Off-Policy Evaluation:
A Distributionally Robust Approach

Speaker: Jie Wang

August 12, 2020

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Outline

- Distributionally Robust Optimization
  - Tractable formulation, history, theory
- A Recent Application in Off-policy Policy Evaluation
  - Tractable formulation, theory, extensions
- Summary

The talk involves contributions from:
Prof. Rui Gao (UT Austin), Prof. Hongyuan Zha (CUHK-SZ), Prof. Xinyun Chen (CUHK-SZ)

Backgroud about Distributionally Robust Optimization: Tractable Formulation and Statistics

Introduction to Stochastic Optimization

Consider the *stochastic optimization problem* as follows:

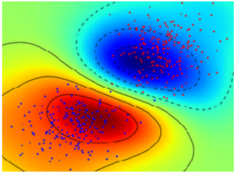
$$\text{maximize}_{x \in \mathcal{X}} \quad \mathbb{E}_{\zeta \sim \mathbb{P}}[h(x, \zeta)] \qquad (1)$$

with $\mathcal{X}$ being convex.

Applications:

Supply Chain Mgmt.

Portfolio Mgmt.



Machine Learning

Introduction to Stochastic Optimization

Consider the *stochastic optimization problem* as follows:

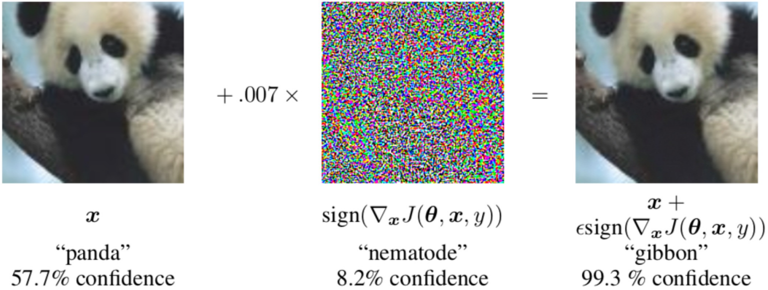
$$\text{maximize}_{x \in \mathcal{X}} \qquad \mathbb{E}_{\zeta \sim \mathbb{P}}[h(x, \zeta)] \tag{2}$$

with $\mathcal{X}$ being convex.

- **Prospective**
  - Expected value is a good measure of performance;
  - Solve by *sample average approximation* (SAA).
- **Challenge**
  - Difficult to know the exact distribution of ζ;
  - Solution can be risky by SAA;
  - SAA may result in sub-optimal solutions.

Risky: Stochastic Optimization with Noises

Adversarial attacks for classification problem [1]:



香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

[1] Ian Goodfellow 2015

Picture for Gibbon

Sub-optimality: the Optimizer's Curse

- Suppose $\hat{\mathbb{P}}_n$ is an unbiased estimator of $\mathbb{P}$:

$$\mathbb{E}_{\otimes}[\hat{\mathbb{P}}_n] = \mathbb{P}.$$

- The optimization results by SAA approach, i.e., $\mathcal{R}_{\text{SAA}}$, tend to be *pessimistic biased*:

$$\begin{aligned}\mathbb{E}_{\otimes}\left[\mathcal{R}_{\text{SAA}}\right] &= \mathbb{E}_{\otimes}\left[\max_{x\in\mathcal{X}} \mathbb{E}_{\zeta\sim\hat{\mathbb{P}}_n}[h(x,\zeta)]\right] \\ &\geq \max_{x\in\mathcal{X}} \mathbb{E}_{\otimes}\left[\mathbb{E}_{\zeta\sim\hat{\mathbb{P}}_n}[h(x,\zeta)]\right] \\ &= \mathcal{R}_{\text{true}}.\end{aligned}$$

Testing Errors for Supervised Learning

Consider the supervised learning problem:

$$\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim \mathbb{P}_{\text{true}}} [\ell(f(x), y)]$$

People tackle this problem by the SAA approach:

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim \hat{\mathbb{P}}_n} [\ell(f_\theta(x), y)], \quad \text{where } \hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(x_i, y_i)}.$$

Decomposition of errors in machine learning [2]:

$$\text{Testing Error} = \begin{cases} \text{Generalization Error (Distributional Uncertainty)} \\ \text{Representation Error} \\ \text{Optimization Error} \end{cases}$$

[2] Ruoyu Sun, Optimization for deep learning: theory and algorithms (2019).

Motivation for DRO: Distributional Uncertainty

- Out-of-Sample performance of SAA:

$$\sup_x \left| \mathbb{E}_{\zeta\sim\mathbb{P}}[h(x,\zeta)] - \mathbb{E}_{\zeta\sim\hat{\mathbb{P}}_n}[h(x,\zeta)] \right|$$

$$\leq C_1\sqrt{\frac{\mathrm{Var}[h(x,\zeta)]}{n}} + C_2 \cdot \frac{1}{n}\mathbb{E}\left[\sup_{x\in\mathcal{X}} \sum_{i=1}^{n} \sigma_i h(x,\zeta_i)\right].$$

- Distributional Uncertainty: it is difficult to obtain $\mathbb{P}$, but related samples or statistical information are available.

How to develop an algorithm that cooperates the distributional uncertainty?

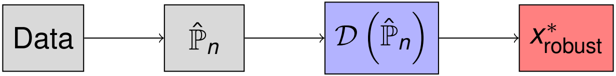
香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Distributionally Robust Optimization

Distributionally Robust Optimization (DRO) model:

$$\text{maximize}_{x \in \mathcal{X}} \quad \min_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\zeta \sim \mathbb{P}}[h(x, \zeta)]$$

where $\mathcal{D}$ denotes a collection of distributions. We call it the ambiguity set.



Distributionally Robust Optimization

Distributionally Robust Optimization (DRO) model:

$$\text{maximize}_{x\in\mathcal{X}} \quad \min_{\mathbb{P}\in\mathcal{D}} \mathbb{E}_{\zeta\sim\mathbb{P}}[h(x,\zeta)]$$

where $\mathcal{D}$ denotes a collection of distributions. We call it the ambiguity set.

Guidance for choosing $\mathcal{D}$:

- Tractability (fast algorithm available);
- Statistical Theoretical Guarantees;
- Numerical Performance (compared with the benchmark cases, such as SAA).

History of DRO

- DRO is first introduced in the context of inventory control problem with a single random demand variable[3].
- DRO with moment bounds[4]:

$$\mathcal{D} = \left\{ \mathbb{P} \middle| \begin{array}{l} (\mathbb{E}_{\mathbb{P}}[\zeta] - \mu_0)^T \Sigma_0^{-1} (\mathbb{E}_{\mathbb{P}}[\zeta] - \mu_0) \leq \gamma_1 \\ \mathbb{E}_{\mathbb{P}}[(\zeta - \mu_0)(\zeta - \mu_0)^T] \preceq \gamma_2 \Sigma_0 \end{array} \right\}$$

- DRO with KL-divergence/f-divergence balls[5]:

$$\mathcal{D} = \left\{ \mathbb{P} \middle| D(\mathbb{P} \| \hat{\mathbb{P}}_n) \leq \gamma \right\},$$

where $D(\cdot, \cdot)$ can be the KL-divergence metric, or f-divergence metric.

[3] Scarf, H. (1958) A Min-Max Solution of an Inventory Problem.

[4] Erick Delage, Y. (2008) Distributionally Robust Optimization under Moment Uncertainty with Application to Data-Driven Problems

[5] Duchi (2016), Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach

Introduction to Wasserstein Distance

- We set the ambiguity set to be

$$\mathcal{D} = \left\{ \mathbb{P} : \ W(\mathbb{P}, \hat{\mathbb{P}}_n) \leq \delta \right\}$$

where $W(\cdot, \cdot)$ refers to the Wasserstein metric:

$$W(\mathbb{P}, \mathbb{Q}) = \sup_{g \in \mathrm{Lip}_1} \left| \int g(x) d\mathbb{P}(x) - \int g(x) d\mathbb{Q}(x) \right|$$

- Wasserstein distance is a *two-sample* formula, and for its approximation, we need samples from both $\mathbb{P}$ and $\mathbb{Q}$.
- If one of $\mathbb{P}$ or $\mathbb{Q}$ is given in an explicit density form, the Wasserstein distance is not convenient to use.

Comparison of Different Probability Metrics

- f-divergence is a *two-density* formula:

$$D_f(\mathbb{P}\|\mathbb{Q}) = \int_\Omega f(d\mathbb{P}/d\mathbb{Q})d\mathbb{Q};$$

- Wasserstein distance is a *two-sample* formula:

$$W(\mathbb{P},\mathbb{Q}) = \sup_{g\in \text{Lip}_1} \left| \int g(x)d\mathbb{P}(x) - \int g(x)d\mathbb{Q}(x) \right|.$$

- Stein discrepancy is a *one-sample-one-density* formula:

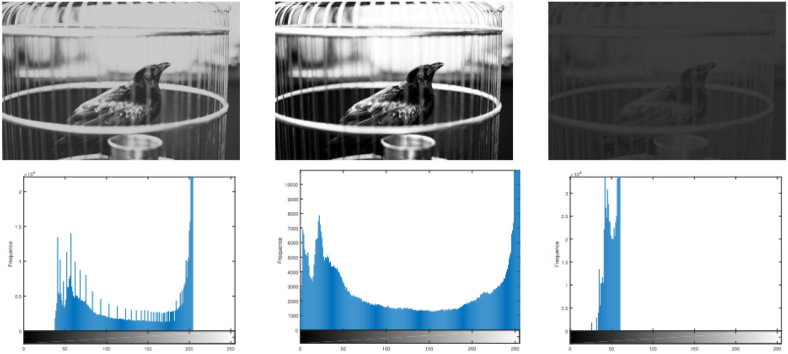
$$S(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \left| \int \mathcal{A}_{\mathbb{P}}[f(x)]d\mathbb{Q}(x) \right|$$

where $\mathcal{A}_{\mathbb{P}}[f(x)] = f(x)\nabla_x \log \mathbb{P}(x) + \nabla_x f(x)$.

Introduction to Wasserstein Distance

By the duality theory in LP,

$$W(\mathbb{P},\mathbb{Q}) = \inf_{\pi}\left\{\mathbb{E}_{\pi}\left[c(\zeta_1,\zeta_2)\right] : \begin{array}{l}\pi \text{ is a distribution of } \zeta_1 \text{ and } \zeta_2 \\ \text{with marginals } \mathbb{P} \text{ and } \mathbb{Q}\end{array}\right\}.$$



(a) Observed image with histogram ν

(b) True image with histogram μ_{true}

(c) Pathological image with histogram μ_{pathol}

FIGURE 1. Three images and their gray-scale histograms. For KL divergence, it holds that $I_{\phi_{KL}}(\mu_{true},\nu) = 5.05 > I_{\phi_{KL}}(\mu_{pathol},\nu) = 2.33$, while in contrast, Wasserstein distance satisfies $W_1(\mu_{true},\nu) = 30.70 < W_1(\mu_{pathol},\nu) = 84.03$.

Statistics Properties for DRO with Wasserstein Distance

Theorem 1

Consider the DRO problem

$$\hat{x}_n = \arg\max_{x \in \mathcal{X}} \left\{ \min_{\mathbb{P} \in \mathcal{D}_n} \mathbb{E}_{\zeta \sim \mathbb{P}}[h(x, \zeta)] \right\}$$

with $\mathcal{D}_n = \{\mathbb{P} : W(\mathbb{P}, \hat{\mathbb{P}}_n) \leq \delta_n\}$ and $\delta_n = O(1/\sqrt{n})$. The following properties hold:

- Asymptotic guarantee: $\mathbb{P}^\infty(\lim_{n \to \infty} \hat{x}_n = x^*) = 1$;
- Finite-sample guarantee: with high probability, $(R_{\text{robust}} - R_{\text{true}})_+ = O(1/n)$;
- Tractability: same complexity class as SAA.

Tractability of DRO with Wasserstein Distance

- The goal is to simplify the DRO problem

$$\min_{x \in \mathcal{X}} \left\{ \sup_{\mathbb{P} \in \mathcal{D}_n} \mathbb{E}[h(x, \zeta)] \right\}$$

Define $\ell(\zeta) := h(x, \zeta)$ for fixed x.

Tractability of DRO with Wasserstein Distance

- The goal is to simplify the DRO problem

$$\min_{x \in \mathcal{X}} \left\{ \sup_{\mathbb{P} \in \mathcal{D}_n} \mathbb{E}[h(x, \zeta)] \right\}$$

Define $\ell(\zeta) := h(x, \zeta)$ for fixed x.

- Reformulate the *worse-case expectation problem*:

$$\begin{aligned} \sup_{\mathbb{P}} \quad & \mathbb{E}_{\zeta \sim \mathbb{P}}[\ell(\zeta)] \\ \text{subject to} \quad & W(\mathbb{P}, \hat{\mathbb{P}}_n) \leq \delta_n \\ \text{where} \quad & W(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \Gamma(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{(\zeta_1, \zeta_2) \sim \pi} \left[c(\zeta_1, \zeta_2) \right]. \end{aligned}$$

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Tractability of DRO with Wasserstein Distance

Assume that the support of $\mathbb{P}$ is $\Xi := \{\zeta_1, \zeta_2, \dots, \zeta_K\}$:

$$
\begin{aligned}
\max_{\mathbb{P}} \quad & \sum_{k=1}^{K} \mathbb{P}(\zeta_k)\ell(\zeta_k) \\
\text{s.t.} \quad & \left\{ \begin{array}{ll}
\min\limits_{\pi \in \mathbb{R}_+^{K\times n}} & \sum_{k=1}^{K}\sum_{i=1}^{n} \pi_{k,i} c(\zeta_k, \hat{\zeta}_i) \\
\text{s.t.} & \sum_{k=1}^{K} \pi_{k,i} = \frac{1}{n},\ \forall i \in [n] \\
& \sum_{i=1}^{n} \pi_{k,i} = \mathbb{P}(\zeta_k),\ \forall k \in [K].
\end{array} \right\} \le \delta_n
\end{aligned}
$$

- Rewrite expectation in the form of summation;
- π is the joint distribution between $\mathbb{P}$ and $\hat{\mathbb{P}}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\hat{\zeta}_i}$.

Tractability of DRO with Wasserstein Distance

Replace the "min" in the constraint as "exist":

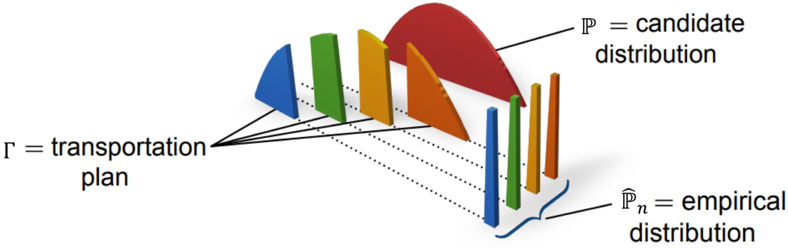
$$
\begin{aligned}
\max_{\mathbb{P}} \quad & \sum_{k=1}^{K} \mathbb{P}(\zeta_k)\ell(\zeta_k) \\
\exists \pi \in \mathbb{R}_+^{K\times n} \text{ such that} \quad & \sum_{k=1}^{K}\sum_{i=1}^{n} \pi_{k,i} c(\zeta_k, \hat{\zeta}_i) \le \delta_n \\
& \sum_{k=1}^{K} \pi_{k,i} = \frac{1}{n}, \ \forall i \in [n] \\
& \sum_{i=1}^{n} \pi_{k,i} = \mathbb{P}(\zeta_k), \ \forall k \in [K].
\end{aligned}
$$

Tractability of DRO with Wasserstein Distance

Reformulate the "feasibility problem" as a LP problem:

$$
\begin{aligned}
\max_{\mathbb{P},\pi\in\mathbb{R}_+^{K\times n}} \quad & \sum_{k=1}^{K} \mathbb{P}(\zeta_k)\ell(\zeta_k) \\
\text{s.t.} \quad & \sum_{k=1}^{K}\sum_{i=1}^{n} \pi_{k,i} c(\zeta_k,\hat{\zeta}_i) \leq \delta_n \\
& \sum_{k=1}^{K} \pi_{k,i} = \frac{1}{n}, \ \forall i \in [n] \\
& \sum_{i=1}^{n} \pi_{k,i} = \mathbb{P}(\zeta_k), \ \forall k \in [K].
\end{aligned}
$$

Representation of worse-case expectation problem



Tractability of DRO with Wasserstein Distance

- Eliminate $\mathbb{P}(\zeta_k)$ shown in the objective function:

$$
\begin{aligned}
\max_{\pi \in \mathbb{R}_+^{K \times n}} \quad & \sum_{k=1}^{K} \sum_{i=1}^{n} \pi_{k,i} \ell(\zeta_k) \\
\text{s.t.} \quad & \sum_{k=1}^{K} \sum_{i=1}^{n} \pi_{k,i} c(\zeta_k, \hat{\zeta}_i) \leq \delta_n \\
& \sum_{k=1}^{K} \pi_{k,i} = \frac{1}{n}, \ \forall i \in [n]
\end{aligned}
$$

Tractability of DRO with Wasserstein Distance

- Eliminate $\mathbb{P}(\zeta_k)$ shown in the objective function:

$$
\begin{aligned}
\max_{\pi \in \mathbb{R}_+^{K\times n}} \quad & \sum_{k=1}^{K}\sum_{i=1}^{n} \pi_{k,i}\ell(\zeta_k) \\
\text{s.t.} \quad & \sum_{k=1}^{K}\sum_{i=1}^{n} \pi_{k,i} c(\zeta_k, \hat{\zeta}_i) \leq \delta_n \\
& \sum_{k=1}^{K} \pi_{k,i} = \frac{1}{n}, \ \forall i \in [n]
\end{aligned}
$$

- By the duality theory for LP,

$$
\begin{aligned}
\inf_{\lambda \geq 0, s_i, i \in [n]} \quad & \lambda \delta_n + \frac{1}{n}\sum_{i=1}^{n} s_i \\
\text{s.t.} \quad & \ell(\zeta) - \lambda \cdot c(\zeta, \hat{\zeta}_i) \leq s_i, \ \forall i \in [n], \forall \xi \in \Xi
\end{aligned}
$$

Tractability of DRO with Wasserstein Distance

- Worse-case expecation problem is a 1-dimensional convex programming:

$$\sup_{\mathbb{P}:\ W(\mathbb{P},\hat{\mathbb{P}}_n)\leq\delta_n} \mathbb{E}_{\zeta\sim\mathbb{P}}[\ell(\zeta)]$$

$$= \inf_{\lambda\geq 0} \lambda\delta_n + \frac{1}{n}\sum_{i=1}^{n}\sup_{\zeta}\left(\ell(\zeta) - \lambda\|\zeta - \hat{\zeta}_i\|\right).$$

Tractability of DRO with Wasserstein Distance

- Worse-case expecation problem is a 1-dimensional convex programming:

$$\sup_{\mathbb{P}:\ W(\mathbb{P},\hat{\mathbb{P}}_n)\leq\delta_n} \mathbb{E}_{\zeta\sim\mathbb{P}}[\ell(\zeta)]$$

$$= \inf_{\lambda\geq 0} \lambda\delta_n + \frac{1}{n}\sum_{i=1}^{n}\sup_{\zeta}\left(\ell(\zeta) - \lambda\|\zeta - \hat{\zeta}_i\|\right).$$

- The DRO problem can be formulated as a single minimization:

$$\inf_{x\in\mathcal{X},\lambda\geq 0} \lambda\delta_n + \frac{1}{n}\sum_{i=1}^{n}\sup_{\zeta}\left(h(x,\zeta) - \lambda\|\zeta - \hat{\zeta}_i\|\right).$$

  - Finite convex program;
  - resulting problem is in the same complexity class as SAA

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

DRO with Wasserstein Distance for Logistic Regression

- Logistic regression suggests solving the ERM problem:

$$\text{minimize} \quad \frac{1}{n}\sum_{i=1}^{n} \ell(x, \xi_i, \lambda_i) := \mathbb{E}_{(\xi,\lambda)\sim\hat{\mathbb{P}}_n}[\ell(x, \xi, \lambda)]$$

$$\text{where} \quad \ell(x, \xi, \lambda) = \log(1 + e^{-\lambda x^T \xi})$$

- DRO suggests solving the problem

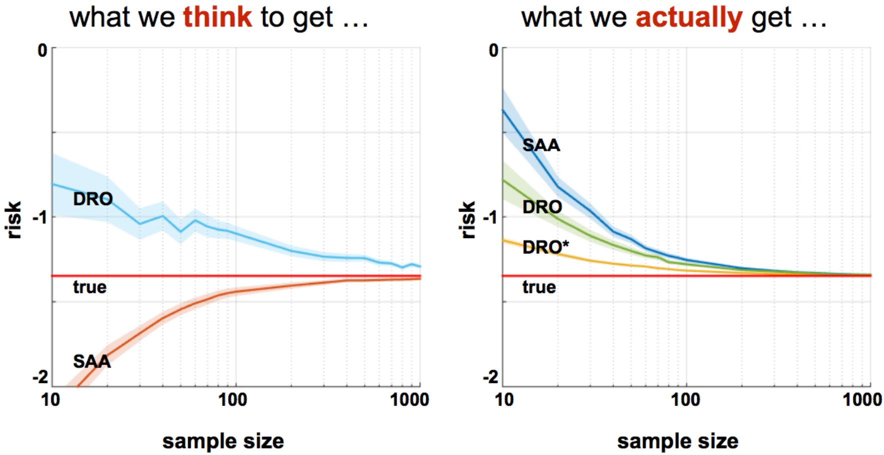
$$\text{minimize} \quad \left\{ \sup_{\mathbb{P}\in\mathcal{D}_n} \mathbb{E}_{(\xi,\lambda)\sim\mathbb{P}}[\ell(x, \xi, \lambda)] \right\}$$

- When labels are assumed to be error-free, DRO reduces to the regularized logistic regression:

$$\min_{x} \frac{1}{N}\sum_{i=1}^{N} \ell(x, \xi_i, \lambda_i) + C \cdot \|x\|_*.$$

Numerical Performance of DRO

Application: portfolio selection problem[6]:



[6] Blanchet (2018), Distributionally Robust Mean-Variance Portfolio

Summary of DRO with Wasserstein Distance

- The DRO model gives solution better than SAA.

Summary of DRO with Wasserstein Distance

- The DRO model gives solution better than SAA.
- The DRO model are tractable.

Summary of DRO with Wasserstein Distance

- The DRO model gives solution better than SAA.
- The DRO model are tractable.
- Well-understood in standard stochastic optimization problem.
  - Extension to general problems, e.g., un-supervised learning, sequential decision problems, etc.
  - Recently we are also applying this technique in multi-hop communication problems. (Ongoing project with Prof. Shenghao Yang)

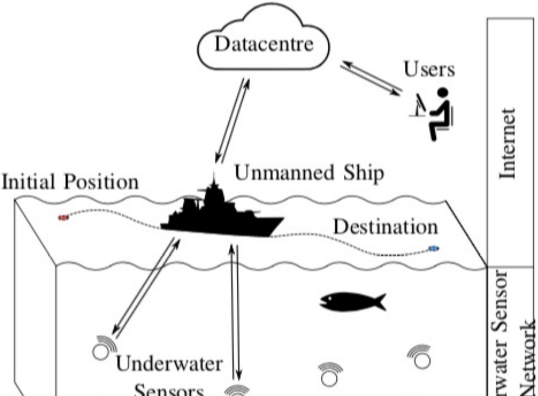
Related References

- Tractability of DRO model:
  - Distributionally Robust Stochastic Optimization with Wasserstein Distance, 2016.
  - Data-driven Robust Optimization with Known Marginal Distributions, 2017.
- Statistical Propeties of DRO model:
  - Wasserstein distributionally robust optimization: Theory and applications in machine learning, 2019.
- Applications of DRO model in supervised learning:
  - Distributionally robust logistic regression
  - Robust Wasserstein profile inference and applications to machine learning
- Introductory Videos about DRO: *https://www.youtube.com/watch?v=b4lJENGAeEA*

Application of Distributionally Robust Optimization in Off-policy Policy Evaluation

Introduction to OPPE

- Data: trajectories collected under a behavior policy π_b;
- Question: What would be the expected reward under target policy π?



香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

MDP Introduction

A MDP Environment: $\langle \mathcal{S}, \mathcal{A}, P, R, d_0 \rangle$ with $\gamma \in (0, 1)$;

- Expected reward:

$$R_\pi := \lim_{T \to \infty} \frac{\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t\right]}{\sum_{t=0}^{T} \gamma^t}$$

  where

$$s_0 \sim d_0, a_t \sim \pi(\cdot \mid s_t), r_t := r(s_t, a_t), s_{t+1} \sim P(\cdot \mid s_t, a_t).$$

MDP Introduction

A MDP Environment: $\langle \mathcal{S}, \mathcal{A}, P, R, d_0 \rangle$ with $\gamma \in (0,1)$;

- Expected reward:

$$R_\pi := \lim_{T\to\infty} \frac{\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t\right]}{\sum_{t=0}^{T} \gamma^t}$$

where

$$s_0 \sim d_0, a_t \sim \pi(\cdot \mid s_t), r_t := r(s_t, a_t), s_{t+1} \sim P(\cdot \mid s_t, a_t).$$

- Average visitation distribution:

$$d_\pi(s) = \lim_{T\to\infty} \frac{\sum_{t=0}^{T} \gamma^t d_{\pi,t}(s)}{\sum_{t=0}^{T} \gamma^t}.$$

It follows that

$$R_\pi = \mathbb{E}_{(s,a)\sim d_\pi}[r(s,a)] = \sum_{s,a} d_\pi(s)\pi(a \mid s) r(s,a).$$

Introduction to OPPE

- Historial data $\{(s_t^i, a_t^i, (s')_t^i)_{t=0}^T\}_{i=1}^N$ induced by the known behavior policy π_b is available:

$$\forall i, s_0 \sim d_0, a_t \sim \pi_b(\cdot \mid s_t), s_{t+1} \sim P(\cdot \mid s_t, a_t), \quad t = 1, \dots, T-1$$

- The goal is to evaluate reward for target policy π:

$$\begin{aligned} R_\pi &= \mathbb{E}_{(s,a)\sim d_\pi}[r(s,a)] = \sum_{s,a} d_\pi(s)\pi(a \mid s)r(s,a) \\ &= \mathbb{E}_{(s,a)\sim d_{\pi_b}}\left[w(s)\beta(s,a)r(s,a)\right], \end{aligned}$$

where $\omega(s) := \frac{d_\pi(s)}{d_{\pi_b}(s)}$ and $\beta(s,a) = \frac{\pi(a|s)}{\pi_b(a|s)}$.

Classical Approach to OPPE

In order to evaluate R_π:

$$R_\pi = \mathbb{E}_{(s,a)\sim d_{\pi_b}}\left[w(s)\beta(s,a)r(s,a)\right],$$

$$\text{with} \quad \omega(s) = \frac{d_\pi(s)}{d_{\pi_b}(s)}, \beta(s,a) = \frac{\pi(a \mid s)}{\pi_b(a \mid s)}$$

- Replace d_{π_b} with its empirical distribution, based on historical data;
- Estimate $\{\omega(s)\}_s$ by making use of the stationary equation:

$$w(s')d_{\pi_b}(s') = (1-\gamma)d_0(s') + \gamma \sum_{s,a} d_{\pi_b}(s,a,s')\beta(s,a)w(s), \quad \forall s'.$$

Classical Approach to OPPE

In order to evaluate R_π:

$$R_\pi = \mathbb{E}_{(s,a)\sim d_{\pi_b}}\left[w(s)\beta(s,a)r(s,a)\right],$$

$$\text{with} \quad \omega(s) = \frac{d_\pi(s)}{d_{\pi_b}(s)}, \beta(s,a) = \frac{\pi(a \mid s)}{\pi_b(a \mid s)}$$

- Replace d_{π_b} with its empirical distribution, based on historical data;
- Estimate $\{\omega(s)\}_s$ by making use of the stationary equation:

$$w(s')d_{\pi_b}(s') = (1-\gamma)d_0(s') + \gamma \sum_{s,a} d_{\pi_b}(s,a,s')\beta(s,a)w(s), \quad \forall s'.$$

  - Substitute $d_{\pi_b}(s,a,s')$ with $d_{\pi_b}(s)\pi_b(a \mid s)P(a,s' \mid s)$ gives

$$d_\pi(s') = (1-\gamma)d_0(s') + \sum_s d_\pi(s)P^\pi(s' \mid s), \quad \forall s'.$$

Challenge for Estimating the Ratio

The importance ratio $\{\omega(s)\}_s$ satisifes stationary equation:

$$\omega(s')d_{\pi_b}(s') = (1-\gamma)d_0(s')+\gamma\sum_{s,a} d_{\pi_b}(s,a,s')\beta(s,a)\omega(s), \quad \forall s' \in \mathcal{S}.$$

- **Challenge**: Only samples from $\{d_{\pi_b}(s,a,s')\}_{s,a,s'}$ are available;
- **Rescue**: Introduce test functions to reduce the variance. [7] The stationary equation holds if and only if for any f,

$$\mathbb{E}_{(s,a,s')\sim d_{\pi_b}}[\omega(s')f(s')-\gamma\beta(s,a)\omega(s)f(s)] = (1-\gamma)\mathbb{E}_{s\sim d_0}[f(s)].$$

[7] Qiang, Liu. Breaking the Curse of Horizon: Infinite-Horizon Off-Policy Estimation

Distributionally Robust Approach to OPPE

We propose the following distributionally robust and optimistic formulation:

$$
\begin{aligned}
\min_{w,\mu} / \max \quad & R_\pi := \sum_{s,a} \mu(s)\pi_b(a \mid s) w(s)\beta(s,a) r(s,a) \\
\text{subject to} \quad & w(s')\mu(s') = (1-\gamma) d_0(s') \\
& \qquad + \gamma \sum_{s,a} \mu(s,a,s')\beta(s,a) w(s), \quad \forall s' \in \mathcal{S} \\
& \mu \in \mathcal{P}.
\end{aligned}
$$

- Joint estimation framework for d_{π_b} and $\omega(s)$;
- Restrict μ, the estiamte for d_{π_b}, within the ambiguity set $\mathcal{P}$;
- Intractable bilinear optimization problem, but:

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Distributionally Robust Approach to OPPE

We propose the following distributionally robust and optimistic formulation:

$$
\begin{aligned}
\min_{w,\mu} / \max \quad & R_\pi := \sum_{s,a} \mu(s)\pi_b(a \mid s) w(s)\beta(s,a) r(s,a) \\
\text{subject to} \quad & w(s')\mu(s') = (1-\gamma) d_0(s') \\
& \qquad + \gamma \sum_{s,a} \mu(s,a,s')\beta(s,a) w(s), \quad \forall s' \in \mathcal{S} \\
& \mu \in \mathcal{P}.
\end{aligned}
$$

- Joint estimation framework for d_{π_b} and $\omega(s)$;
- Restrict μ, the estiamte for d_{π_b}, within the ambiguity set $\mathcal{P}$;
- Intractable bilinear optimization problem, but:
  - w can be uniquely determined for fixed μ.

Tractable Formulation to Robust OPPE

- By the change of variable $\kappa(s) = \mu(s)w(s)$, the max-max problem can be equivalently formulated as:

$$
\begin{aligned}
\max_{\kappa,\mu} \quad & \sum_s \kappa(s) \sum_a \pi(a \mid s) r(s,a) \\
\text{subject to} \quad & \kappa(s') = (1-\gamma) d_0(s') \\
& \qquad + \gamma \sum_s \kappa(s) \left[\sum_a \frac{\mu(s,a,s')}{\mu(s)} \beta(s,a) \right], \quad \forall s' \in \mathcal{S} \\
& \mu \in \mathcal{P}
\end{aligned}
$$

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Tractable Formulation to Robust OPPE

- By the change of variable $\kappa(s) = \mu(s)w(s)$, the max-max problem can be equivalently formulated as:

$$
\begin{aligned}
\max_{\kappa,\mu} \quad & \sum_s \kappa(s) \sum_a \pi(a \mid s) r(s,a) \\
\text{subject to} \quad & \kappa(s') = (1-\gamma) d_0(s') \\
& \qquad + \gamma \sum_s \kappa(s) \left[\sum_a \frac{\mu(s,a,s')}{\mu(s)} \beta(s,a) \right], \quad \forall s' \in \mathcal{S} \\
& \mu \in \mathcal{P}
\end{aligned}
$$

- Special design of ambiguity set $\mathcal{P}$ to ensure tractability:

$$
\begin{aligned}
\mathcal{P} &= \otimes_{s \in \mathcal{S}} \mathcal{P}_s \\
&= \otimes_{s \in \mathcal{S}} \left\{ \mu(\cdot, \cdot \mid s) : \; W\big(\mu(\cdot, \cdot \mid s), \hat{\mu}(\cdot, \cdot \mid s)\big) \le \vartheta_s \right\}.
\end{aligned}
$$

Tractable Formulation to Robust OPPE

Taking the duality for the inner maximization problem, we have

$$
\begin{aligned}
\text{Max}_{\mu}\text{Min}_{v} \quad & (1-\gamma)\sum_{s} v(s)d_0(s) \\
\text{subject to} \quad & v(s) \geq \sum_{a} \pi(a \mid s) r(s,a) \\
& \quad + \gamma \sum_{(a,s')} \mu(a,s' \mid s) v(s') \beta(s,a), \quad \forall s \\
& \mu \in \mathcal{P} = \otimes_{s \in \mathcal{S}} \Big\{ \mu(\cdot,\cdot \mid s) : \; W\big(\mu(\cdot,\cdot \mid s), \hat{\mu}(\cdot,\cdot \mid s)\big) \leq \vartheta_s \Big\}.
\end{aligned}
$$

Tractable Formulation to Robust OPPE

Applying the s-rectangularity of $\mathcal{P}$, we have

$$\begin{aligned}\text{Min}_v \quad & (1-\gamma)\sum_s v(s)d_0(s) \\ \text{subject to} \quad & v(s) \geq \sum_a \pi(a \mid s)r(s,a) \\ & \quad + \gamma \max_{\mu(\cdot,\cdot|s)\in\mathcal{P}_s} \sum_{(a,s')} \mu(a,s' \mid s)v(s')\beta(s,a), \quad \forall s \\ & \mathcal{P}_s = \left\{\mu(\cdot,\cdot \mid s) : \; W\big(\mu(\cdot,\cdot \mid s), \hat{\mu}(\cdot,\cdot \mid s)\big) \leq \vartheta_s\right\}.\end{aligned}$$

- Based on the fact that the uncertainty within constriants is uncoupled.

Tractable Formulation to Robust OPPE

Lemma: LP with Fixed Point Equation

Suppose that f is a component-wise non-decreasing contraction mapping with the unique fixed point x^*. Then for fixed $c \in \mathbb{R}^n_+$,

$$\max \left\{ c^T x : \quad x \in \mathbb{R}^n_+, x \leq f(x) \right\} = c^T x^*.$$

- Example: the policy evaluation problem in standard MDP reduces to the following LP problem:

$$\begin{array}{ll} \text{minimize} & (1-\gamma) \sum_s v(s) d_0(s) \\ \text{subject to} & v(s) \geq \mathcal{T}[v](s) \\ \text{with} & \mathcal{T}[v](s) = r_\pi(s) + \gamma \sum_{s'} v(s) \sum_a \pi(a \mid s) P(s' \mid s, a) \end{array}$$

Tractable Formulation to Robust OPPE

- By making use of this technique lemma, we argue at optimality the constraint is tight:

$$\begin{aligned} \min_{v} \quad & (1-\gamma)\sum_{s} v(s)d_0(s) \\ \text{s.t.} \quad & v(s) \geq \sum_{a} \pi(a \mid s) r(s,a) + \gamma V(s), \ \forall s \in \mathcal{S}, \end{aligned}$$

$$\textit{where} \quad V(s) := \max_{\mu(\cdot,\cdot|s)\in\mathcal{P}_s} \sum_{(a,s')} \mu(a,s' \mid s) v(s') \beta(s,a)$$

- The solution can be obtained by solving the fixed-point equation

$$v(s) = \sum_{a} \pi(a \mid s) r(s,a) + \gamma V(s), \ \forall s \in \mathcal{S}.$$

Algorithm for Optimistic Value Iteration

For each iteration:

- For each $s \in \mathcal{S}$, compute $V(s)$ by:

$$
\begin{aligned}
V(s) &= \max_{\mu(\cdot,\cdot|s)\in\mathcal{P}_s} \sum_{(a,s')} \mu(a, s' \mid s) v(s') \beta(s, a) \\
&= \min_{\lambda \geq 0} \Bigg\{ \lambda \vartheta_s + \\
&\qquad \frac{1}{n_s} \sum_{i=1}^{n_s} \max_{a\in\mathcal{A}, s'\in\mathcal{S}} \left\{ v(s')\beta(s, a) - \lambda c((a, s'), (a_i, s_i')) \right\} \Bigg\}
\end{aligned}
$$

- For each $s \in \mathcal{S}$, update

$$
v(s) \leftarrow \sum_a \pi(a \mid s) r(s, a) + \gamma \cdot V(s)
$$

Theoretical Gurantees for Robust OPPE

Lemma: Sensitivity Analysis for Value Iteration

- Denote by $\mathcal{T}$ the Bellman operator with the true conditional probability $d_{\pi_b}(a, s' \mid s)$:

$$\begin{aligned} & \mathcal{T}[v](s) = \textstyle\sum_a \pi(a \mid s) r(s, a) + \gamma \sum_{s'} P^{\text{true}}_{s,s'} v(s') \\ \text{with} \quad & P^{\text{true}}_{s,s'} := \textstyle\sum_a d_{\pi_b}(a, s' \mid s) \beta(s, a) \end{aligned}$$

- Denote by $\tilde{\mathcal{T}}$ a perturbation of $\mathcal{T}$ so that
  - $\tilde{\mathcal{T}}[v](s) = T[v](s) + \epsilon_v(s)$;
  - $\epsilon_v(s) \leq \epsilon(s)$ for all $s \in \mathcal{S}$ and v.

Let $v^*, \tilde{v}^*$ be the solutions to the fixed point of $\mathcal{T}$ and $\tilde{\mathcal{T}}$ respectively. Then

$$\tilde{v}^* - v^* \leq \left(I - \gamma P^{\text{true}}\right)^{-1} \epsilon.$$

Implications for the Lemma

- Our algorithm is simply the perturbation of the underlying Bellman operator:

$$\begin{aligned} v(s) &= \sum_a \pi(a \mid s) r(s,a) + \gamma V(s), \ \forall s \in \mathcal{S} \\ V(s) &= \max_{\mu(\cdot,\cdot|s) \in \mathcal{P}_s} \sum_{s'} \left[\mu(a, s' \mid s) \beta(s,a) \right] v(s') \\ &\approx \sum_{s'} P^{\text{true}}_{s,s'} v(s') \\ \mathcal{P}_s &= \left\{ \mu(\cdot, \cdot \mid s) : \ W\big(\mu(\cdot,\cdot \mid s), \hat{\mu}(\cdot,\cdot \mid s)\big) \le \vartheta_s \right\}. \end{aligned}$$

- Build the uniform bound for the perturbation gives the theoretical gurantees.

Proof for the Lemma

- Define $\tilde{v}^{(k)}$ as the k-th iteration point for the approximate value iteration algorithm, then we have the relation

$$
\begin{aligned}
\tilde{v}^{(k+1)} - v^* &= \tilde{T}[\tilde{v}^{(k)}] - \mathcal{T}[v^*] \\
&= \mathcal{T}[\tilde{v}^{(k)}] - \mathcal{T}[v^*] + \epsilon_{\tilde{v}^{(k)}} \\
&\leq \mathcal{T}[\tilde{v}^{(k)}] - \mathcal{T}[v^*] + \epsilon \\
&= \gamma P^{\text{true}}(\tilde{v}^{(k)} - v^*) + \epsilon
\end{aligned}
$$

- Applying the relation inductively, we have

$$
\tilde{v}^{(n)} - v^* \leq \sum_{k=0}^{n-1} \gamma^{n-k-1} (P^{\text{true}})^{n-k-1} \epsilon + \gamma^n (P^{\text{true}})^n (\tilde{v}^{(0)} - v^*)
$$

Taking the limit $n \to \infty$ completes the proof.

Uniform Bound for Perturbation

- The underlying true value function is returned by solving the fixed point equation

$$v(s) = \sum_a \pi(a \mid s) r(s,a) + \gamma \sum_{(a,s')} d_{\pi_b}(a, s' \mid s)[\beta(s,a) v(s')], \quad \forall s.$$

- The optimistic/robust value iteration is to solve

$$v(s) = \sum_a \pi(a \mid s) r(s,a) + \gamma \max_{\mu(\cdot,\cdot|s)\in\mathcal{P}_s} / \min \sum_{(a,s')} \mu(a, s' \mid s)[\beta(s,a) v(s')$$

Uniform Bound for Perturbation

- The underlying true value function is returned by solving the fixed point equation

$$v(s) = \sum_a \pi(a \mid s) r(s,a) + \gamma \sum_{(a,s')} d_{\pi_b}(a, s' \mid s)[\beta(s,a) v(s')], \quad \forall s.$$

- The optimistic/robust value iteration is to solve

$$v(s) = \sum_a \pi(a \mid s) r(s,a) + \gamma \max_{\mu(\cdot,\cdot|s) \in \mathcal{P}_s} / \min \sum_{(a,s')} \mu(a, s' \mid s)[\beta(s,a) v(s')$$

- Define $f(a,s') = \beta(s,a) v(s')$ for fixed s. Then with high probability,

$$\mathbb{E}_{\mathbb{P}_{\text{true}}}[f(a,s')] \leq \max_{\mathbb{P}:\ W(\mathbb{P}, \hat{\mathbb{P}}_n)} [f(a,s')] + \frac{6}{n}$$

$$\mathbb{E}_{\mathbb{P}_{\text{true}}}[f(a,s')] \geq \min_{\mathbb{P}:\ W(\mathbb{P}, \hat{\mathbb{P}}_n)} [f(a,s')] - \frac{6}{n}$$

Theoretical Gurantees for Robust OPPE

Theorem 2: Non-asymptotic Confidence Bounds

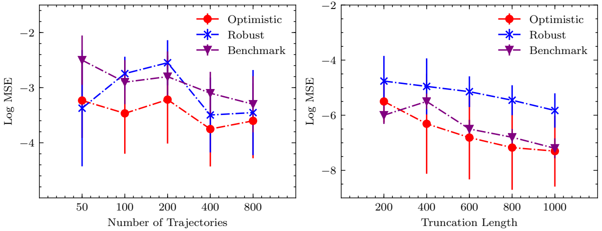
Denote $R_{\text{optimistic}}$ and R_{robust} as the reward for optimistic/robust estimate for the underlying reward R_π. With high probability,

$$R_\pi \leq R_{\text{optimistic}} + \frac{6}{n} \sum_{s\in\mathcal{S},s'\in\mathcal{S}} (I - \gamma P^{\text{true}})^{-1}_{s,s'} d_0(s),$$

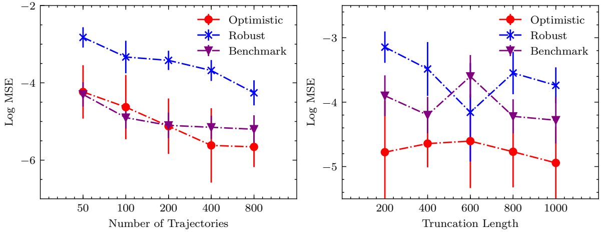
$$R_\pi \geq R_{\text{robust}} - \frac{6}{n} \sum_{s\in\mathcal{S},s'\in\mathcal{S}} (I - \gamma P^{\text{true}})^{-1}_{s,s'} d_0(s).$$

Moreover, $R_{\text{optimistic}} - R_{\text{robust}} = O(1/\sqrt{n})$.

Numerical Simulation



(a) Machine Replacement Problem



(b) Healthcare Management Problem

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Numerical Simulation

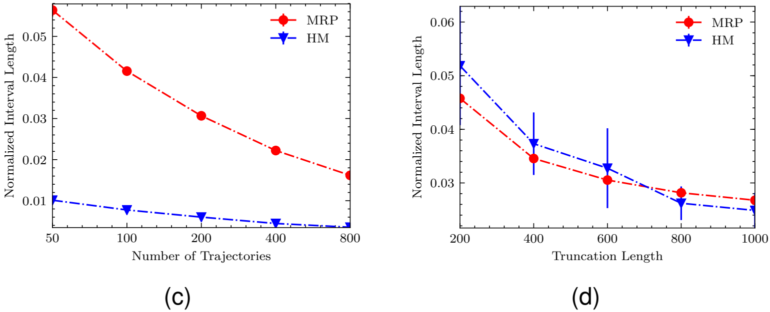


Figure: Plots for the normalized interval length with respect to number of trajectories and length of truncation.

Conclusion

- Our contributions involve:
  - Exact tractable reformulations for the distributionally robust and optimistic off-policy evaluation.
  - First non-asymptotic confidence interval estimate for infinite-horizon OPPE.
  - Generalization bound for Wasserstein distributionally robust optimization in discrete space.

Conclusion

- Our contributions involve:
  - Exact tractable reformulations for the distributionally robust and optimistic off-policy evaluation.
  - First non-asymptotic confidence interval estimate for infinite-horizon OPPE.
  - Generalization bound for Wasserstein distributionally robust optimization in discrete space.
- Future work would be:
  - Extend its applicability into general problems;
  - Design more efficient algorithm to solve the problem faster.

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen