Distributionally Robust Optimization: Theory and Applications

Speaker: Jie Wang

August 1, 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 Adaptive Recoding
  - Tractable formulation
- A Recent Application in Off-policy Policy Evaluation
  - Tractable formulation, theory, extensions
- Summary

The talk involves contributions from (in random order):
Rui Gao, Hongyuan Zha, Xinyun Chen, Shenghao Yang, Zhiyuan Jia, Hoover H. F. Yin

Backgroud about Distributionally Robust Optimization: Tractable Formulation and Statistics

Introduction to Stochastic Optimization

Consider the *stochastic optimization problem* as follows:

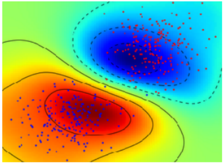
$$\text{maximize}_{x \in \mathcal{X}} \qquad \mathbb{E}_{\zeta \sim \mathbb{P}}[h(x, \zeta)] \tag{1}$$

with $\mathcal{X}$ being convex.

Applications:

Supply Chain Mgmt.

Portfolio Mgmt.



Machine Learning

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

Introduction to Stochastic Optimization

Consider the *stochastic optimization problem* as follows:

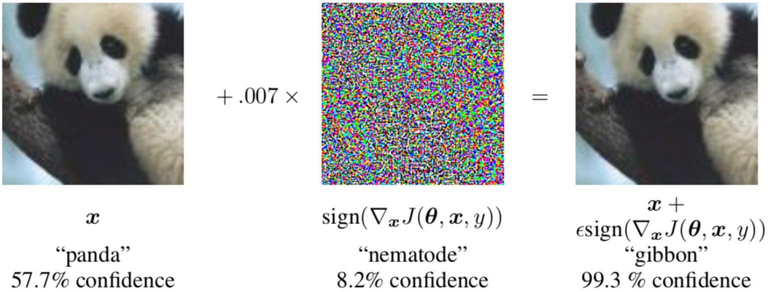
$$\text{maximize}_{x \in \mathcal{X}} \qquad \mathbb{E}_{\zeta \sim \mathbb{P}}[h(x, \zeta)] \tag{2}$$

with $\mathcal{X}$ being convex.

- **Prospective**
  - expected value is a good measure of performance;
  - simply solve by *sample average approximation* (SAA).
- **Challenge**
  - difficult to know the exact distribution of ζ;
  - SAA may result in sub-optimal solutions;
  - solution can be risky even know the distribution.

Stochastic Optimization with Noises

Adversarial attacks for classification problem [1]:



[1] Ian Goodfellow 2015

Picture for Gibbon

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]:

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

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

Motivation for DRO: Distributional Uncertainty

- Poor performance of SAA: with high probability,

$$\left| \mathbb{E}_{\zeta \sim \mathbb{P}_{\text{true}}}[h(x,\zeta)] - \frac{1}{n}\sum_{i=1}^{n} h(x,\zeta_i) \right| \leq O\left(\sqrt{\frac{\text{Var}[h(x,\zeta)]}{n}} \right).$$

- Distributional Uncertainty: Exact distribution of the random variables is difficult to obtain, but observed samples and other statistical information is available.

How to develop an algorithm that cooperates the distributional uncertainty?

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 becnmark 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:

$$D_f(P\|Q) = \int_\Omega f(dP/dQ)dQ;$$

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

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

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

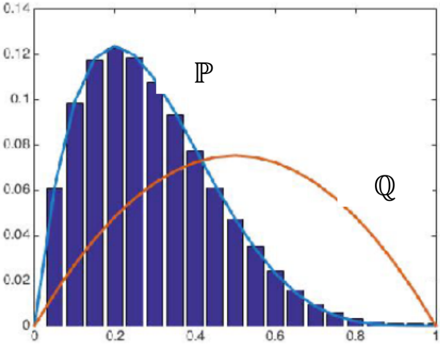
$$S(p,q) = \sup_{f\in\mathcal{F}} \left| \int \mathcal{A}_p[f(x)]dq(x) \right|$$

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

Introduction to Wasserstein Distance

By the duality theory in LP,

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



$W(\mathbb{P},\mathbb{Q})$ quantifies the minimum cost of moving $\mathbb{P}$ to $\mathbb{Q}$.

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

Statistics Properties for DRO with Wasserstein Distance

Theorem 1

Consider the DRO problem

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

with $\mathcal{D}_n = \{\mathbb{P} : W(\mathbb{P}, \hat{\mathbb{P}}_n) \leq \delta_n\}$, $\delta_n = O(1/\sqrt{n})$, 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 : DRO is in the same complexity class as SAA.

Tractability of DRO with Wasserstein Distance

Worse-case expectation problem:

$$\sup_{\mathbb{P}\in\mathcal{D}_n} \mathbb{E}_{\zeta\sim\mathbb{P}}[\ell(\zeta)],$$

where

$$\mathcal{D}_n = \{\mathbb{P} : \; W(\mathbb{P}, \hat{\mathbb{P}}_n) \leq \delta_n\}$$

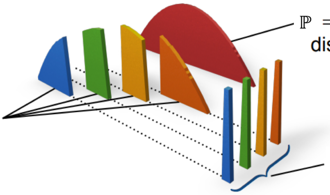
with

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

Tractability of DRO with Wasserstein Distance

Reformulation:

$$\begin{aligned}
\sup_{\Gamma, \mathbb{P}} \quad & \int \ell(\zeta) d\mathbb{P}(\zeta) \\
\text{subject to} \quad & \iint \|\zeta - \zeta'\| \Gamma(d\zeta, d\zeta') \leq \delta_n \\
& \Gamma \text{ is a joint distribution of } \zeta, \zeta', \\
& \text{with marginals } \mathbb{P} \text{ and } \hat{\mathbb{P}}_n, \text{ respectively}
\end{aligned}$$



Tractability of DRO with Wasserstein Distance

Decompose Γ into $\frac{1}{n}\sum_{i=1}^{n}\delta_{\zeta_i}\otimes\mathbb{P}_i$:

$$\begin{array}{ll} \sup\limits_{\mathbb{P}_i, i=1,2,\ldots,n} & \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\int \ell(\zeta)d\mathbb{P}_i(\zeta) \\ \text{subject to} & \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\int \|\zeta-\hat{\zeta}_i\|d\mathbb{P}_i(\zeta)\leq\delta_n \end{array}$$

Tractability of DRO with Wasserstein Distance

Apply the duality theory in linear programming:

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

When combining with the outer minimization over $x \in \mathcal{X}$,

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

- Finite convex program;
- Problem size grows polynomially in input data;
- 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

- Consider the feature-label training dataset $\{(\xi_i, \lambda_i)\}_{i=1}^n$ generated from $\mathbb{P}$, and consider the logistic regression:

$$\min_x \frac{1}{n}\sum_{i=1}^n \ell(x, \xi_i, \lambda_i), \quad \text{where } \ell(x, \xi, \lambda) = \log(1 + e^{-\lambda x^T \xi}).$$

- DRO suggests solving the problem

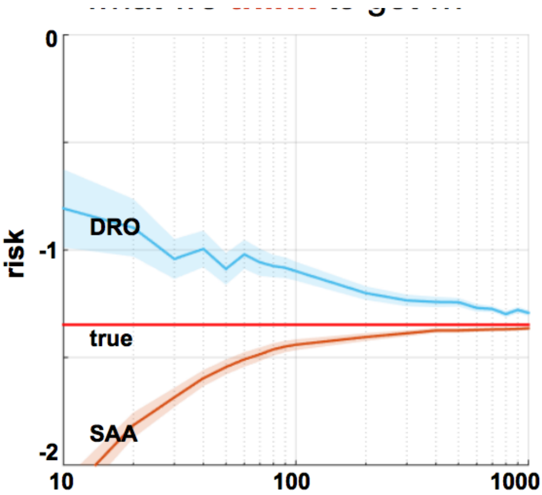
$$\min_x \sup_{\mathbb{P}\in\mathcal{D}_n} \mathbb{E}_{\mathbb{P}}[\ell(x, \xi, \lambda)].$$

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

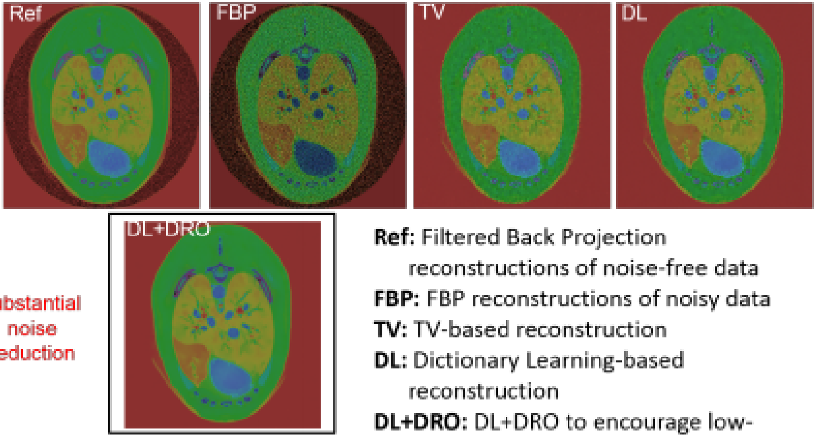
$$\min_x \frac{1}{N}\sum_{i=1}^N \ell(x, \xi_i, \lambda_i) + C \cdot \|x\|_*.$$

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

Performance of DRO in Supervised Learning



Performance of DRO in Medical Application



Substantial noise reduction

Ref: Filtered Back Projection reconstructions of noise-free data
FBP: FBP reconstructions of noisy data
TV: TV-based reconstruction
DL: Dictionary Learning-based reconstruction
DL+DRO: DL+DRO to encourage low-rankness and robustness

Summary of DRO with Wasserstein Distance

- The DRO model gives a solution with statistical guarantees better than SAA approach. The ambiguity set for possible distributions can be constructed from historical training data.

Summary of DRO with Wasserstein Distance

- The DRO model gives a solution with statistical guarantees better than SAA approach. The ambiguity set for possible distributions can be constructed from historical training data.
- The DRO model are tractable, and sometimes with the same complexity class as SAA.

Summary of DRO with Wasserstein Distance

- The DRO model gives a solution with statistical guarantees better than SAA approach. The ambiguity set for possible distributions can be constructed from historical training data.
- The DRO model are tractable, and sometimes with the same complexity class as SAA.
- This approach in standard stochastic optimization is well-understood for utilizing the data uncertainty. We wish to extend its applicability into more general optimization problems, such as semi-supervised learning, reinforcement learning, etc.

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=b4IJENGAeEA*

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

Application of Distributionally Robust Optimization in Adaptive Network Coding

Introduction to Adaptive Network Coding

- The adpative recoding scheme suffices to solve the following optimization problem:

$$\max_{t_r \geq 0, \forall r \in [M]} \quad \sum_{r=0}^{M} h_r E_r(t_r)$$

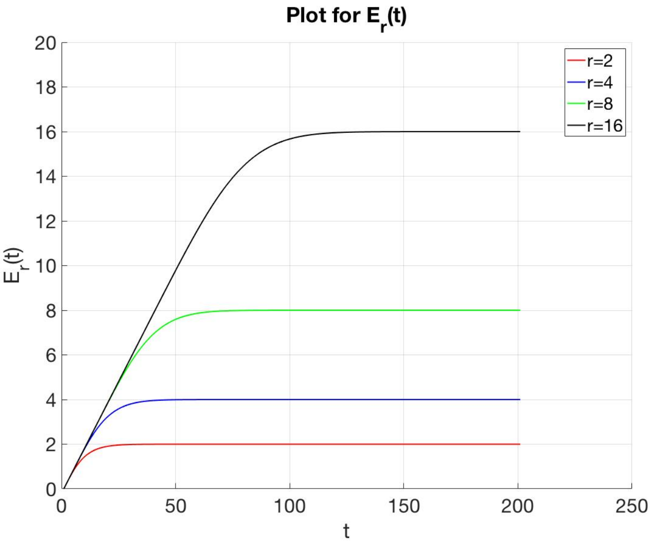
$$\text{subject to} \quad \sum_{r=0}^{M} h_r t_r \leq t_{\text{avg}}$$

  where $\{h_r\}_{r \in [M]}, t_{\text{avg}}$ are parameters of this problem.
- The function $E_r(t)$ is a concave, monotone increasing function:

$$E_r(t) = \sum_{i=0}^{t} \Pr\left(\sum_{j=1}^{t} Z_j = i\right) \sum_{j=0}^{\min\{i,r\}} j\zeta_j^{i,r}, \quad t \in \mathbb{N},$$

  and for $t \in \mathbb{R}$, $E_r(t) = \epsilon E_r(\lfloor t \rfloor + 1) + (1 - \epsilon) E_r(\lfloor t \rfloor)$.

Properties of the Optimization Problem



Properties of the Optimization Problem

$$\max_{t_r \geq 0, \forall r \in [M]} \quad \sum_{r=0}^{M} h_r E_r(t_r)$$

$$\text{subject to} \quad \sum_{r=0}^{M} h_r t_r \leq t_{\text{avg}}$$

- At optimality the inequality constraint is tight;
- At optimality all except one $t_r, r \in [M]$ are integers.
- To solve the problem from primal, a greedy algorithm similar to bin-packing can be developed:

$$\begin{array}{ll} \text{minimize} & \sum_{j \in J} x_j \\ \text{subject to} & \sum_{j \in J} a_j x_j \geq n \end{array}$$

Solving the Optimization from Duality

$$
\begin{aligned}
\max_{t_r \geq 0, \forall r \in [M]} \quad & \sum_{r=0}^{M} h_r E_r(t_r) \\
\text{subject to} \quad & \sum_{r=0}^{M} h_r t_r \leq t_{\text{avg}}
\end{aligned}
$$

The dual is a one-dimensional convex programming problem:

$$
\min_{\lambda \geq 0} \ \lambda t_{\text{avg}} + \sum_{r \in [M]} h_r \sup_{t_r \geq 0} \left(E_r(t_r) - \lambda t_r \right).
$$

- Dual sub-gradient? Cannot gurantee convergence.

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

Solving the Optimization from Duality

$$\begin{aligned} \max_{t_r \geq 0, \forall r \in [M]} \quad & \sum_{r=0}^{M} h_r E_r(t_r) \\ \text{subject to} \quad & \sum_{r=0}^{M} h_r t_r \leq t_{\text{avg}} \end{aligned}$$

The dual is a one-dimensional convex programming problem:

$$\min_{\lambda \geq 0} \ \lambda t_{\text{avg}} + \sum_{r \in [M]} h_r \sup_{t_r \geq 0} \Big(E_r(t_r) - \lambda t_r \Big).$$

- Dual sub-gradient? Cannot gurantee convergence.
- Bisection algorithm, with each step solving for the inner supremum problem at optimality.

Main Challenge for SAA

$$\max_{t_r \geq 0, \forall r \in [M]} \quad \sum_{r=0}^{M} h_r E_r(t_r)$$
$$\text{subject to} \quad \sum_{r=0}^{M} h_r t_r \leq t_{avg}$$

- $\{h_r\}$ denotes the rank distribution. But we cannot obtain the exact distribution, but only some samples $\{r_i\}_{i=1}^{N}$.

$$\max_{t_r \geq 0, \forall r \in [M]} \quad \sum_{r=0}^{M} \hat{h}_r E_r(t_r)$$
$$\text{subject to} \quad \sum_{r=0}^{M} \hat{h}_r t_r \leq t_{avg}$$

Main Challenge for SAA

$$\max_{t_r \geq 0, \forall r \in [M]} \quad \sum_{r=0}^{M} h_r E_r(t_r)$$
$$\text{subject to} \quad \sum_{r=0}^{M} h_r t_r \leq t_{\text{avg}}$$

- $\{h_r\}$ denotes the rank distribution. But we cannot obtain the exact distribution, but only some samples $\{r_i\}_{i=1}^{N}$.

$$\max_{t_r \geq 0, \forall r \in [M]} \quad \sum_{r=0}^{M} \hat{h}_r E_r(t_r)$$
$$\text{subject to} \quad \sum_{r=0}^{M} \hat{h}_r t_r \leq t_{\text{avg}}$$

- Optimizer's curse: Solution to SAA may even not be feasible in the original problem.

DRO formulation for Adaptive Network Coding

$$\max_{\{t_r\}_{r\in[M]}\geq 0} \quad \min_{\mathrm{h}:W(\mathrm{h},\hat{\mathrm{h}})\leq\varepsilon_1} \quad \mathbb{E}_{r\sim\mathrm{h}}[E_r(t_r)]$$

$$\text{Subject to} \quad \sup_{\mathrm{h}:W(\mathrm{h},\hat{\mathrm{h}})\leq\varepsilon_2} \quad \mathbb{E}_{r\sim\mathrm{h}}[t_r] \leq t_{\mathrm{avg}}.$$

where $W(\cdot,\cdot)$ is a Wasserstein metric:

$$W(\mu,\nu) = \min_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} c(x,y)\mathrm{d}\gamma(x,y).$$

- The objective is to pick the decision variable to optimize for the "worse-case" distribution;
- The constraint is to ensure the obtained solution is feasible in the underlying problem.

Tractable formulation for DRO problem

$$\max_{\{t_r\}_{r\in[M]}\in\mathcal{T}}\left\{\min_{\mathrm{h}:W(\mathrm{h},\hat{\mathrm{h}})\leq\varepsilon_1}\mathbb{E}_{r\sim\mathrm{h}}[E_r(t_r)]\right\}$$

$$\text{where}\quad \mathcal{T}=\left\{\{t_r\}_{r\in[M]}:\ t_r\geq 0,\ \sup_{\mathrm{h}:W(\mathrm{h},\hat{\mathrm{h}})\leq\varepsilon_2}\mathbb{E}_{r\sim\mathrm{h}}[t_r]\leq t_{\mathrm{avg}}\right\}.$$

Reformulation about the Objective Function

$$\min_{\mathrm{h}:W(\mathrm{h},\hat{\mathrm{h}})\leq\varepsilon_1} \mathbb{E}_{r\sim\mathrm{h}}[E_r(t_r)] = \sup_{\lambda_1\geq 0}\left\{-\lambda_1\varepsilon_1+\frac{1}{N}\sum_{i\in[N]}\inf_{r\in[M]}\left(E_r(t_r)+\lambda_1 c(r,\hat{r}_i)\right)\right\}$$

Reformulation about the Constraint

By the standard duality result,

$$\sup_{\mathrm{h}:W(\mathrm{h},\hat{\mathrm{h}})\le\varepsilon_2} \mathbb{E}_{r\sim\mathrm{h}}[t_r] = \inf_{\lambda_2\ge 0} \lambda_2\varepsilon_2 + \frac{1}{N}\sum_{i\in[N]} \sup_{r\in[M]} \Big(t_r - \lambda_2 c(r,\hat{r}_i) \Big).$$

It follows that Hence, $\mathcal{T}$ can be reformulated as

$$\mathcal{T} = \left\{ \{t_r\}_{r\in[M]} : \ t_r \ge 0, \sup_{\mathrm{h}:W(\mathrm{h},\hat{\mathrm{h}})\le\varepsilon_2} \mathbb{E}_{r\sim\mathrm{h}}[t_r] \le t_{\text{avg}} \right\}$$

$$= \Bigg\{ \{t_r\}_{r\in[M]} : \ t_r \ge 0,$$

$$\lambda_2\varepsilon_2 + \frac{1}{N}\sum_{i\in[N]} \sup_{r\in[M]} \Big(t_r - \lambda_2 c(r,\hat{r}_i) \Big) \le t_{\text{avg}} \text{for some } \lambda_2 \ge 0 \Bigg\}$$

Tractable formulation for DRO problem

The original DRO problem admits the following tractable formulation:

$$
\begin{aligned}
\max_{\substack{t_r \geq 0,\ r \in [M] \\ \lambda_1 \geq 0, \lambda_2 \geq 0}} \quad & -\lambda_1 \varepsilon_1 + \frac{1}{N} \sum_{i \in [N]} \inf_{r \in [M]} \Big(E_r(t_r) + \lambda_1 c(r, \hat{r}_i) \Big) \\
\text{s.t.} \quad & \lambda_2 \varepsilon_2 + \frac{1}{N} \sum_{i \in [N]} \sup_{r \in [M]} \Big(t_r - \lambda_2 c(r, \hat{r}_i) \Big) \leq t_{\text{avg}}
\end{aligned}
$$

- At optimality, is the solution still be almost deterministic?
- How to develop algorithm to solve this problem efficiently?

Application of Distributionally Robust Optimization in Off-policy Policy Evaluation

Introduction to OPPE

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

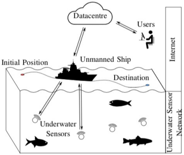
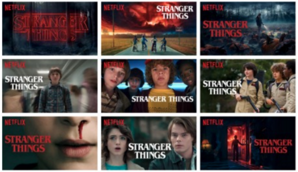


Fig. 1. Internet of Underwater Things

(a) Unmanned Data Collection



(b) Artwork Optimization at Netflix

MDP Introducttion

A MDP Environment:

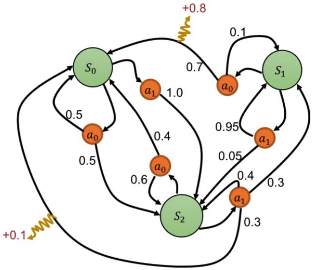
$\langle \mathcal{S}, \mathcal{A}, P, R, \gamma, d_0 \rangle$;

- Expected reward:

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

- 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

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

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

Introduction to OPPE

The expected reward R_π can be expressed in the expectation form

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

where $d_\pi(s,a) := d_\pi(s)\pi(a \mid s)$, and ω denotes the marginalized importance ratio:

$$w(s) := \frac{d_\pi(s)}{d_{\pi_b}(s)}, \quad \beta(s,a) := \frac{\pi(a \mid s)}{\pi_b(a \mid s)}.$$

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

Classical Approach to OPPE

In order to evaluate the reward for target policy π,

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

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

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

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

Challenge for Estimating the Ratio

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

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

- Challenge: Estimating $\{d_{\pi_b}(s,a,s')\}_{s,a,s'}$ based on historial data is not accurate;
- Rescue: Introduce the test function to reduce the variance.[6] The stationary equation holds if and only if for any f,

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

[6] 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:

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

where μ is the estimate for the underlying distribution d_{π_b}, and

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

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:

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

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

Tractable Formulation to Robust OPPE

Taking the duality for the inner maximization problem, we have

$$
\begin{aligned}
\text{Max}_{\mu}\text{Min}_{v} \quad & (1-\gamma)\sum_{s} v(s)d_0(s) \\
\text{subject to} \quad & v(s) \geq \sum_{a} \pi(a \mid s) r(s,a) \\
& \qquad + \gamma \sum_{(a,s')} \mu(a,s' \mid s) v(s') \beta(s,a), \quad \forall s \\
& \mu \in \mathcal{P}
\end{aligned}
$$

Tractable Formulation to Robust OPPE

Theorem 2: Refomulation of LP with contraction mapping constraints

Suppose that f is a component-wise non-decreasing contraction mapping with the unique fixed point x^*. The optimization problem

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

Tractable Formulation to Robust OPPE

By making use of this technique result, the max-max problem can be reformulated as

$$
\begin{aligned}
\min_{v} \quad & (1-\gamma)\sum_{s} v(s) d_0(s) \\
\text{s.t.} \quad & v(s) \geq \sum_{a} \pi(a \mid s) r(s,a) + \gamma V(s), \ \forall s \in \mathcal{S}, \\
\textit{where} \quad & V(s) := \max_{\mu(\cdot,\cdot|s) \in \mathcal{P}_s} \sum_{(a,s')} \mu(a, s' \mid s) v(s') \beta(s,a)
\end{aligned}
$$

At optimality the constraint is tight. The solution can be obtained by solving the fixed-point equation

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

Theoretical Gurantees for Robust OPPE

Lemma

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

$$\mathcal{T}[v](s) = \sum_a \pi(a \mid s) r(s, a) + \gamma \sum_{(a,s')} d_{\pi_b}(a, s' \mid s) v(s') \beta(s, a).$$

Denote by $\tilde{\mathcal{T}}$ a perturbation of $\mathcal{T}$ so that $\tilde{\mathcal{T}}[v](s) = T[v](s) + \epsilon_v(s)$. Assume there exist $\epsilon = (\epsilon(s))_{s \in \mathcal{S}}$ such that $\epsilon_v(s) \le \epsilon(s)$ 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

$$\tilde{v}^* - v^* \le \left(I - \gamma P^{true}\right)^{-1} \epsilon,$$

where $P^{true} \in \mathcal{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is defined as $P^{true}_{s,s'} := \sum_a d_{\pi_b s}(a, s' \mid s) \beta(s, a)$, and the inequality is interpreted component-wise.

Theoretical Gurantees for Robust OPPE

Theorem 3: Non-asymptotic Confidence Bounds

With high probability,

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

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

Conclusion

- Powerful tool for stochastic optimization problems;
- Computationally tractable, elegant theoretical gurantees;
- Future work:
  - Extend its applicability into sequential decision problems, such as reinforcement learning, and Off-line Policy-improvement problem;
  - Incorporate model uncertainty to solve more problems in Network Coding;
  - Design more efficient algorithm to solve the problem faster.

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