Policy Learning under Biased Sample Selection

Roshni Sahoo
Joint work with Lihua Lei, Stefan Wager
Stanford University

Cornell Causal Reading Group Meeting 10/5/23

Practitioners often use data from a **study (train) population** to learn decision rules that can be deployed on a **target (test) population**.

However, the study population may differ from the target population due to **sampling bias**.

Examples:

- Evaluations of educational interventions [Bell et al., 2016].
- Clinical trials for anti-depressants [Wang et al., 2018].

Outline

1. Supervised learning under biased sampling with a **minimax loss criterion**.

2. Policy learning under biased sample selection with a **maxmin welfare** and **minimax regret criteria**.

1. Supervised Learning

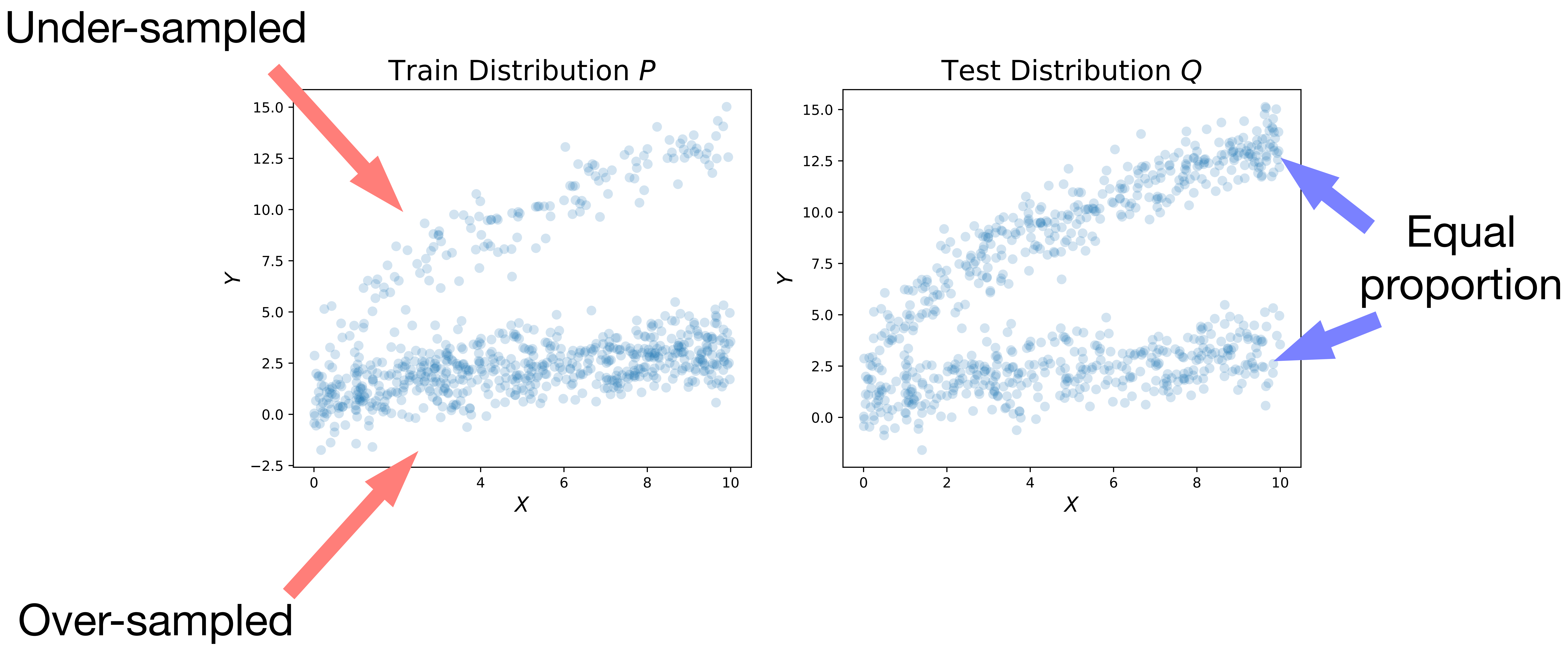
Machine learning model $h : \mathscr{X} \to \mathscr{Y}$

How to learn a good h?

Empirical Risk Minimization (ERM)! Given train distribution P, loss function L

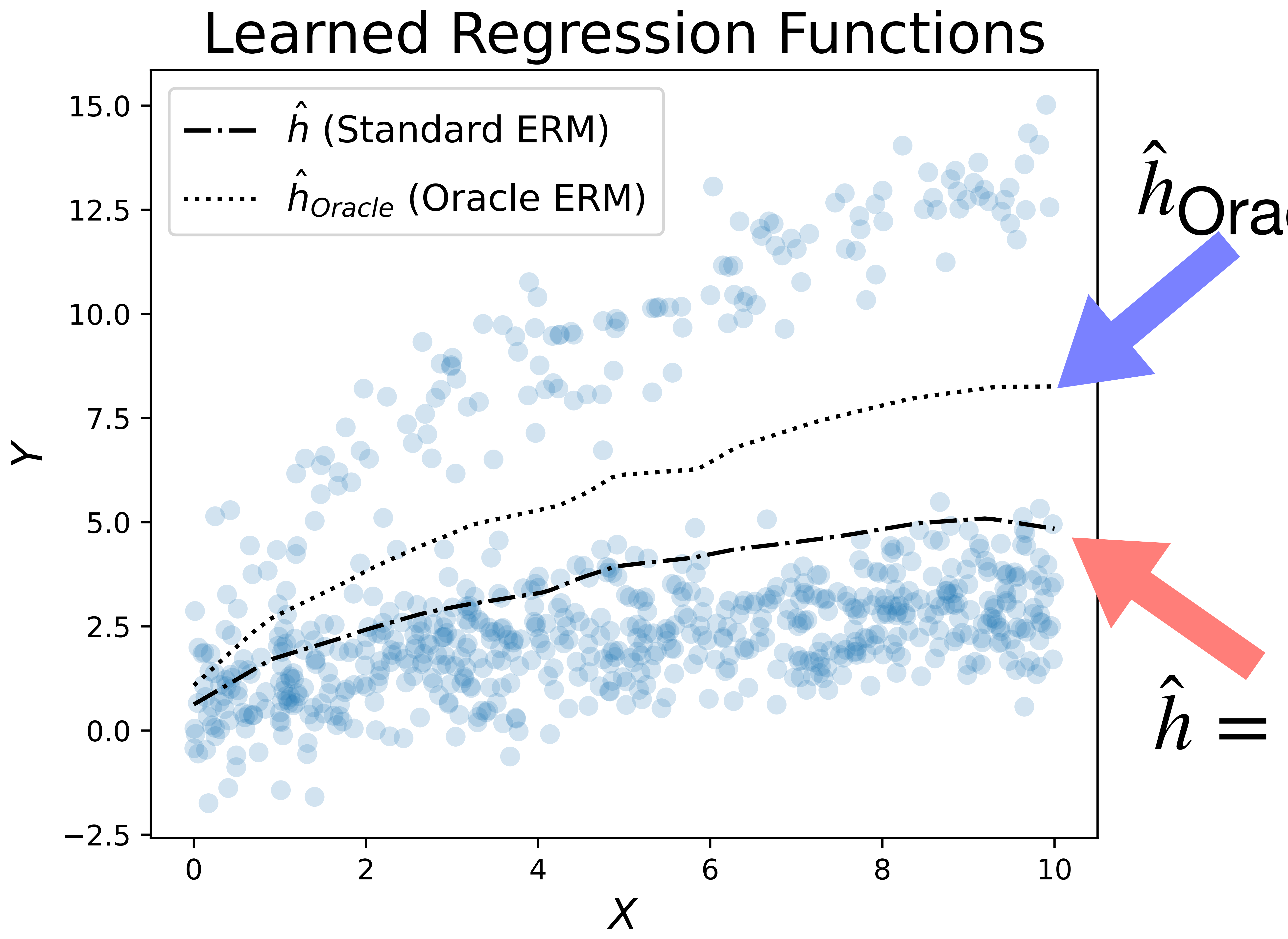
$$\hat{h} = \text{argmin}_h \hat{\mathbb{E}}_P[L(h(X), Y)]$$

However, under sampling bias, the test distribution Q may differ from P.


Under-sampled
Over-sampled
Train Distribution P
Test Distribution Q
Equal proportion
X
Y

Standard ERM is suboptimal in the presence of sampling bias.

$$L(h(X), Y) = (Y - h(X))^2$$



$$\hat{h}_{\text{Oracle}} = \text{argmin}_h \hat{\mathbb{E}}_Q[L(h(X), Y)]$$

$$\hat{h} = \text{argmin}_h \hat{\mathbb{E}}_P[L(h(X), Y)]$$

Types of Sampling Bias

Selection Mechanism: Every unit i in test population Q is associated with $S_i \in \{0,1\}$, which indicates whether unit i is in the train population.

Sample selection probability of unit i is given by $\mathbb{E}[S_i \mid X_i, Y_i]$.

1) Ignorable Selection: Selection probabilities only depend on observable attributes

$$\mathbb{E}[S_i \mid X_i, Y_i] = \mathbb{E}[S_i \mid X_i].$$

2) Non-ignorable Selection: Selection probabilities depends on **observables and unobservables**!

Examples

- Consider a medical study that aims to recruit participants.
  - Younger people may be more likely to participate than older people.

    Ignorable Selection

  - People who live farther from a hospital are less likely to participate.

    Non-ignorable Selection

Setting

Denote the **full test distribution** $(X, Y, S) \sim F$, where X are covariates, Y are outcomes, $S \in \{0,1\}$ are **unobservable**, binary selection indicators.

Our ideal model minimizes the loss under the true test distribution:

$$h^*_Q = \text{argmin}_h \mathbb{E}_Q[L(h(X), Y)] \quad Q = F_{X,Y}$$

Challenge: We cannot access i.i.d. samples from Q.
We can only access $P = F_{X,Y|S=1}$.

Biased Sampling

Assumption (Γ*-biased sampling*): The strength of the sampling bias is controlled by $\Gamma \geq 1$,

$$\Gamma^{-1} \leq \mathbb{E}_F[S \mid X, Y] / \mathbb{E}_F[S \mid X] \leq \Gamma.$$

Interpretations**:**

1) X can affect the probability of sample selection arbitrarily much BUT we limit the amount of **unexplained variation** in this probability.

2) Can think of Γ as governing the level of ignorable selection.

Minimax Learning under Biased Sampling

Challenge: The true test distribution is unknown and given P, there are many possible test distributions under Γ-biased sampling.

Let $\mathcal{S}_\Gamma(P, Q_X)$ be the family of test distributions that

1) Can generate P via Γ-biased sampling,
2) Have covariate distribution Q_X.

Idea: Apply DRO (distributionally robust optimization)! [Ben-Tal et al., 2013]

For any Q_X, we aim to solve

$$\text{argmin}_h \sup_{Q \in \mathcal{S}_\Gamma(P, Q_X)} \mathbb{E}_Q[L(h(X), Y)].$$

Bottom Line Up Front

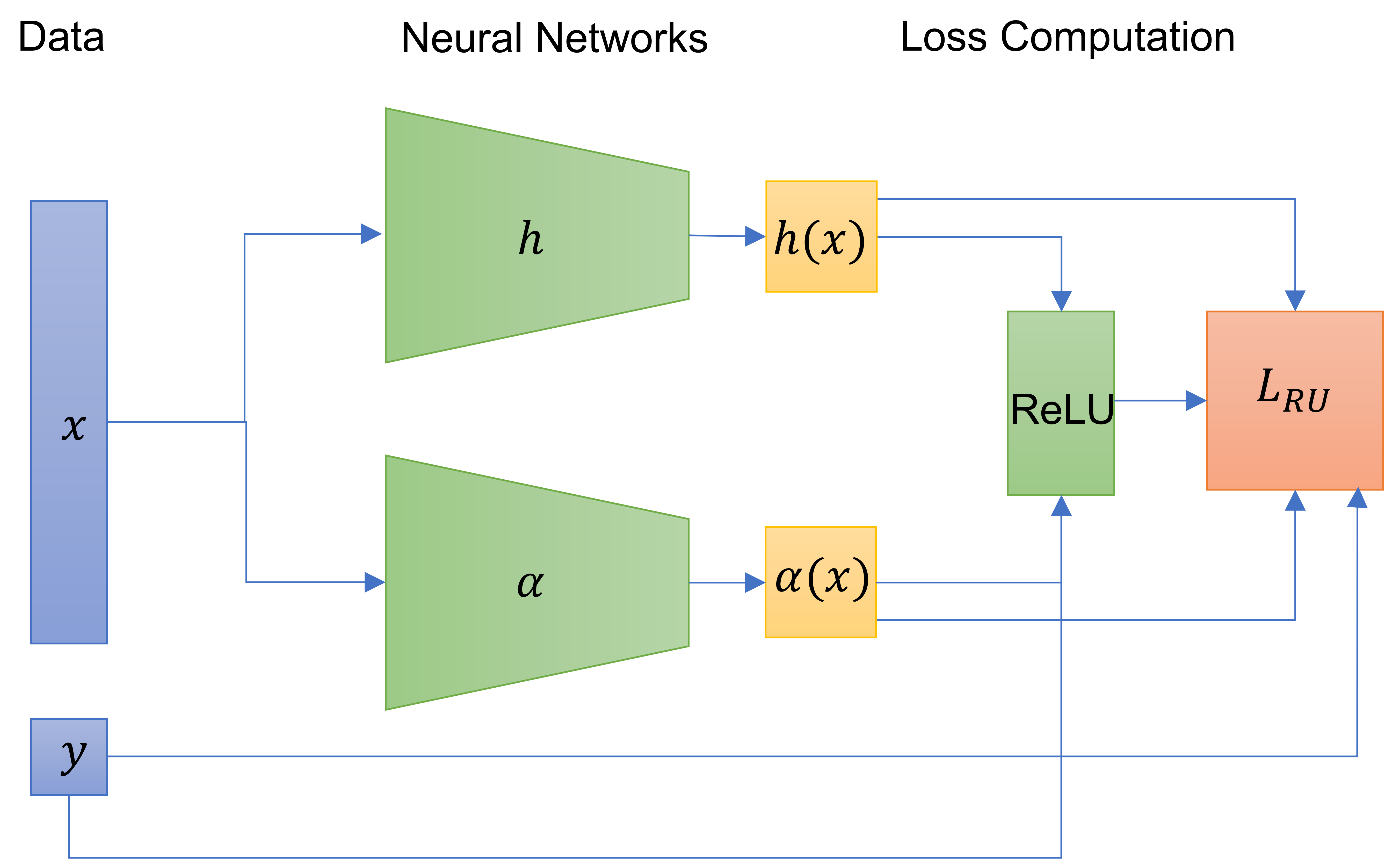
We propose a procedure called **RU Regression** that solves our worst-case risk minimization problem for **any** Q_X such that $Q_X \ll P_X$.

Given a loss function L and $\Gamma > 1$, we define the Rockafellar-Uryasev (RU) loss

$$L_{\text{RU}}^{\Gamma}(z, a, y) = \Gamma^{-1} \cdot L(z, y) + (1 - \Gamma^{-1}) \cdot a + (\Gamma - \Gamma^{-1}) \cdot (L(z, y) - a)_{+}.$$

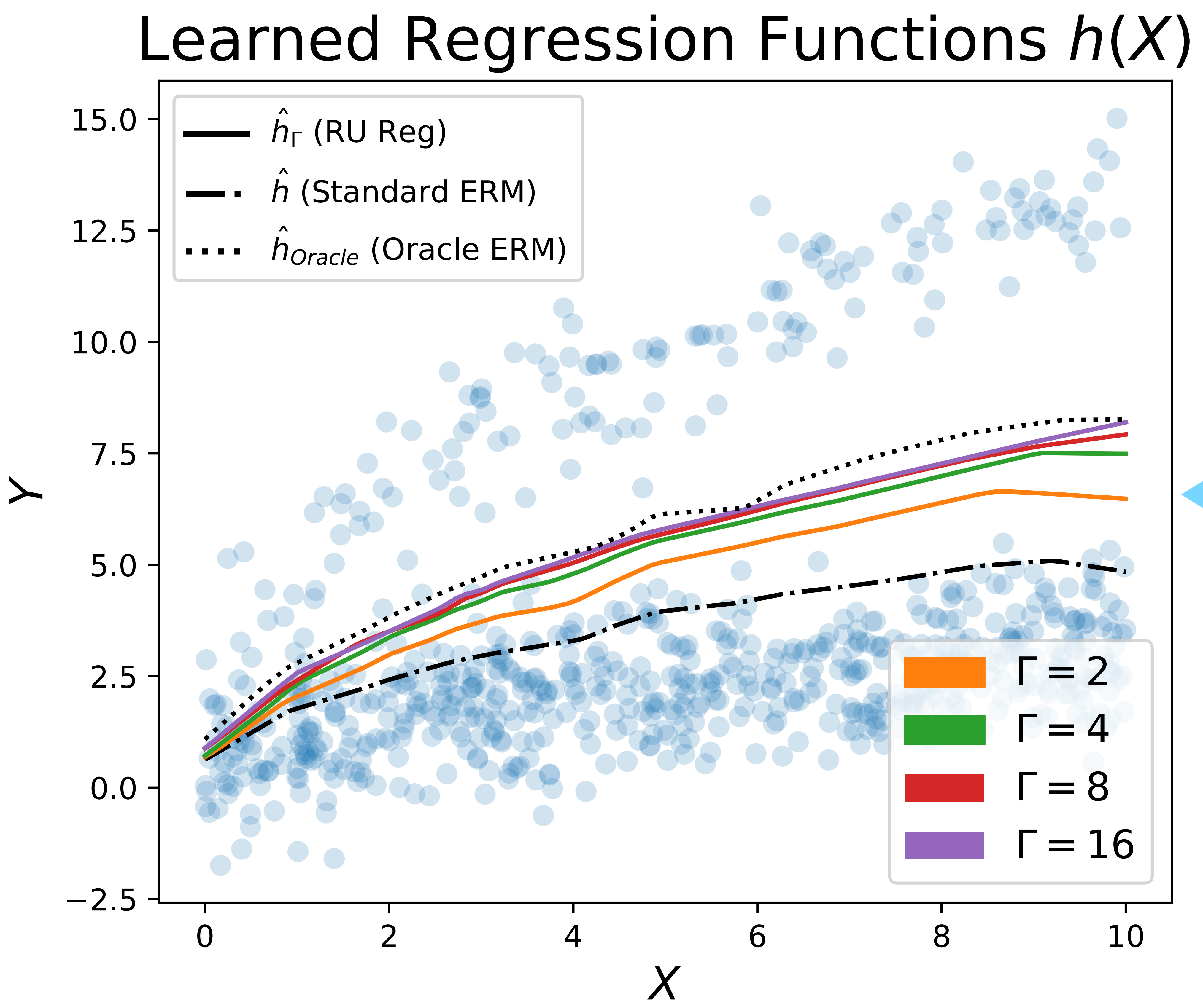
RU Regression solves

$$(h_{\Gamma}^{*}, \alpha_{\Gamma}^{*}) \in \text{argmin}_{(h,\alpha)\in L^2(P_X,\mathcal{X})\times L^2(P_X,\mathcal{X})} \mathbb{E}_P[L_{\text{RU}}^{\Gamma}(h(X), \alpha(X), Y)].$$



Jointly train two neural networks, one for each of h and α, using the RU loss with a standard optimization algorithm like SGD.

Back to the Toy Example



Ours

Some intuition on where RU Regression comes from…

Another way to express the robustness set:

$$\mathcal{S}_\Gamma(P, Q_X) = \left\{ Q \,\middle|\, \Gamma^{-1} \leq \frac{dQ_{Y|X=x}(y)}{dP_{Y|X=x}(y)} \leq \Gamma \quad \forall x, y, \text{ and } F_X = Q_X \right\}.$$

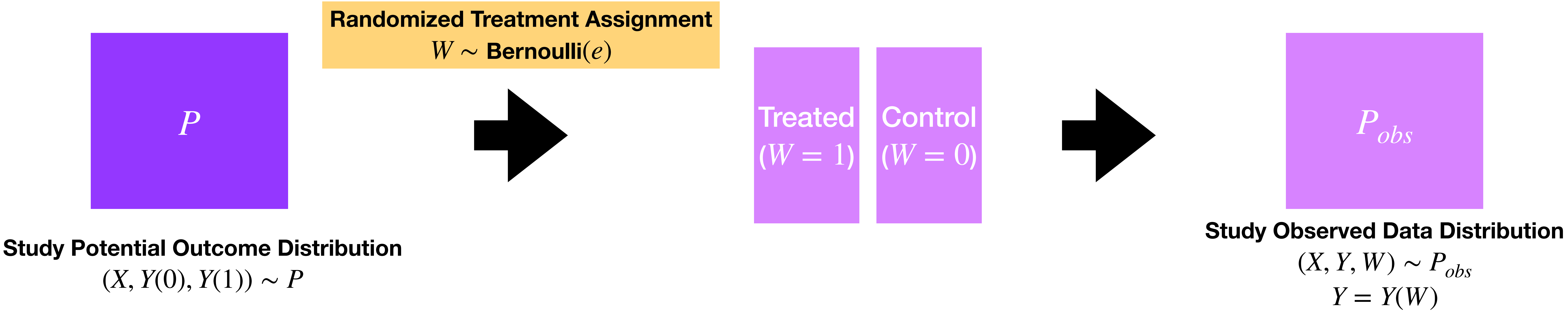
Conditionally on x, the worst-case distribution **upweights examples with high loss** by Γ and **downweights examples with low loss** by Γ^{-1}.

$$dQ^*_{Y|X=x}(y) = \begin{cases} \Gamma \cdot dP_{Y|X=x}(y) & \text{if } L(h(x), Y) \geq q_{\eta(\Gamma)}(L(h(x), Y)) \\ \Gamma^{-1} \cdot dP_{Y|X=x}(y) & \text{o.w.} \end{cases}.$$

The function $\alpha(x)$ in RU Regression implicitly learns the threshold $q_{\eta(\Gamma)}(L(h(x), Y))$ where the worst-case distribution switches from unweighting to downweighting for each x.

2. Policy Learning

Refresher on Policy Learning (No Sampling Bias)



Aim to learn a policy $\pi : \mathscr{X} \to \{0,1\}$ from policy class Π that maximizes the welfare

$$V_P(\pi) = \mathbb{E}_P[Y(\pi(X))].$$

When Π is unconstrained, the optimal policy is

$$\pi_{non-robust}(X) = \mathbb{I}(\tau(X) \geq 0),$$

where τ is the **CATE function**: $\tau(x) = \mathbb{E}_P[Y(1) - Y(0) \mid X = x]$.

Can think of learning policies from RCT data as an **offline contextual bandit problem.**

Policy Learning = Supervised Learning?

Recent works demonstrate that we can learn policies through (modified) supervised learning algorithms (Kitagawa & Tetenov, 2018; Athey & Wager, 2021; Mbakop & Tabord-Meehan, 2021).

What about policy learning under **biased sample selection**?

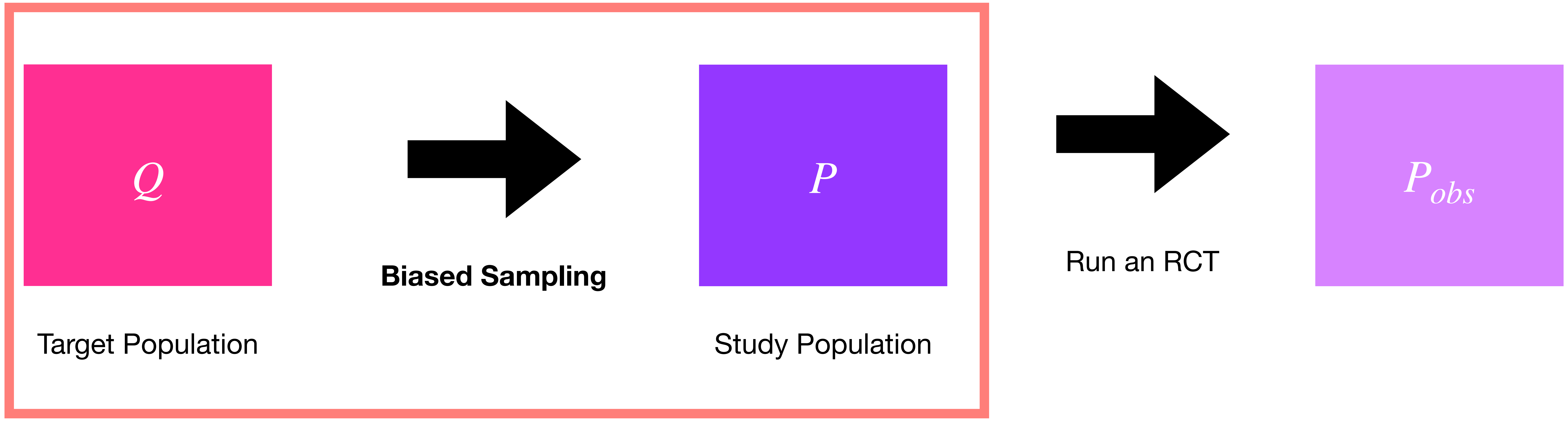
Challenge #1: Reducing to supervised learning is delicate.

Challenge #2: Maximin welfare is generally not considered a good criterion for treatment choice problem; minimax regret is often preferred (Savage, 1951; Manski, 2011).

Data-Generating Process under Sampling Bias

P, Q are potential outcome distributions, i.e. distributions over $(X, Y(0), Y(1))$.

P_{obs} is an observed data distribution, i.e. a distribution over (X, Y, W).



Assumption: RCT is well-executed.

Setting

Denote the **full target distribution** $(X, Y(0), Y(1), S) \sim F$.

1. The target potential outcome distribution Q is $F_{X,Y(0),Y(1)}$.
2. The study potential outcome distribution P is $F_{X,Y(0),Y(1)|S=1}$.
3. We run an RCT on P to generate P_{obs}.

We are interested in learning a policy that attains high welfare under Q:

$$V_Q(\pi) = \mathbb{E}_Q[Y(\pi(X))].$$

We assume S obeys Γ-biased sampling:

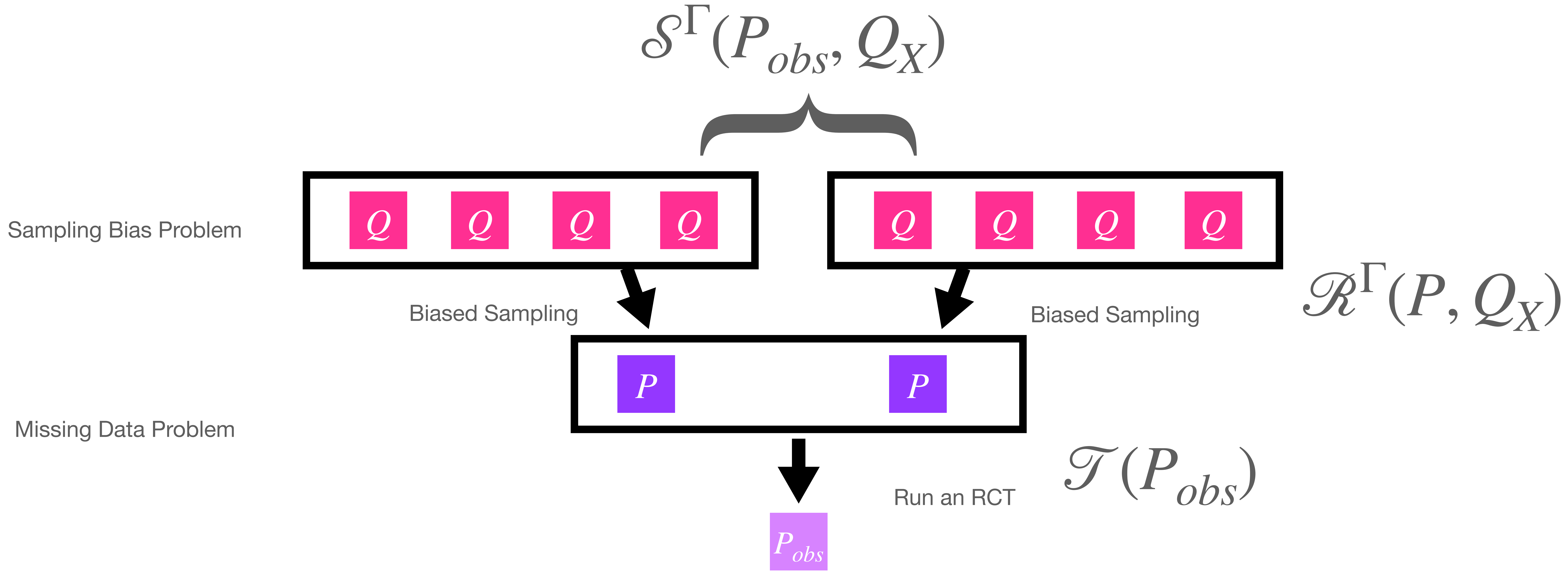
$$\Gamma^{-1} \leq \mathbb{E}_F[S \mid X, Y(0), Y(1)]/\mathbb{E}_F[S \mid X] \leq \Gamma.$$

Biased Sampling to RCT

If we has access to the study potential outcome distribution P, we could proceed as before in the supervised learning case.

However, we only have access to P_{obs}, so we must define our robustness set as $\mathcal{S}^{\Gamma}(P_{obs}, Q_X)$.

Robustness Set for Policy Learning



How to measure performance?

Many possible objectives to consider:

Max-min [Adjaho and Christensen, 2022, Mu et. al., 2021, Savage 1951, Si et. al, 2022, Wald 1950]

$$\sup_{\pi \in \Pi} \inf_{Q \in \mathcal{S}_\Gamma(P_{obs}, Q_X)} \mathbb{E}_Q[Y(\pi(X))].$$

Max-min gain over a baseline [Ben-Michael et. al. 2021, Kallus and Zhou et. al. 2021]

$$\sup_{\pi \in \Pi} \inf_{Q \in \mathcal{S}_\Gamma(P_{obs}, Q_X)} \mathbb{E}_Q[Y(\pi(X))] - \mathbb{E}_Q[Y(\pi_0(X))].$$

Minimax regret [Manski 2004, Savage 1951]

$$\inf_{\pi \in \Pi} \sup_{Q \in \mathcal{S}_\Gamma(P_{obs}, Q_X)} R_Q(\pi), \text{ where } R_Q(\pi) = \sup_{\pi' \in \Pi} \mathbb{E}_Q[Y(\pi'(X))] - \mathbb{E}_Q[Y(\pi(X))].$$

Preliminaries

Policy class Π - unconstrained, binary-valued functions.

Our identification results depend on the **conditional value-at-risk (CVaR)** of the outcomes. The $\eta -$ CVaR of a random variable Z is given by

$$\text{CVaR}_{\eta}(Z) = \mathbb{E}[Z \mid Z \geq q_{\eta}(Z)],$$

where $q_{\eta}(Z)$ is the η-th quantile of Z.

Optimal Policies

Optimal policies of these objectives are identifiable under P_{obs} , and we have **closed-form expressions** for them!

Max-min

$$\pi^*_{maxmin}(x) = \mathbb{I}(\tau(x) \geq H_\Gamma(x))$$

Max-min Gain

$$\pi^*_{gain}(x) = \mathbb{I}(\pi_0(x) = 0)\mathbb{I}(\tau(x) \geq H^+_\Gamma(x)) + \mathbb{I}(\pi_0(x) = 1)\mathbb{I}(\tau(x) \geq H^-_\Gamma(x))$$

Minimax Regret

$$\pi^*_{regret}(x) = \mathbb{I}(\tau(x) \geq (H^+_\Gamma(x) + H^-_\Gamma(x))/2)$$

Can think of $H_\Gamma(\cdot), H^+_\Gamma(\cdot), H^-_\Gamma(\cdot)$ as identifiable nuisance parameters that depend on CVaR$_{\zeta(\Gamma)}(Y(w) \mid X = x)$, CVaR$_{\zeta(\Gamma)}(-Y(w) \mid X = x)$ for $w \in \{0,1\}$, where $\zeta(\Gamma) = \dfrac{1}{\Gamma + 1}$.

How to learn the optimal policies?

Naive two-stage approach:

1)Estimate $\tau(\cdot), H_\Gamma(\cdot), H_\Gamma^+(\cdot), H_\Gamma^-(\cdot)$ using data from P_{obs}.

2)Plug them into closed-form expressions from for the optimal policies

Can we learn the optimal policies directly?

Yes! We can learn the optimal max-min and max-min gain policies **in one step** using RU Regression (does not require separate estimation of nuisance parameters!).

Loss Minimization Approach

Theorem: We can specify $v_{maxmin}(z; x, y, w)$ so that RU Regression yields $\pi^*_{\Gamma,maxmin}(x)$.
Similarly, we can specify $v_{gain}(z; x, y, w)$ so that RU Regression yields $\pi^*_{\Gamma,gain}(x)$.

1) Given v and $\Gamma > 1$, define the RU loss [Sahoo et. al., 2022]

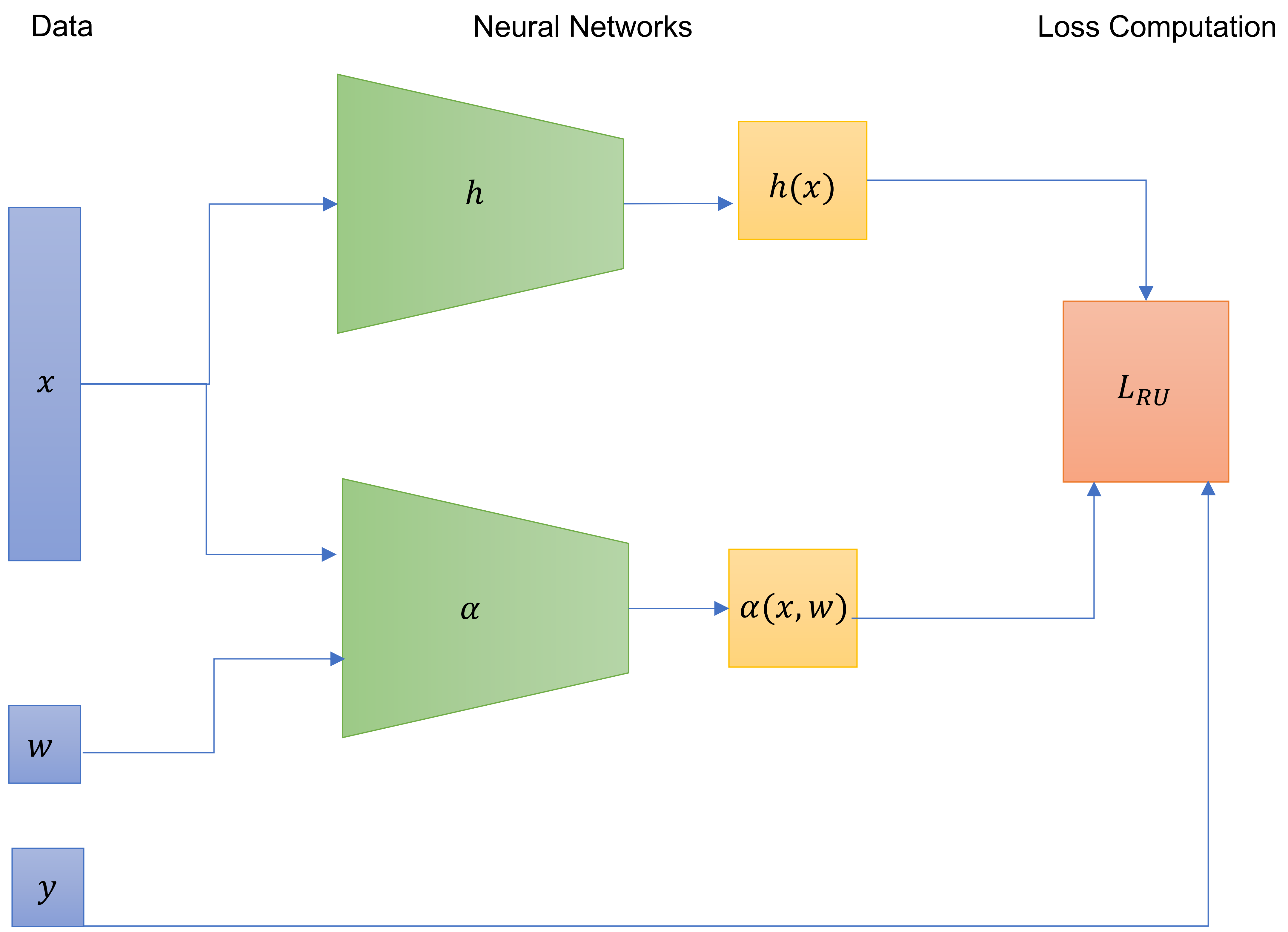
$$L^{\Gamma}_{RU}(z, a; x, y, w) = \Gamma^{-1}(-v(z; x, y, w)) + (1 - \Gamma^{-1}) \cdot a + (\Gamma - \Gamma^{-1})(-v(z; x, y, w) - a)_+$$

2) Solve the RU Regression problem.

$$(h_\Gamma, \alpha_\Gamma) \in \text{arginf}_{(h,\alpha)\in\mathcal{H}\times\mathcal{A}} \mathbb{E}_P[L_{RU}(h(X), \alpha(X, W), Y)].$$

3) Return the policy $\mathbb{I}\left(h_\Gamma(x) \geq \frac{1}{2}\right)$.

RU Regression for Policy Learning



Super similar to supervised learning case, except

1. Auxiliary function α takes in X, W.
2. Restrict the function h to output [0,1] with sigmoid activation.

Conclusions

1. In many settings, we need to learn decision rules from data that may be a biased sample from the population of interest.

2. We considered methods for learning with robust guarantees under biased sample selection.

3. The learning criterion we use matters; and **non-robust**, **maxmin**, **maxmin gain**, and **minimax regret** decision rules are generally not the same.

4. RU Regression is a simple and practical avenue to learning decision rules from biased data using deep learning.

Happy to chat and collaborate :)

rsahoo@stanford.edu
roshni714.github.io