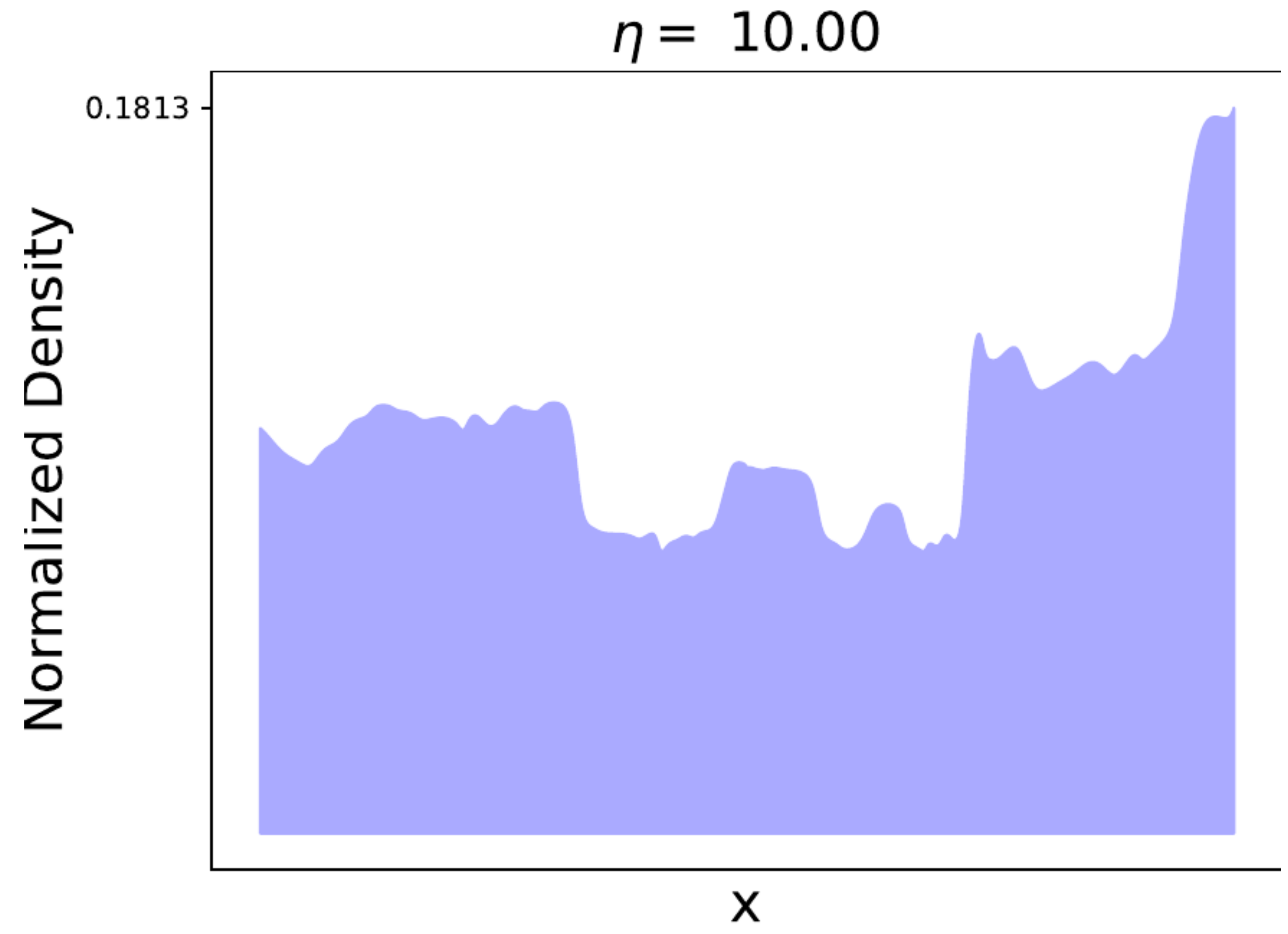
Recovery of Worst-case Distribution

$$(\mathbb{P}^*, \gamma^*) = \underset{\mathbb{P},\gamma}{\operatorname{argmax}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] - \mathbb{D}_f(\gamma, \gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \mathbf{d}(x,y) \le \rho \end{array} \right\}$$



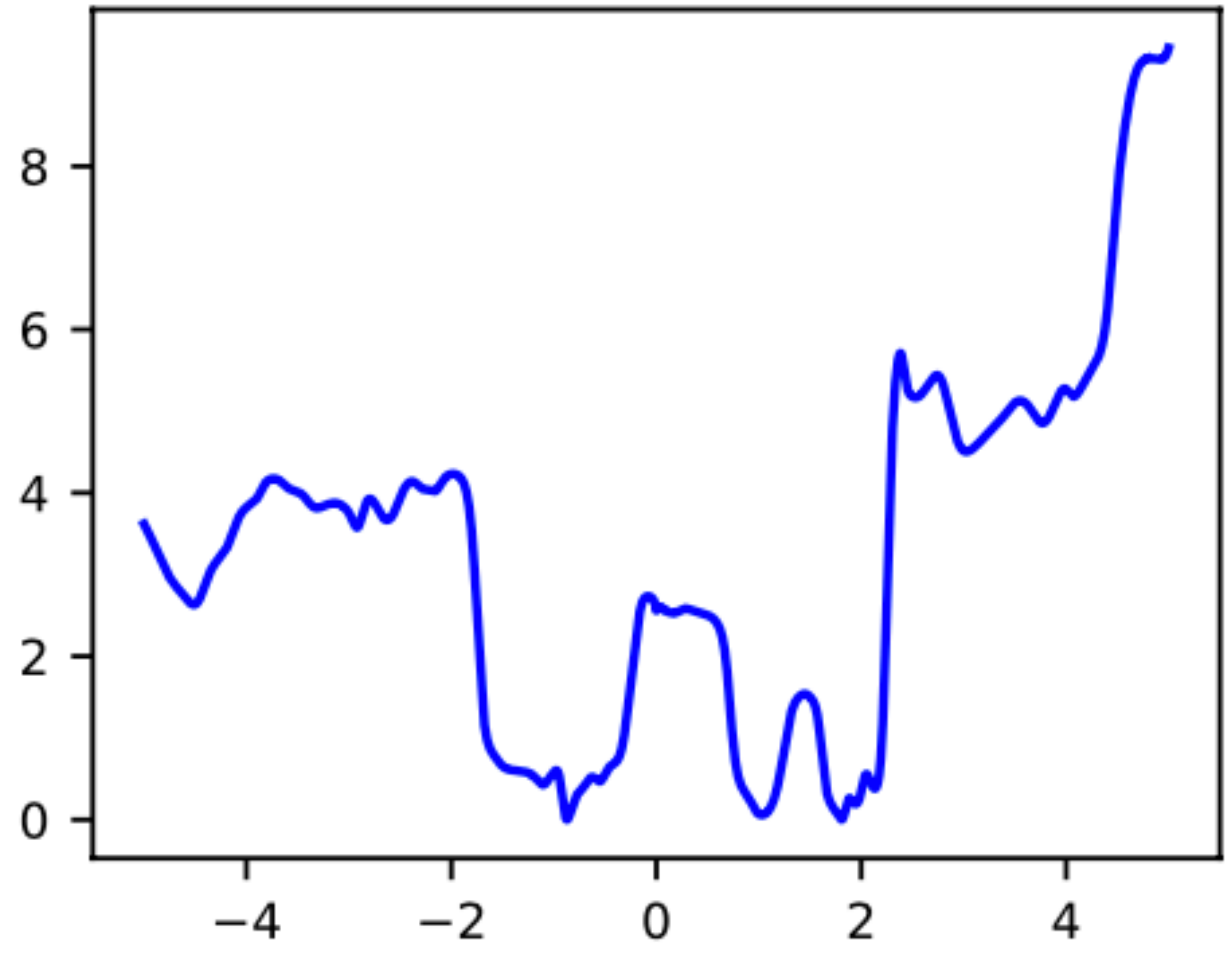
Landscape of 3-layer neural network

Entropic Regularization:



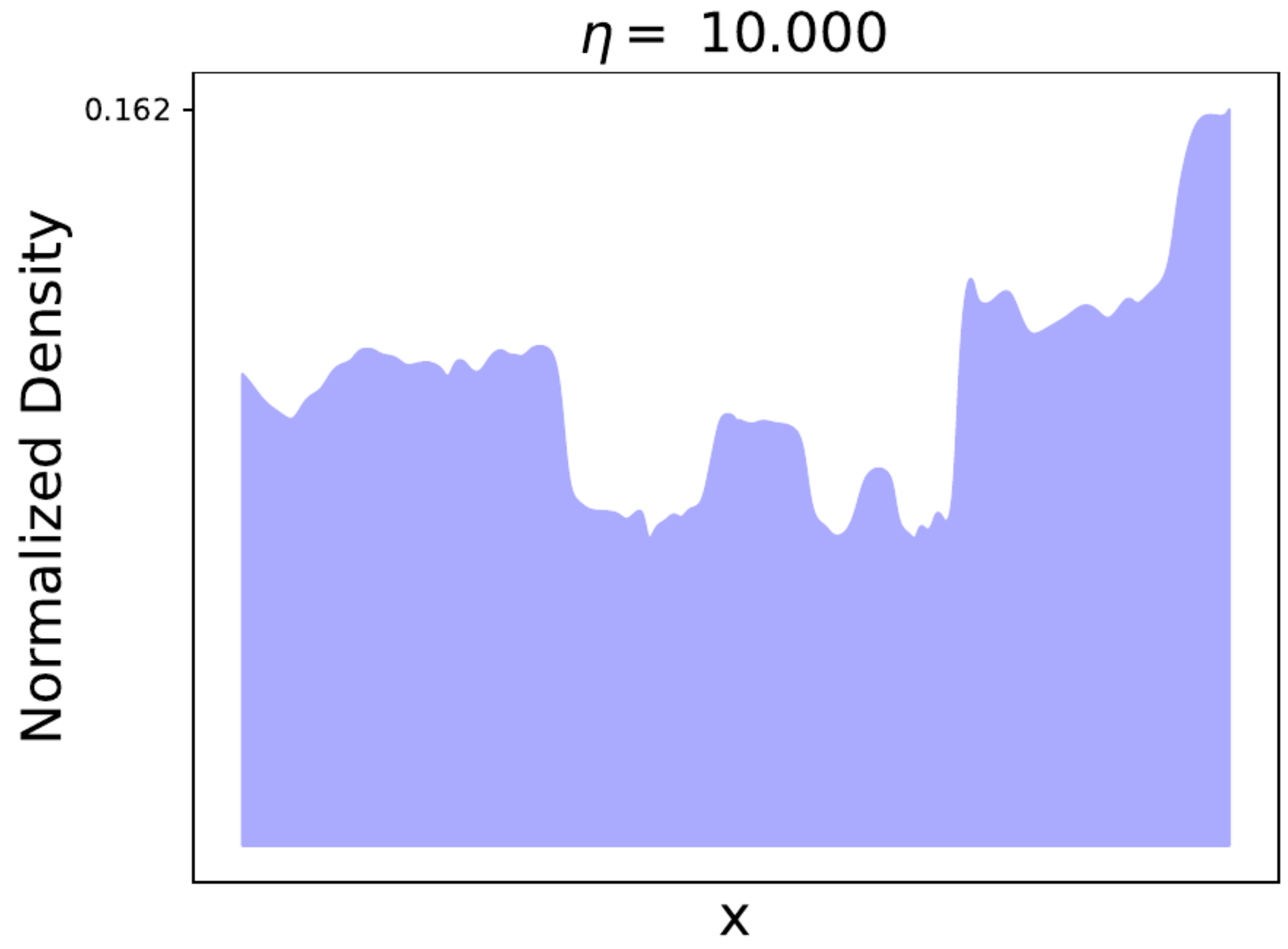
Recovery of Worst-case Distribution

$$(\mathbb{P}^*, \gamma^*) = \underset{\mathbb{P},\gamma}{\operatorname{argmax}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] - \mathbb{D}_f(\gamma, \gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \mathbf{d}(x,y) \le \rho \end{array} \right\}$$



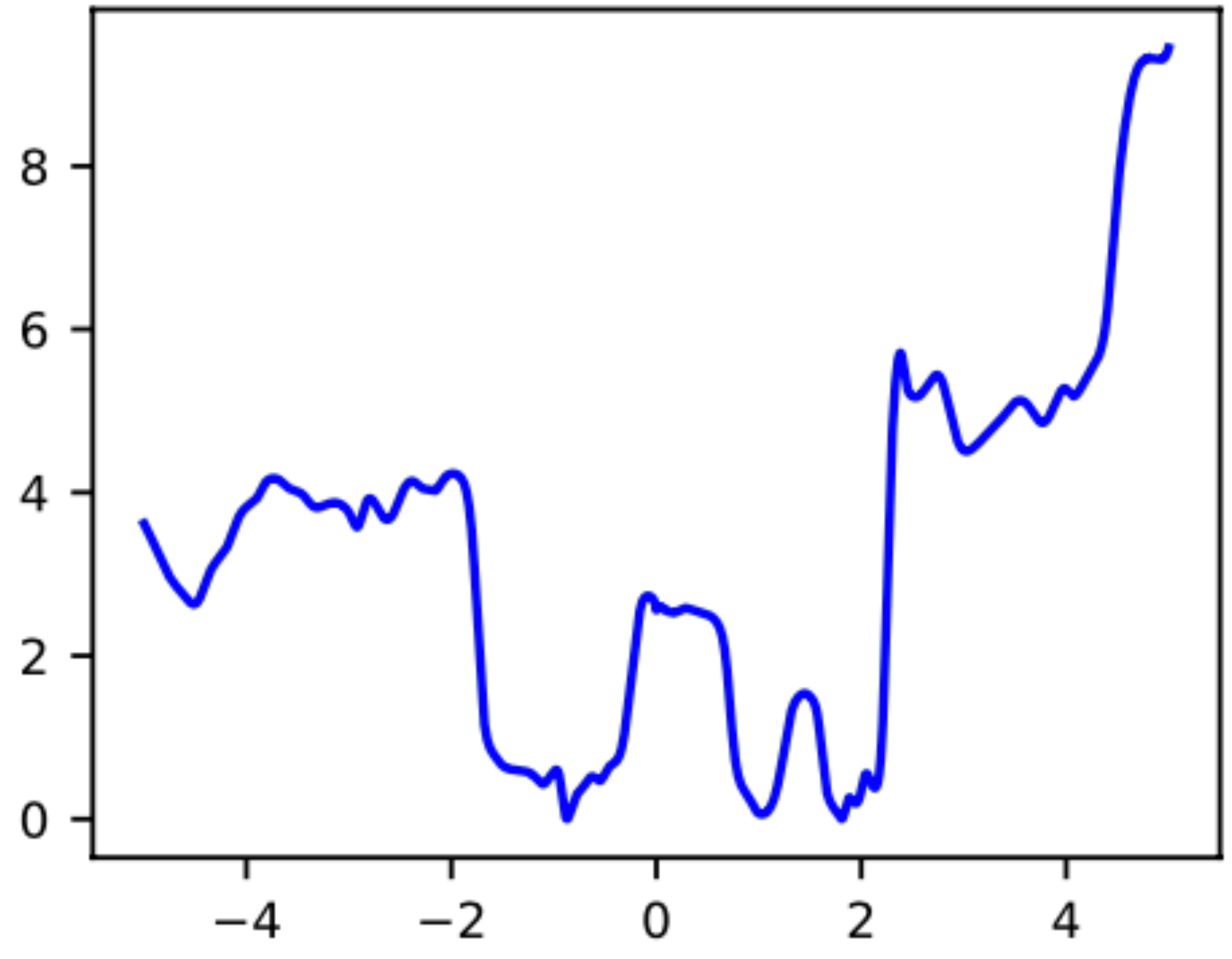
Landscape of 3-layer neural network

Quadratic Regularization:



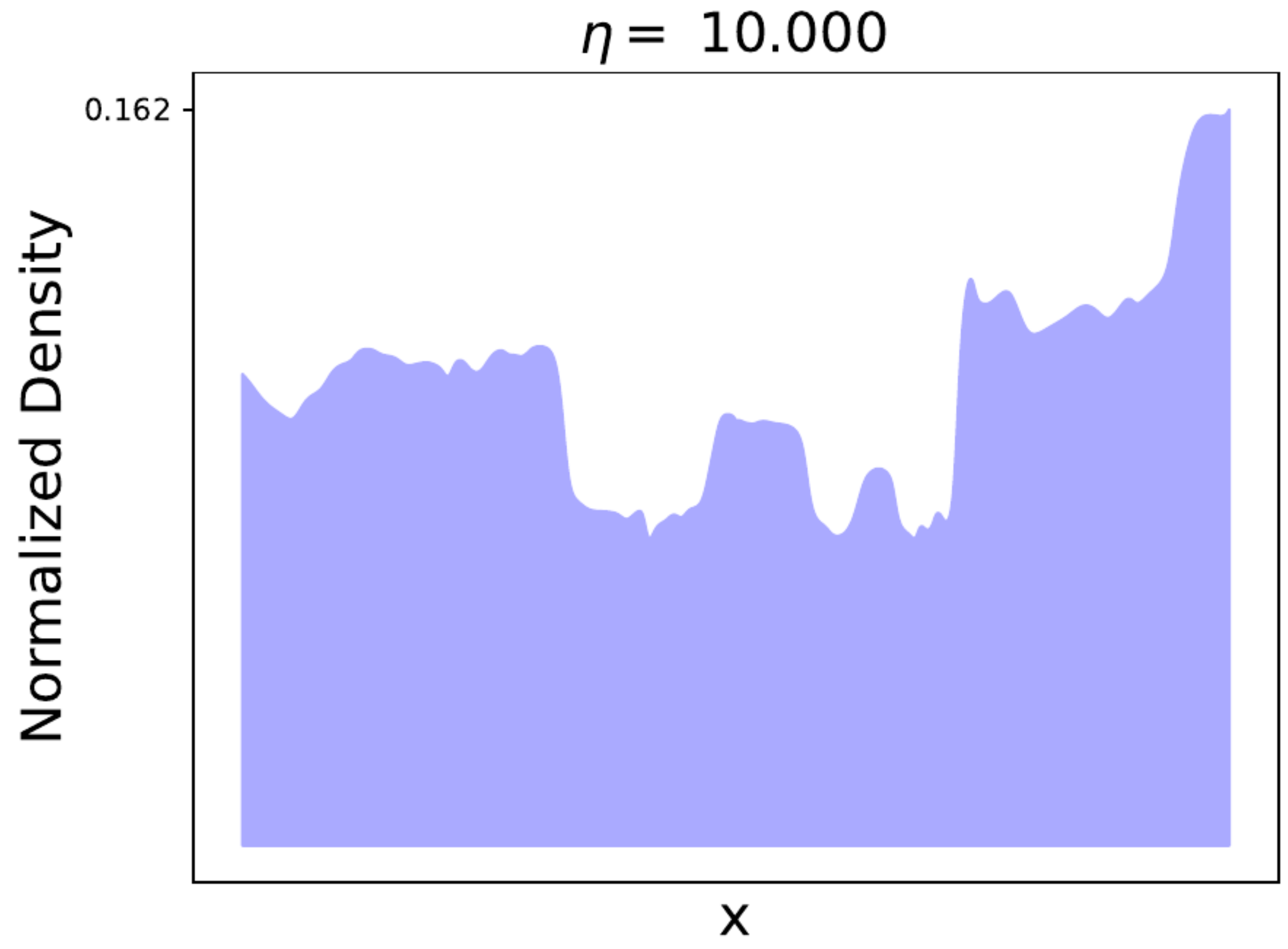
Recovery of Worst-case Distribution

$$(\mathbb{P}^*, \gamma^*) = \underset{\mathbb{P},\gamma}{\operatorname{argmax}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] - \mathbb{D}_f(\gamma, \gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \mathbf{d}(x,y) \le \rho \end{array} \right\}$$



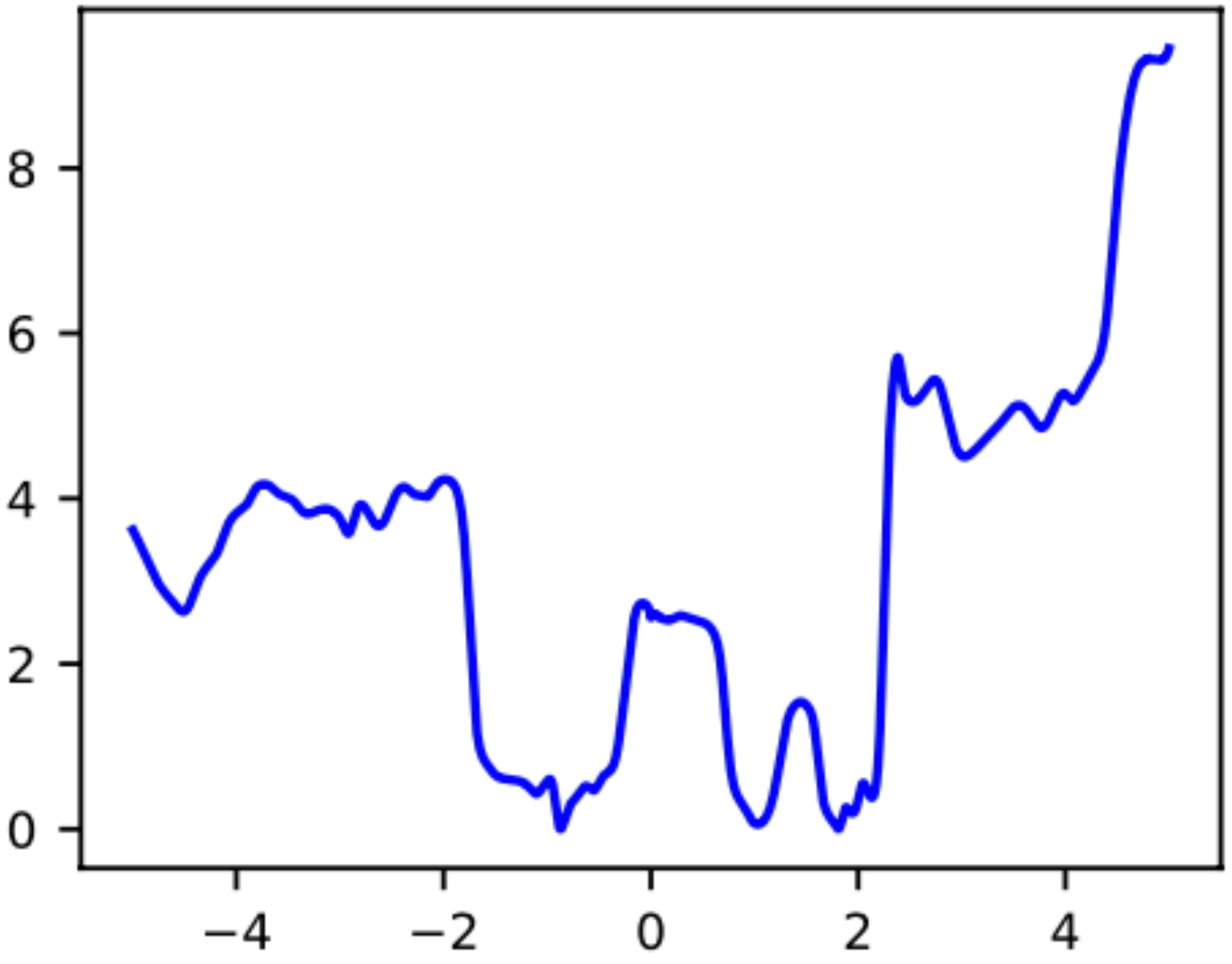
Landscape of 3-layer neural network

Quadratic Regularization:



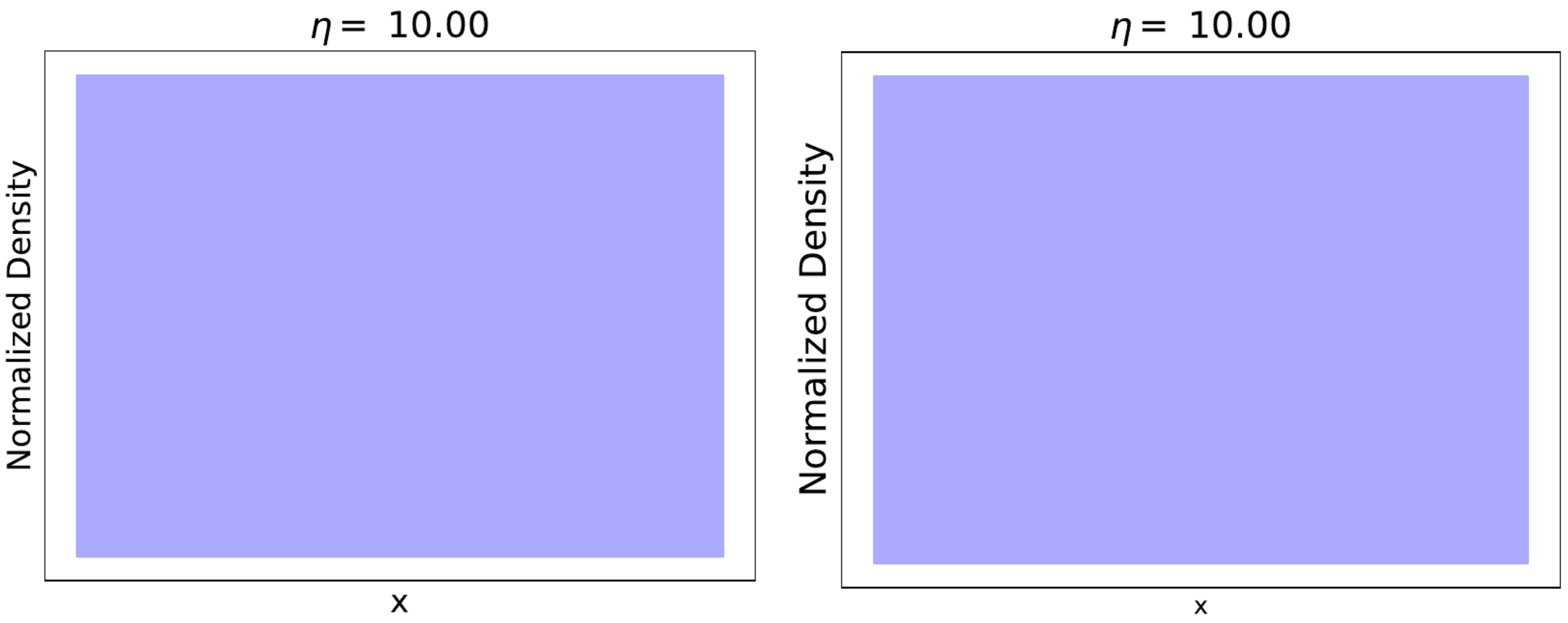
Recovery of Worst-case Distribution

$$(\mathbb{P}^*, \gamma^*) = \underset{\mathbb{P},\gamma}{\operatorname{argmax}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] - \mathbb{D}_f(\gamma,\gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \mathbf{d}(x,y) \le \rho \end{array} \right\}$$



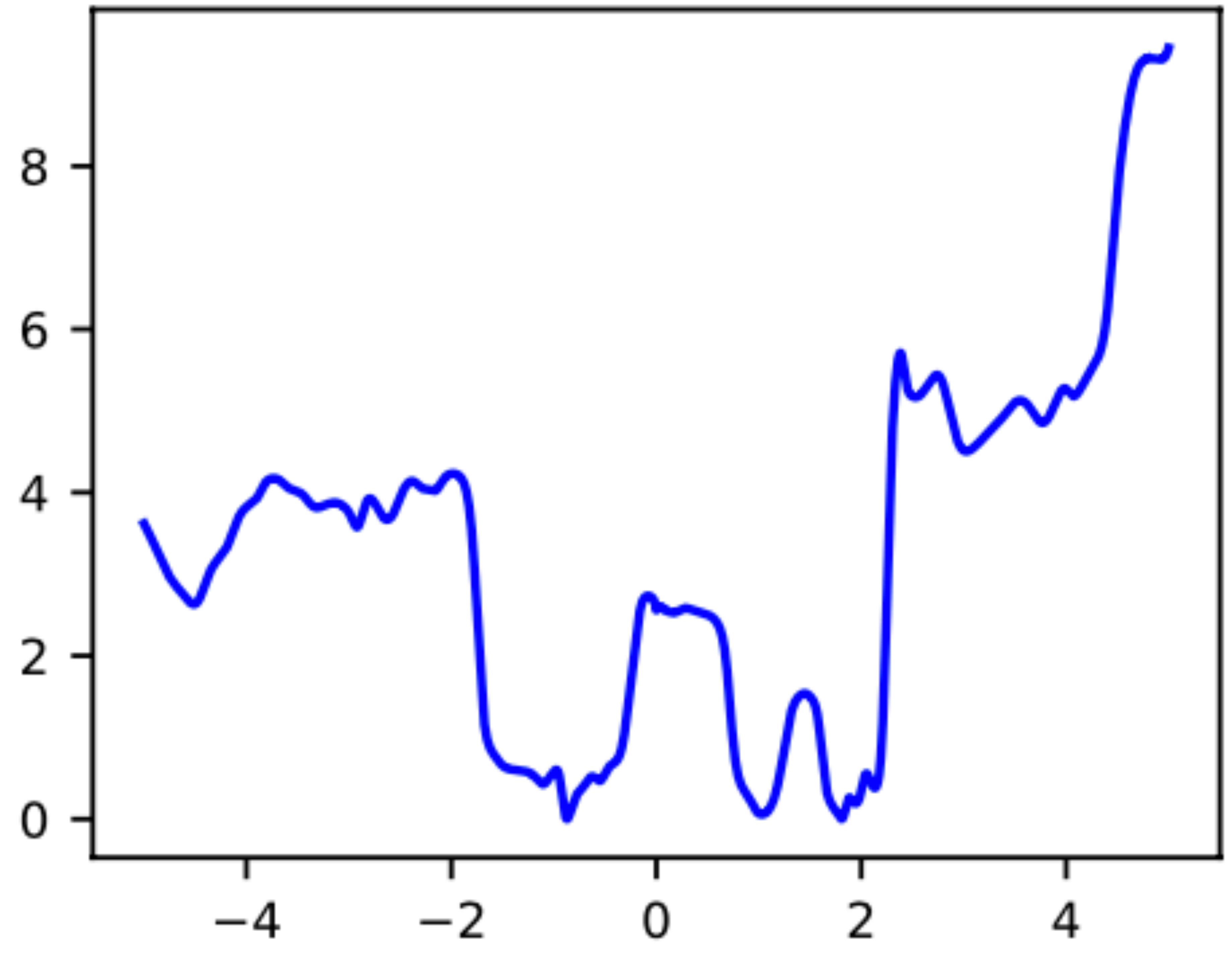
Landscape of 3-layer neural network

Absolute Value/Hinge Loss Regularization:



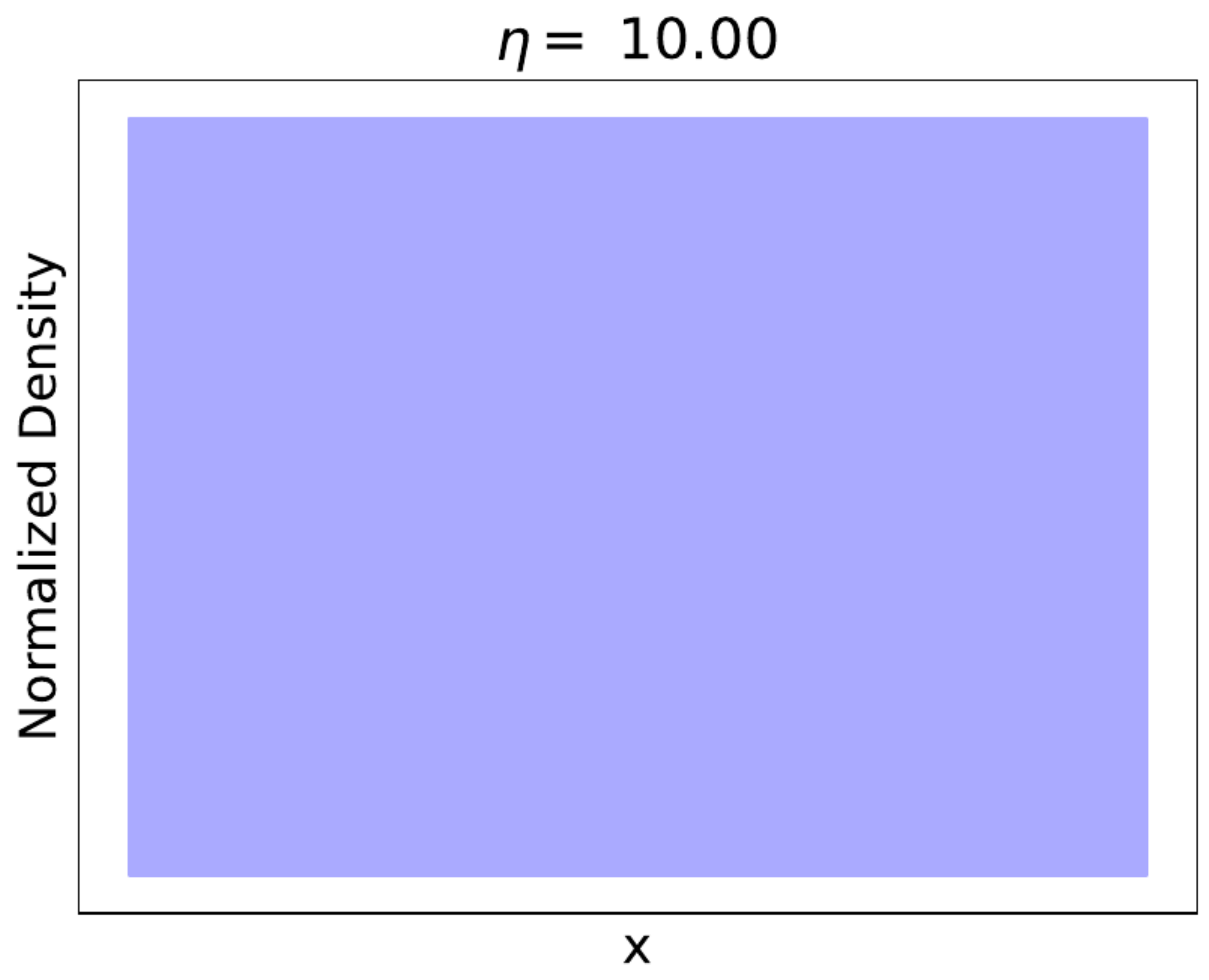
Recovery of Worst-case Distribution

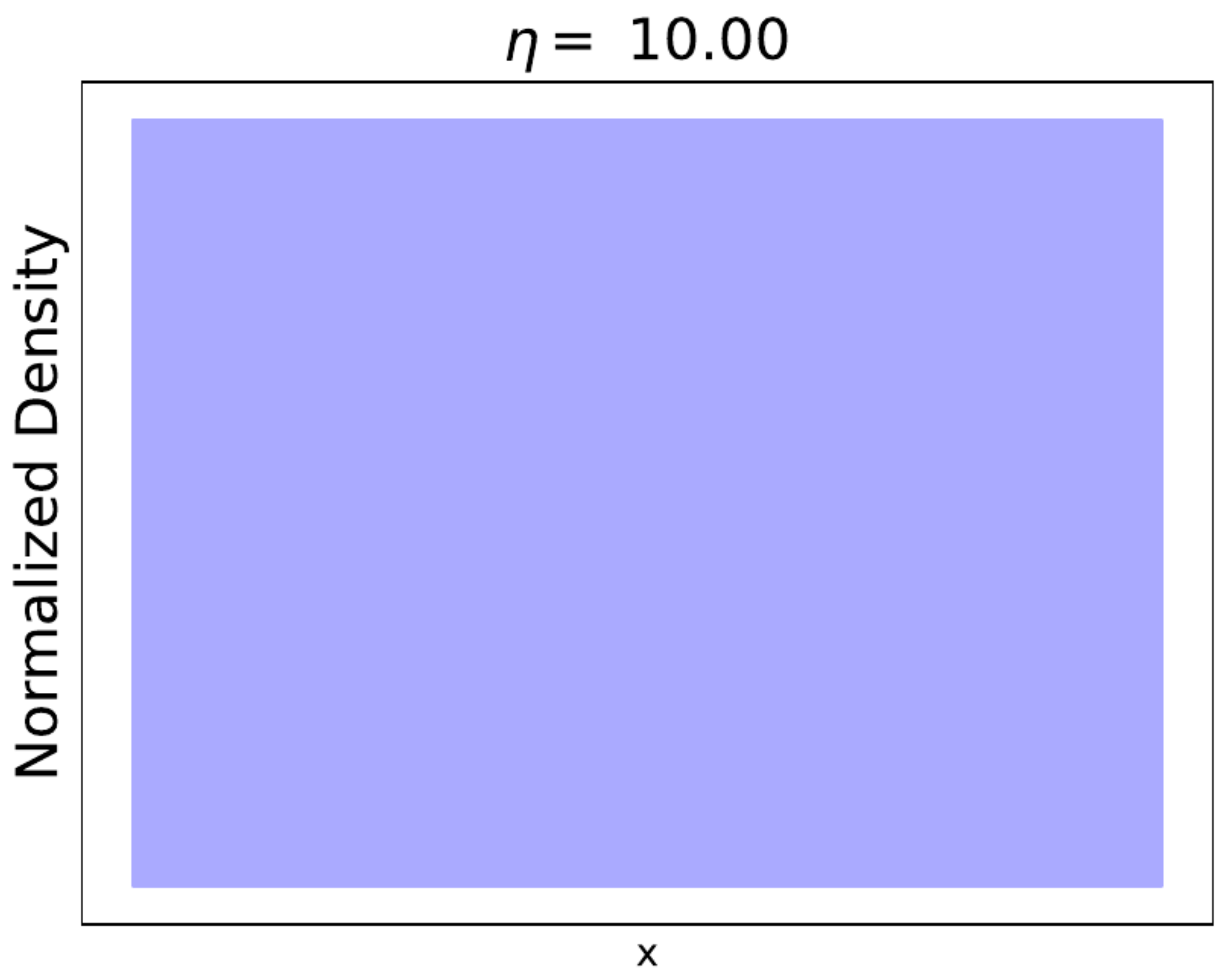
$$(\mathbb{P}^*, \gamma^*) = \underset{\mathbb{P},\gamma}{\operatorname{argmax}} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] - \color{red}{\mathbb{D}_f(\gamma,\gamma_0)} : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \mathbf{d}(x,y) \le \rho \end{array} \right\}$$



Landscape of 3-layer neural network

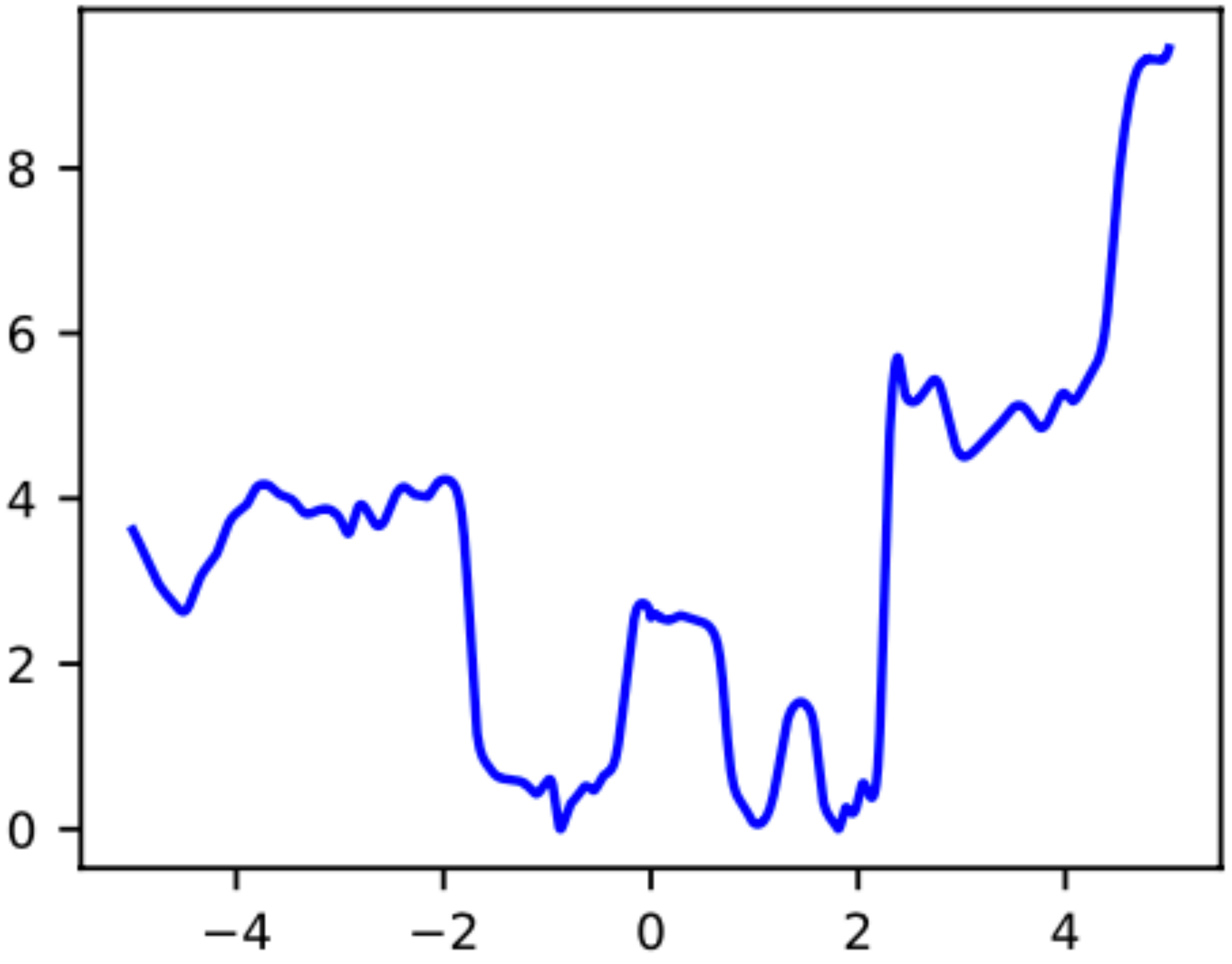
Absolute Value/Hinge Loss Regularization:





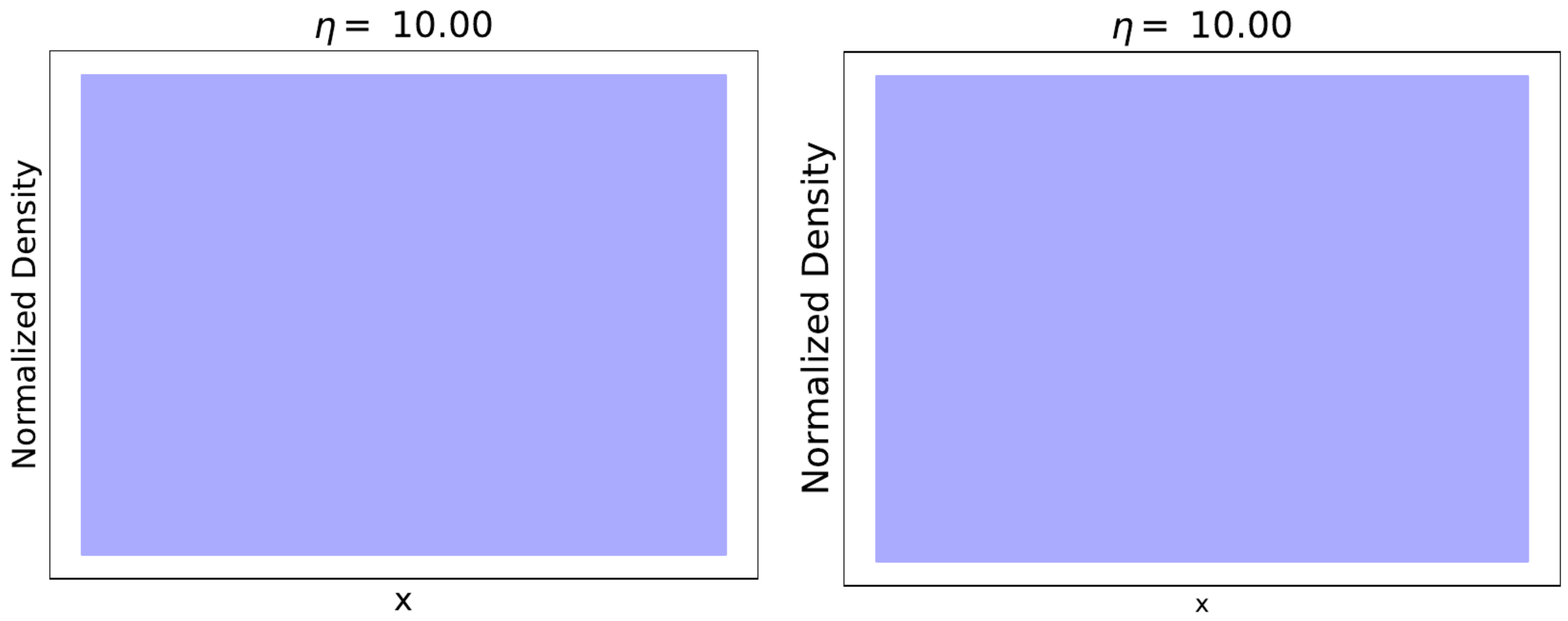
Recovery of Worst-case Distribution

$$(\mathbb{P}^*, \gamma^*) = \operatorname*{argmax}_{\mathbb{P}, \gamma} \left\{ \mathbb{E}_{z \sim \mathbb{P}}[\ell(z;\theta)] - \mathbb{D}_f(\gamma, \gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \mathbf{d}(x, y) \leq \rho \end{array} \right\}$$



Landscape of 3-layer neural network

Absolute Value/Hinge Loss Regularization:



3. Algorithm Design

Tractable Algorithm

- **Ideal formulation:**

$$\min_{\theta\in\Theta} \mathbb{E}_{z\sim\mathbb{P}_n} \left[\inf_{\mu\in\mathbb{R}} \left\{ \mu + \mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[(\eta f)^*(\ell(z';\theta) - \mu)] \right\} \right]$$

Tractable Algorithm

- **Approximation:**

$$\min_{\theta\in\Theta} \mathbb{E}_{\substack{\bullet\ z\sim\mathbb{P}_n \\ \bullet\ \{z_i'\}_{i\in[2^l]}\sim\mathbb{Q}_{z,\rho}}} \left[\inf_{\mu\in\mathbb{R}} \left\{ \mu + \frac{1}{2^l} \sum_{i\in[2^l]} [(\eta f)^*(\ell(z_i';\theta) - \mu)] \right\} \right]$$

Tractable Algorithm

- **Ideal formulation:**

$$\min_{\theta\in\Theta} \boxed{\mathbb{E}_{z\sim\mathbb{P}_n}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[(\eta f)^*(\ell(z';\theta)-\mu)]\right\}\right]} \triangleq F(\theta)$$

- **Approximation:**

$$\min_{\theta\in\Theta} \boxed{\mathbb{E}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\frac{1}{2^l}\sum_{i\in[2^l]}[(\eta f)^*(\ell(z_i';\theta)-\mu)]\right\}\right]} \triangleq F^l(\theta)$$

- $z \sim \mathbb{P}_n$
- $\{z_i'\}_{i\in[2^l]} \sim \mathbb{Q}_{z,\rho}$

Tractable Algorithm

- **Approximation:**

$$\min_{\theta\in\Theta} \mathbb{E}_{\substack{\bullet\ z\sim\mathbb{P}_n \\ \bullet\ \{z_i'\}_{i\in[2^l]}\sim\mathbb{Q}_{z,\rho}}} \boxed{\left[\inf_{\mu\in\mathbb{R}} \left\{\mu + \frac{1}{2^l}\sum_{i\in[2^l]} [(\eta f)^*\ \ell(z_i';\theta) - \mu)]\right\}\right]} \triangleq F^l(\theta)$$

SGD with Naive Estimator: Fix large $l \equiv L$,

If converged? $\xrightarrow{\text{False}}$ **Sample $z\sim\mathbb{P}_n$ and $\{z_i'\}_{i\in[2^L]}\sim\mathbb{Q}_{z,\rho}$** $\rightarrow$ **Generate Sample Estimate of $\nabla F^L(\theta)$ and Update** $\rightarrow$ **If converged?**

Tractable Algorithm

- **Approximation:**

$$\min_{\theta\in\Theta}\ \mathbb{E}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\frac{1}{2^l}\sum_{i\in[2^l]}\left[(\eta f)^*\left(\ell(z_i';\theta)-\mu\right)\right]\right\}\right] \triangleq F^l(\theta)$$

- $z \sim \mathbb{P}_n$
- $\{z_i'\}_{i\in[2^l]} \sim \mathbb{Q}_{z,\rho}$

SGD with Random Sampling Estimator:

If converged?

False

Sample $\hat{L}$ with $P(\hat{L}=i) \propto 2^{-i}, i=0,\ldots,L$

Sample $z \sim \mathbb{P}_n$ and $\{z_i'\}_{i\in[2^{\hat{L}}]} \sim \mathbb{Q}_{z,\rho}$

Generate Sample Estimate of $P(\hat{L})^{-1}(\nabla F^{\hat{L}}(\theta) - \nabla F^{\hat{L}-1}(\theta))$ **and Update**

Tractable Algorithm

$$\mathbb{E}_{\hat{L}}\left[P(\hat{L})^{-1}(\nabla F^{\hat{L}}(\theta) - \nabla F^{\hat{L}-1}(\theta))\right] = \sum_{\hat{L}} \nabla F^{\hat{L}}(\theta) - \nabla F^{\hat{L}-1}(\theta) = \nabla F^{L}(\theta)$$

SGD with Random Sampling Estimator:

Complexity for Optimizing F

$$\min_{\theta\in\Theta} \boxed{\mathbb{E}_{z\sim\mathbb{P}_n}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[(\eta f)^*(\ell(z';\theta)-\mu)]\right\}\right]} \triangleq F(\theta)$$

Algorithm	Naive Estimator		Random Sampling Estimator	
Loss $\ell(z,\cdot)$	Convex		Convex	
Choice of f-divergence	Arbitrary		Arbitrary	
Complexity	$\tilde{O}(\delta^{-3})$		$\tilde{O}(\delta^{-2})$	

Complexity for Optimizing F

$$\min_{\theta\in\Theta} \boxed{\mathbb{E}_{z\sim\mathbb{P}_n}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\mathbb{E}_{z'\sim\mathbb{Q}_{z,\rho}}[(\eta f)^*(\ell(z';\theta)-\mu)]\right\}\right]} \triangleq F(\theta)$$

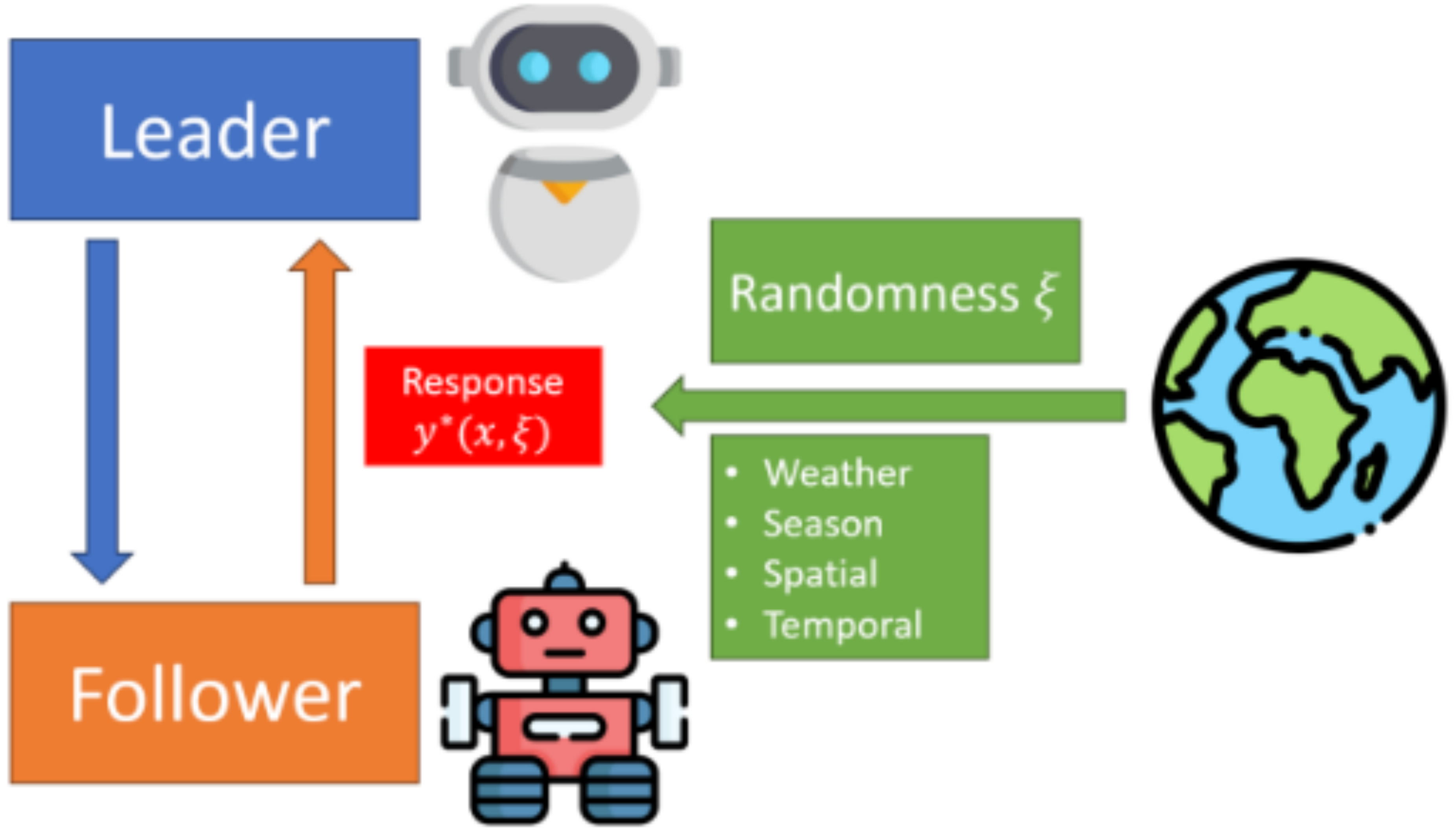
Algorithm	Naive Estimator		Random Sampling Estimator	
Loss $\ell(z,\cdot)$	Convex	Nonconvex Smooth	Convex	Nonconvex Smooth
Choice of f-divergence	Arbitrary	**KL-divergence**	Arbitrary	**KL-divergence**
Complexity	$\tilde{O}(\delta^{-3})$	$\tilde{O}(\delta^{-6})$	$\tilde{O}(\delta^{-2})$	$\tilde{O}(\delta^{-4})$

General Optimization Results

- **Goal**: $\min_{\theta} F(\theta)$, whereas **unbiased** gradient of $F(\theta)$ is **not available**!
- **Assumption**:
  - Gradient of approximation objective F^l is easy to obtain
  - $|F^l(\theta) - F(\theta)| = O(2^{-l})$ or $\|\nabla F^l(\theta) - \nabla F(\theta)\|^2 = O(2^{-l})$
- **Examples**:

Contextual Bilevel Optimization

$$\begin{aligned}\text{minimize} \quad & F(\theta) \triangleq \mathbb{E}_{\xi}\left[f(\theta, y^*(\theta;\xi))\right] \\ \text{where} \quad & y^*(\theta;\xi) \triangleq \operatorname{argmin}_y \mathbb{E}_{\eta\sim\mathbb{P}_{\eta|\xi}}\left[g(x,y;\eta,\xi)\right], \quad \forall \xi\end{aligned}$$



Hu Y, **Wang J**, Xie Y, Krause A, Kuhn D (2023) Contextual stochastic bilevel optimization. *NIPS'23*

Hu Y, **Wang J**, Chen X, He N (2024) Multi-level Monte-Carlo Gradient Methods for Stochastic Optimization with Biased Oracles. *arXiv preprint*

General Optimization Results

- **Goal**: min $F(\theta)$, whe... **able**!
- **As**...
  - ...
  - ... $\|^2 = O(2^{-l})$

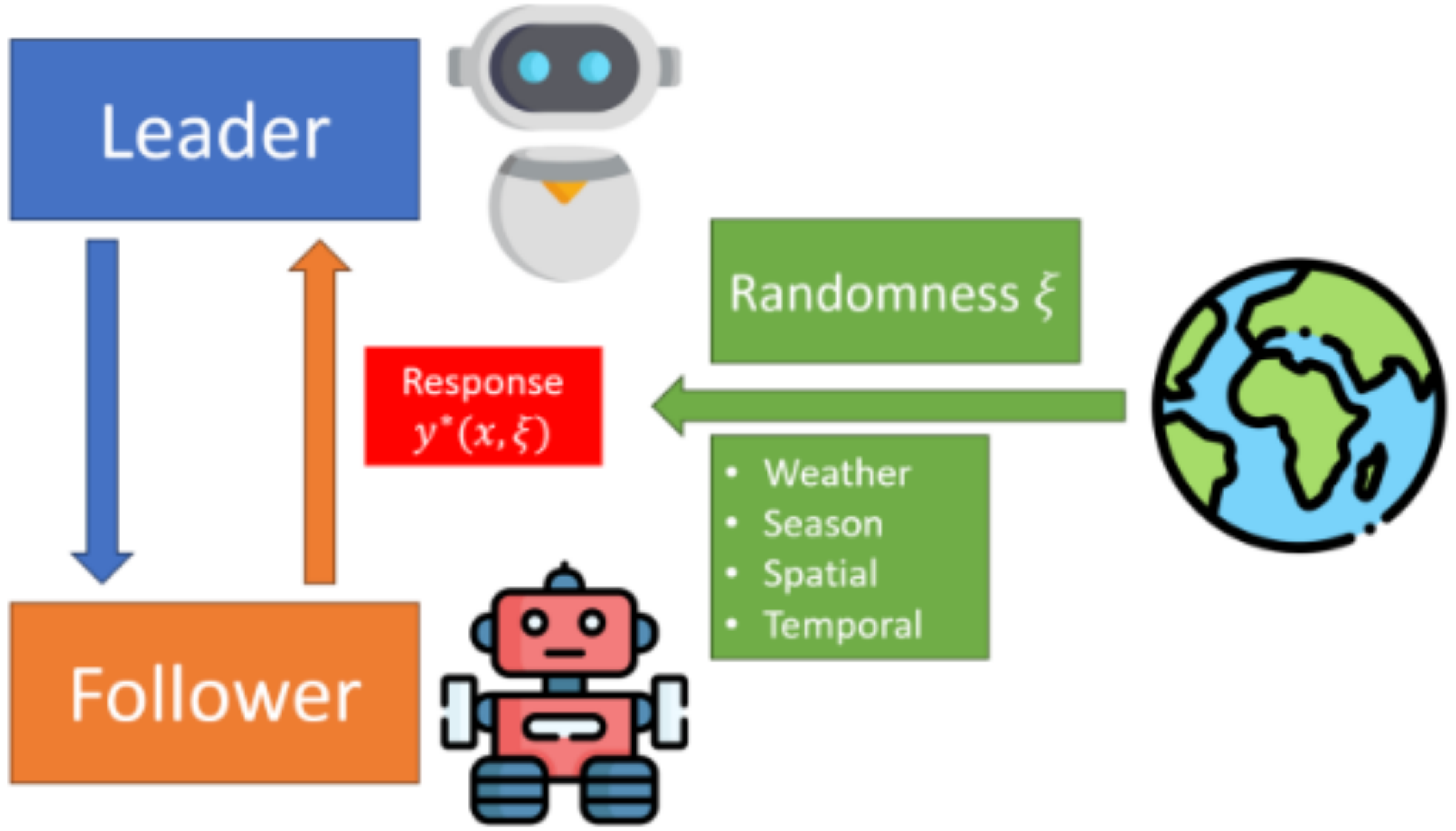
Random Sampling Gradient Estimator Achieves Optimal Complexity on those examples!

- **Examples**:

Contextual Bilevel Optimization

$$\text{minimize} \quad F(\theta) \triangleq \mathbb{E}_{\xi}\left[f(\theta, y^*(\theta;\xi))\right]$$

$$\text{where} \quad y^*(\theta;\xi) \triangleq \operatorname{argmin}_y \mathbb{E}_{\eta\sim\mathbb{P}_{\eta|\xi}}\left[g(x, y; \eta, \xi)\right], \quad \forall \xi$$



Hu Y, **Wang J**, Xie Y, Krause A, Kuhn D (2023) Contextual stochastic bilevel optimization. *NIPS'23*

Hu Y, **Wang J**, Chen X, He N (2024) Multi-level Monte-Carlo Gradient Methods for Stochastic Optimization with Biased Oracles. *arXiv preprint*

4. Statistical Analysis

Regularization Effects

- Regularized adversarial learning:

$$\min_{\theta\in\Theta}\ \sup_{\mathbb{P},\gamma}\left\{\mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]-\eta\mathbb{D}_f(\gamma,\gamma_0)\ :\ \begin{array}{l}\gamma\in\Gamma(\mathbb{P}_n.\mathbb{P})\\ \gamma\text{-esssup}\ \|x-y\|\le\rho\end{array}\right\}$$

- Regularization Effects:

$$\boxed{(\text{Regularized Adversarial Learning})\approx\min_{\theta\in\Theta}\left\{\mathbb{E}_{z\sim\mathbb{P}_n}[\ell(z;\theta)]+\text{Regularization}\right\}}$$

Regularization Effects

- Regularized adversarial learning:

$$\min_{\theta\in\Theta} \sup_{\mathbb{P},\gamma} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] - \eta\mathbb{D}_f(\gamma,\gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n.\mathbb{P}) \\ \gamma\text{-esssup } \|x-y\| \leq \rho \end{array} \right\}$$

- **Case 1:** $\rho/\eta \to \infty$

$$(\text{Regularized Adversarial Learning}) \approx \min_{\theta\in\Theta} \left\{ \mathbb{E}_{z\sim\mathbb{P}_n}[\ell(z;\theta)] + \rho \cdot \boxed{\mathbb{E}_{z\sim\mathbb{P}_n}[\|\nabla\ell(z;\theta)\|]} \right\}$$

  - Recovers regularization for **∞-type Wasserstein DRO**!
  - Hedge against adversarial attack

Regularization Effects

- Regularized adversarial learning:

$$\min_{\theta\in\Theta} \sup_{\mathbb{P},\gamma} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] - \eta \mathbb{D}_f(\gamma,\gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup } \|x-y\| \le \rho \end{array} \right\}$$

- **Case 2:** $\rho/\eta \to 0$

$$\text{(Regularized Adversarial Learning)} \approx \min_{\theta\in\Theta} \left\{ \mathbb{E}_{z\sim\mathbb{P}_n}[\ell(z;\theta)] + \frac{\rho^2}{2\eta f''(1)} \cdot \mathbb{E}_{z\sim\mathbb{P}_n}[\mathbb{V}\mathrm{ar}_{b\sim\beta}[\nabla \ell(z;\theta)^\top b]] \right\}$$

  - Relates to regularization for ***f*-divergence DRO**!
  - Hedge against white noise attack

Regularization Effects

- Regularized adversarial learning:

$$\min_{\theta\in\Theta} \sup_{\mathbb{P},\gamma} \left\{ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] - \eta \mathbb{D}_f(\gamma, \gamma_0) : \begin{array}{l} \gamma \in \Gamma(\mathbb{P}_n, \mathbb{P}) \\ \gamma\text{-esssup} \, \|x - y\| \le \rho \end{array} \right\}$$

- **Case 3:** $\rho/\eta \to C$

$$(\text{Regularized Adversarial Learning}) \approx \min_{\theta\in\Theta} \left\{ \mathbb{E}_{z\sim\mathbb{P}_n}[\ell(z;\theta)] + \rho \cdot \mathbb{E}_{x\sim\mathbb{P}_n}\left[\inf_{\mu\in\mathbb{R}} \left\{ \mu + \frac{1}{C}\mathbb{E}_{b\sim\beta}[f^*(C\cdot(\nabla\ell(z;\theta)^\top b - \mu))] \right\} \right] \right\}$$

  - Relates to **optimized certainty equivalent regularization**
  - Interpolates between **gradient norm** and **variance regularization**!

5. Numerical Study and Conclusion

Numerical Study: MNIST Classification

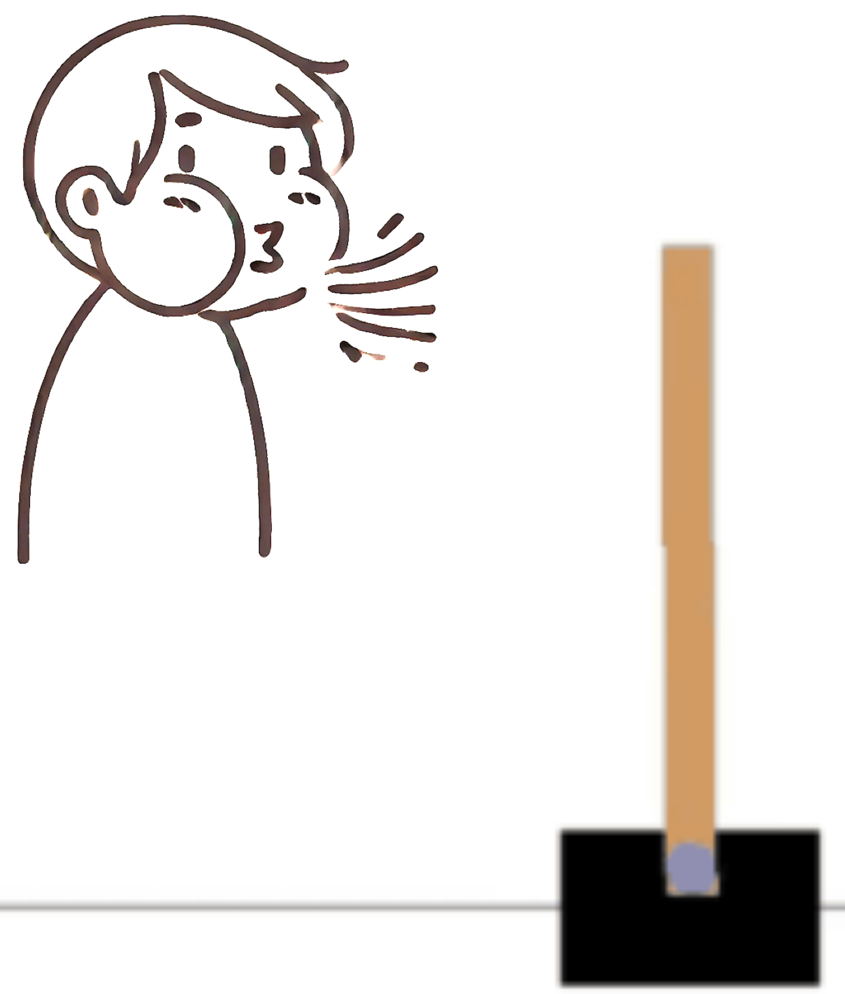
- Goal: Classification with convolutional network and ELU activation
- Training data: MNIST handwritten digits with $6 \cdot 10^4$ samples
- Testing data: digits with 10^4 samples, perturbed by random/adversarial ℓ_2/ℓ_∞ noise

Numerical Study: Reliable Reinforcement Learning

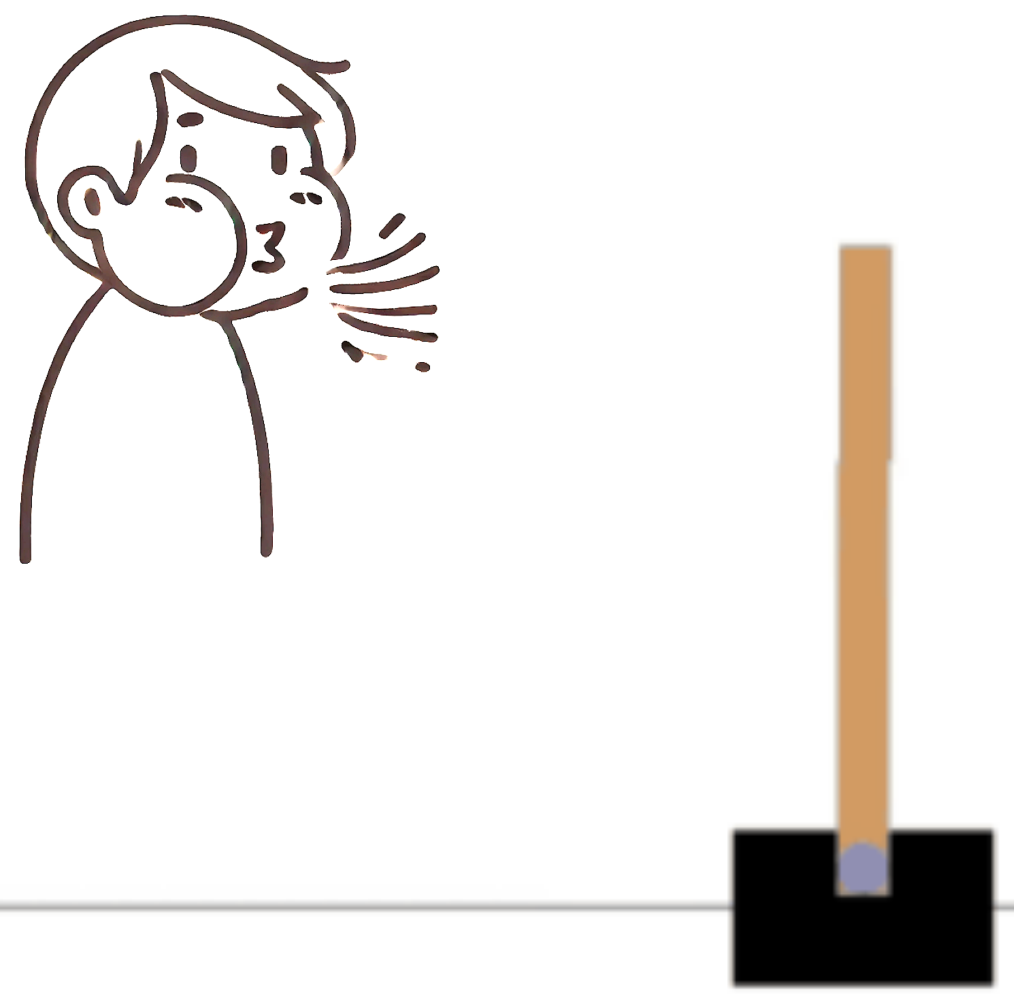
- **Standard Q-learning:** $Q(s^t, a^t) \leftarrow (1-\alpha_t)Q(s^t, a^t) + \alpha_t r(s^t, a^t)$

$$-\gamma\alpha_t \boxed{\min_a \,(-Q(s^{t+1}, a)),} \qquad s^{t+1} \sim \mathbb{P}(\,\cdot \mid s^t, a^t)$$

random agent



trained agent

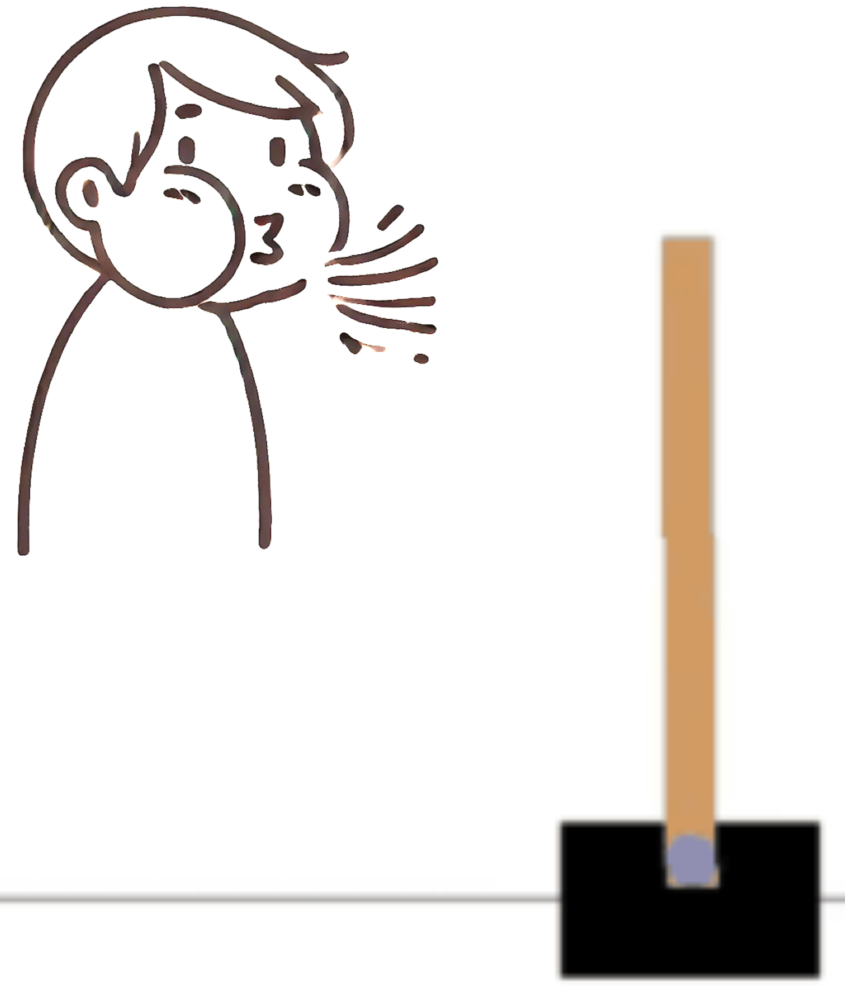


Numerical Study: Reliable Reinforcement Learning

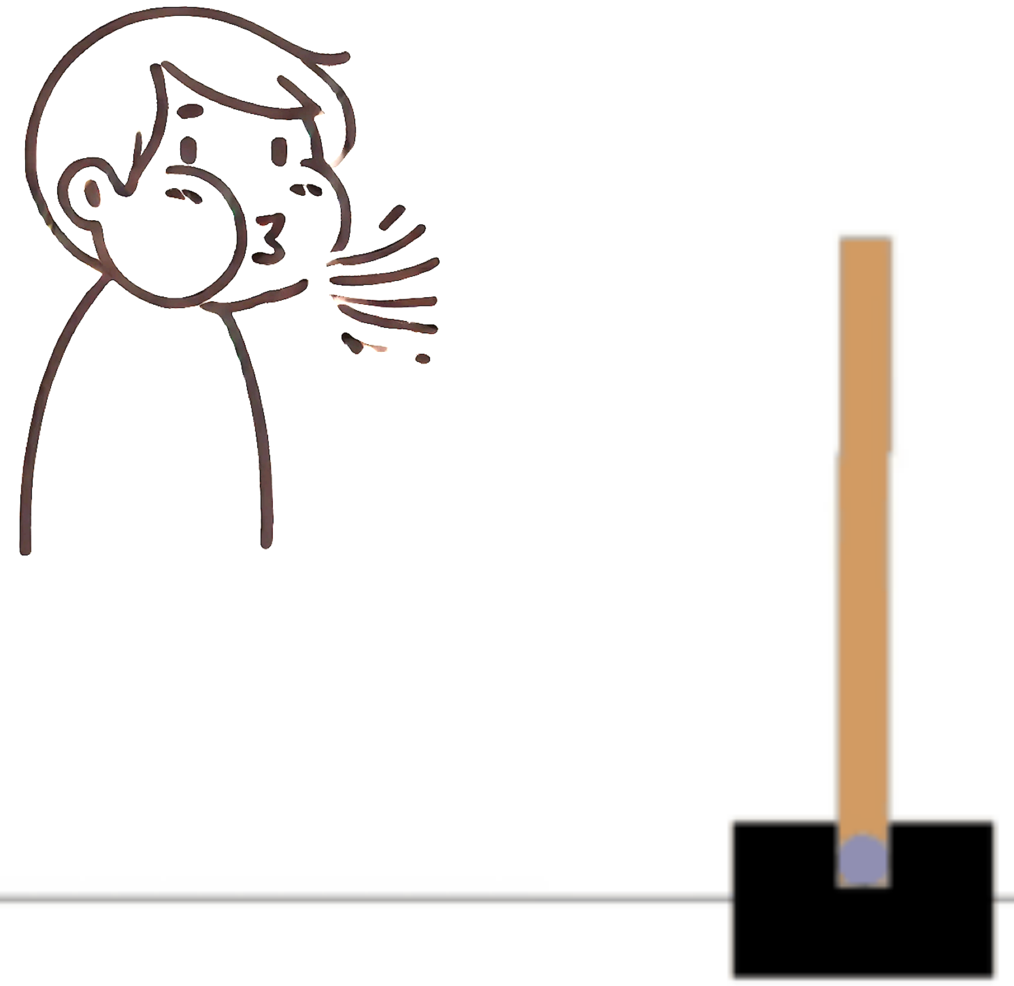
- **Standard Q-learning:** $Q(s^t, a^t) \leftarrow (1 - \alpha_t)Q(s^t, a^t) + \alpha_t r(s^t, a^t)$

$$-\gamma\alpha_t \boxed{\min_a \, (-Q(s^{t+1}, a)),} \qquad s^{t+1} \sim \mathbb{P}(\,\cdot \mid s^t, a^t)$$

random agent



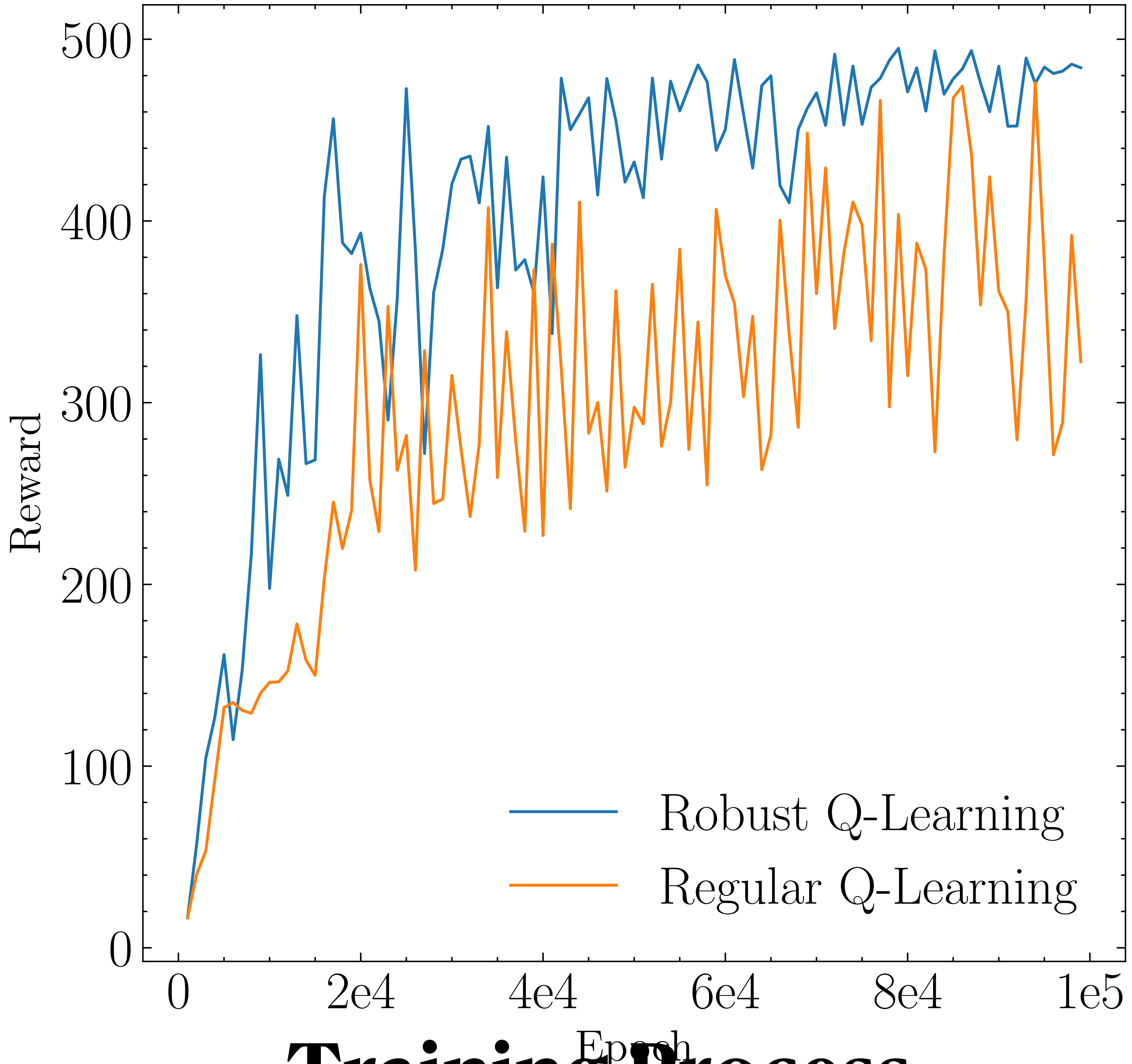
trained agent



Numerical Study: Reliable Reinforcement Learning

Environment	Regular	Robust
Original MDP	469.42±19.03	**487.11±9.09**
Perturbed MDP (Heavy Pole)	187.63±29.40	**394.12±12.01**
Perturbed MDP (Short Pole)	355.54±28.89	**443.17±9.98**
Perturbed MDP (Strong Gravity)	271.41±20.70	**418.42±13.64**

Reward by Regular and Robust Q-Learning



Training Process

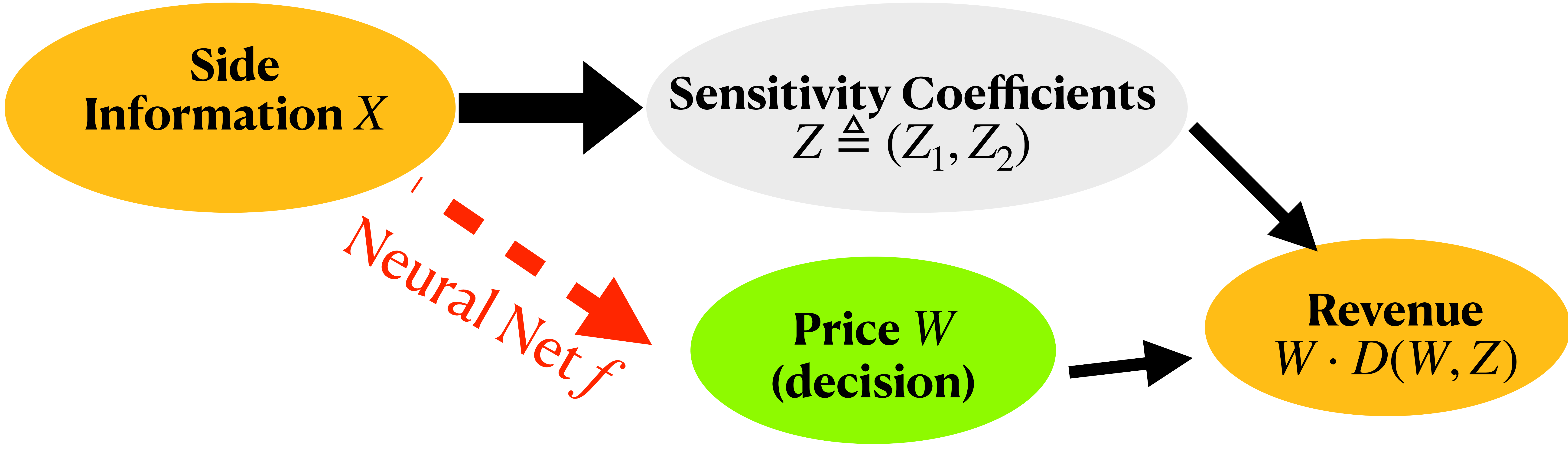
Data-Driven Personalized Pricing

- Linear demand model:

$$D(w, z) = z_1 w + z_2$$

Price

Sensitivity coefficients



$$\inf_f \ \mathbb{E}_{(X,Z)\sim\mathbb{P}^0} \left[-f(X) \cdot D(f(X), Z) \right] \qquad \text{(Nominal Problem)}$$

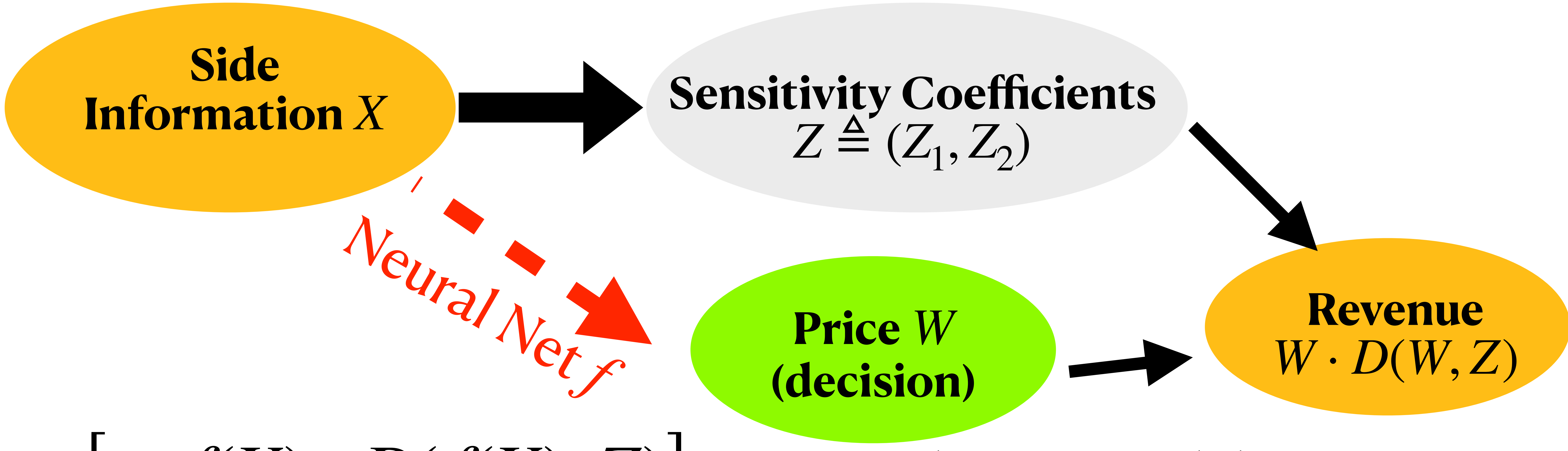
Numerical Study: Robust Contextual Learning

- Linear demand model:

$$D(w, z) = z_1 w + z_2$$

Price

Sensitivity coefficients



$$\inf_{f} \sup_{\mathbb{P}:\, \mathscr{C}_\infty(\mathbb{P},\mathbb{P}^0)\le\rho} \mathbb{E}_{(X,Z)\sim\mathbb{P}} \left[-f(X) \cdot D(f(X), Z) \right] \quad \text{(Robust Problem)}$$

$$= \inf_{f} \left\{ \mathbb{E}_{x\sim\mathbb{P}^0} \left[\sup_{x'\in\mathbb{B}_\rho(x)} \mathbb{E}_{z\sim\mathbb{P}^0(\cdot|X=x)}[-f(x')\cdot D(f(x'), z)] \right] \right\}$$

Yang, J., Zhang, L., Chen, N., Gao, R., & Hu, M. (2022). Decision-making with side information: A causal transport robust approach. *Optimization Online*.

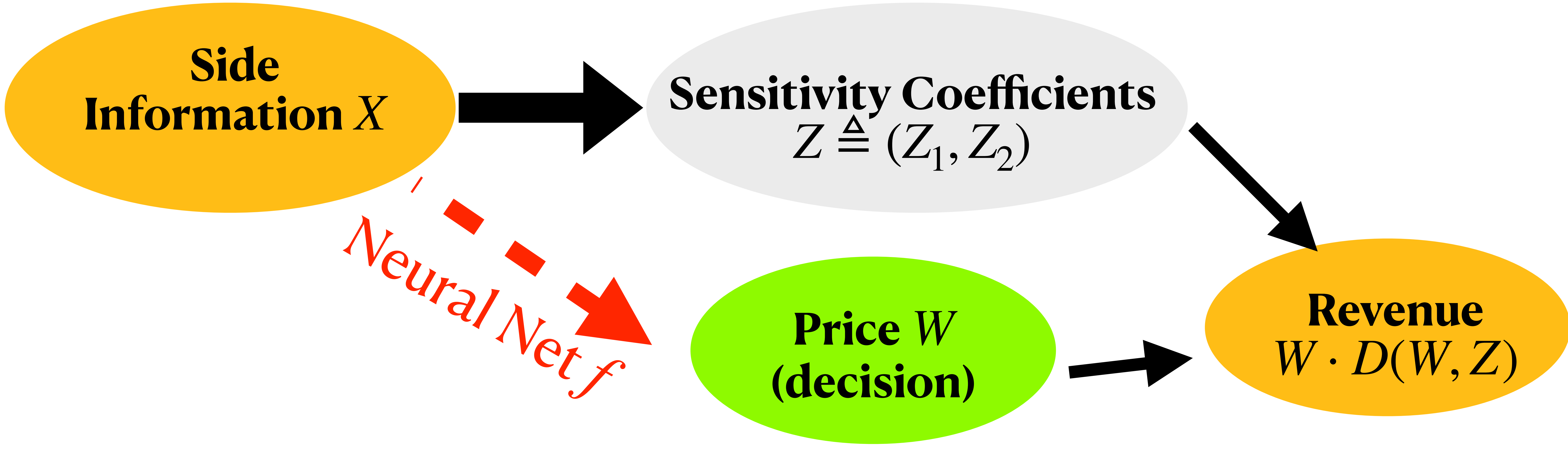
Numerical Study: Robust Contextual Learning

- Linear demand model:

$$D(w, z) = z_1 w + z_2$$

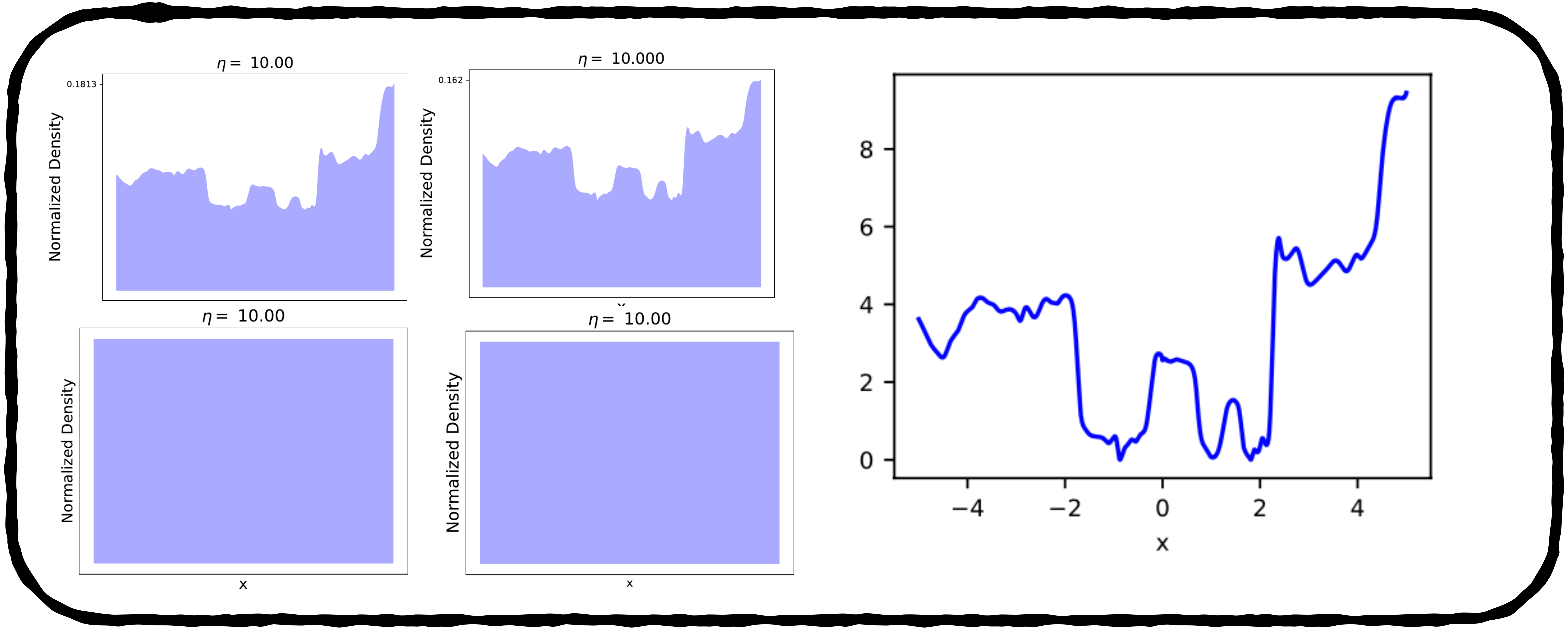
Price

Sensitivity coefficients



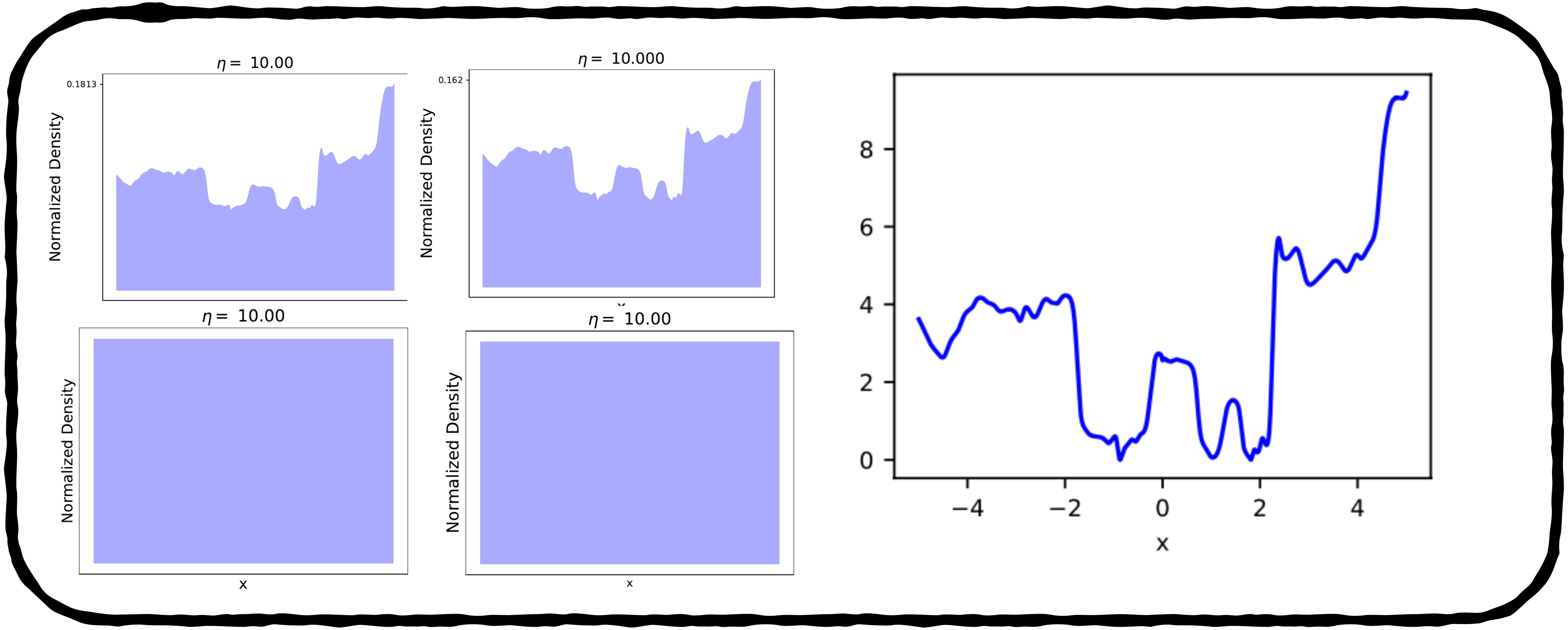
Conclusion

- f-divergence regularization for adversarial robust learning (∞-Wasserstein DRO)



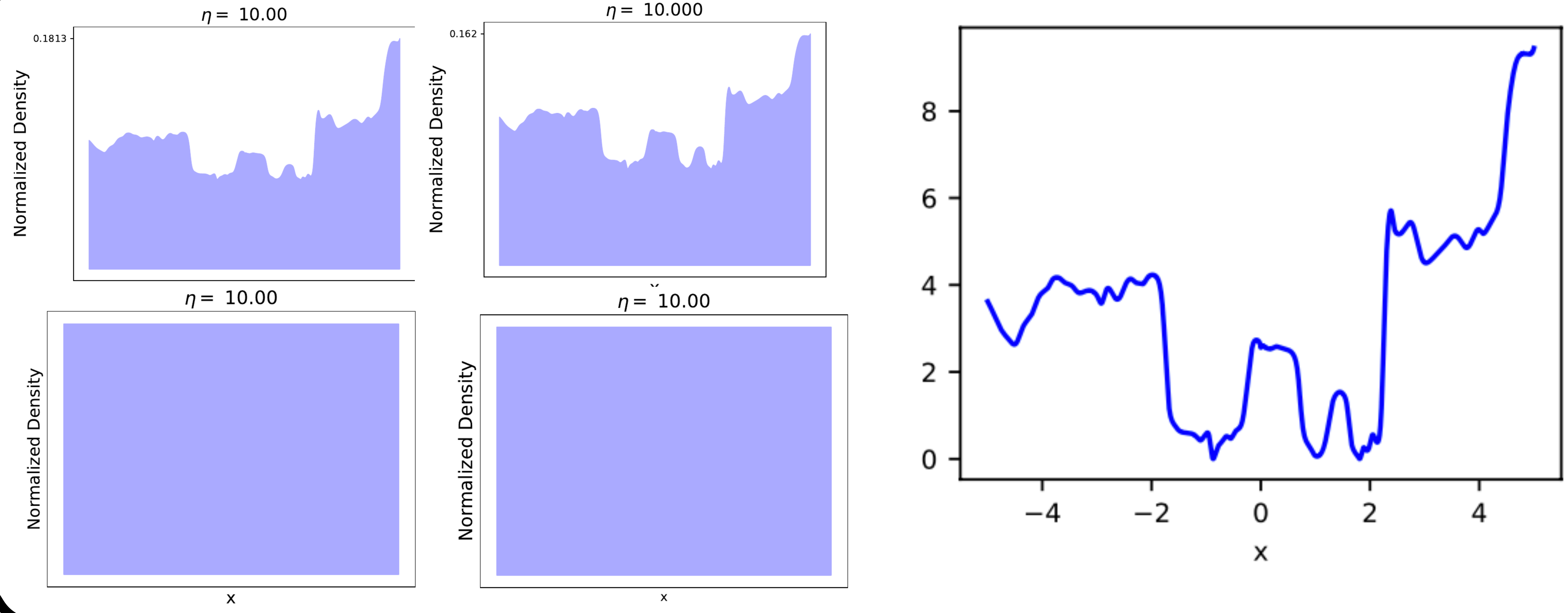
Conclusion

- f-divergence regularization for adversarial robust learning (∞-Wasserstein DRO)



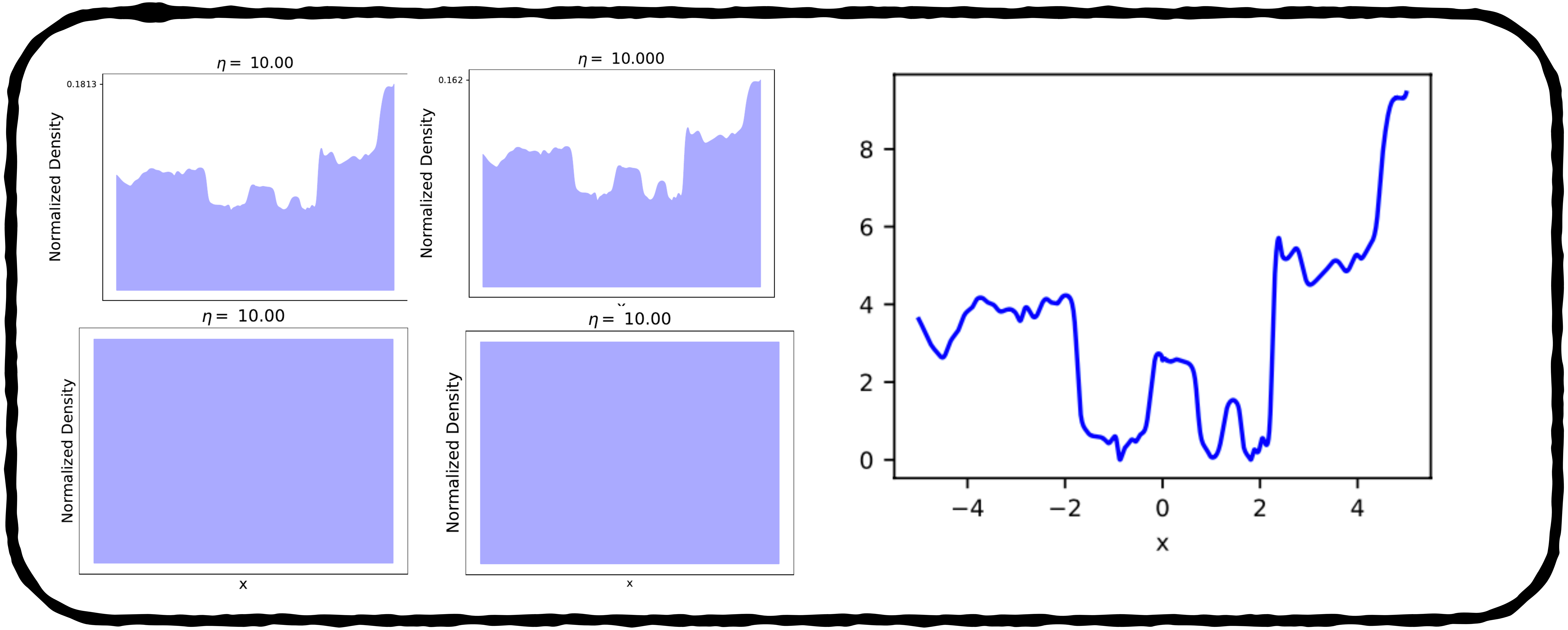
Conclusion

- f-divergence regularization for adversarial robust learning (∞-Wasserstein DRO)



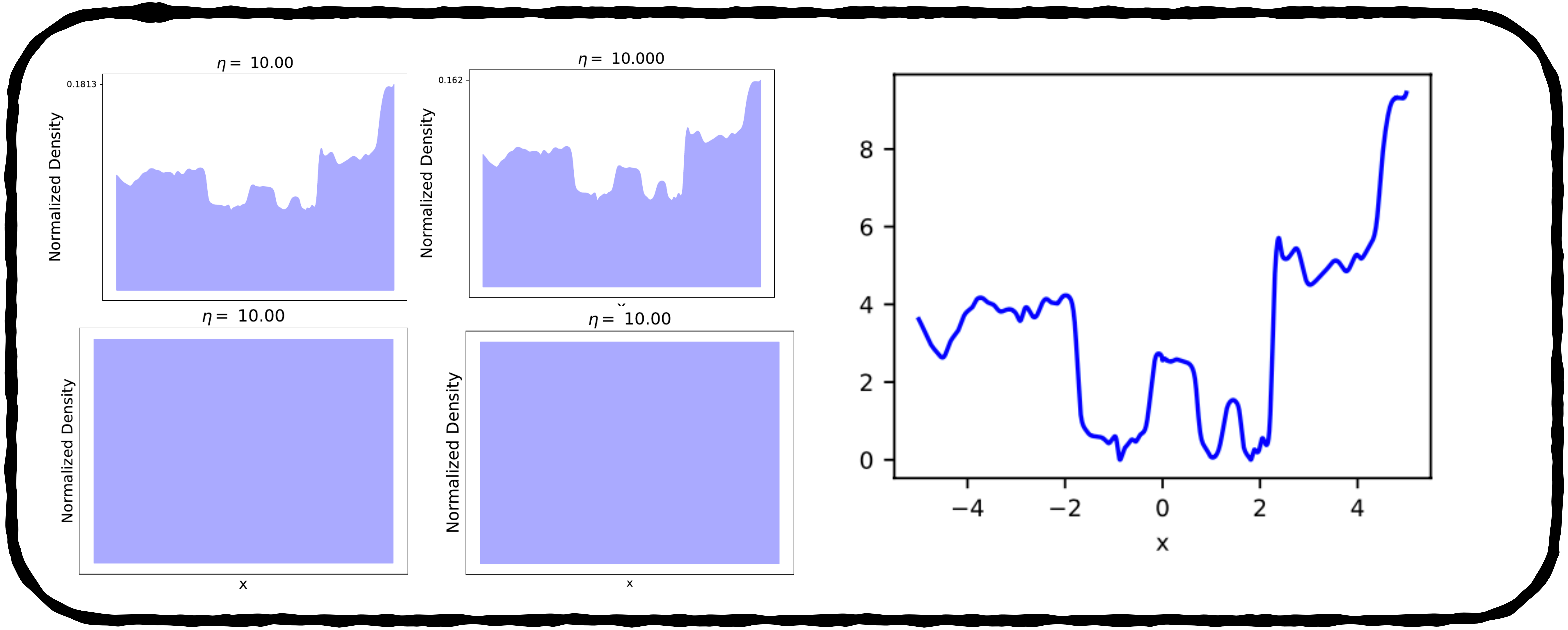
Conclusion

- f**-divergence regularization for adversarial robust learning (∞-Wasserstein DRO)**



Conclusion

- f-divergence regularization for adversarial robust learning (∞-Wasserstein DRO)



Conclusion

- f-divergence regularization for adversarial robust learning (∞-Wasserstein DRO)
- Efficient Algorithm using Multi-level Monte Carlo Sampling

Algorithm	Loss	Choice of Divergence	Complexity
Random Sampling	Convex/Nonconvex Smooth	Arbitrary/KL-Divergence	$\tilde{O}(\delta^{-2})/\tilde{O}(\delta^{-4})$

Conclusion

- f-divergence regularization for adversarial robust learning (∞-Wasserstein DRO)
- Efficient Algorithm using Multi-level Monte Carlo Sampling
- Regularization effects under different scaling regimes of ρ/η

Related References

- **Wang J**, Gao R, Xie Y (2024) Regularization for Adversarial Robust Learning. *arXiv preprint arXiv:2109.11926*
- **Wang J**, Gao R, Xie Y (2021) Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926 [Major Revision at Operations Research]*
- Hu Y, **Wang J**, Xie Y, Krause A, Kuhn D (2023) Contextual stochastic bilevel optimization. *Advances in Neural Information Processing Systems* 36
- Hu Y, **Wang J**, Chen X, He N (2024) Multi-level Monte-Carlo Gradient Methods for Stochastic Optimization with Biased Oracles. *arXiv preprint arXiv:2408.11084*
- Sinha A, Namkoong H, Volpi R, Duchi J (2018) Certifying some distributional robustness with principled adversarial training. *ICLR 2018*
- Gao R, Chen X, Kleywegt AJ (2022) Wasserstein distributionally robust optimization and variation regularization. *Operations Research*