Tikhonov Regularization is Optimal Transport Robust under Martingale Constraints

Jiajin Li

Department of Management Science and Engineering

Stanford University

[**Joint work** with Sirui Lin, Jose Blanchet, Viet Anh Nguyen]

Oct 17th, 2022

Outline

Empirical Risk Minimization

- Training dataset: $\{X_i\}_{i=1}^N$ i.i.d. drawn from $\mathbb{P}^\star$;

Empirical Risk Minimization

- Training dataset: $\{X_i\}_{i=1}^N$ i.i.d. drawn from $\mathbb{P}^\star$;
- As the true distribution $\mathbb{P}^\star$ is typically not known, one considers the empirical risk minimization (ERM) problem

$$\inf_\beta \left\{ \mathbb{E}_{X\sim\hat{\mathbb{P}}}[\ell(f_\beta(X))] = \frac{1}{N}\sum_{i=1}^N \ell(f_\beta(X_i)) \right\},$$

where

$$\hat{\mathbb{P}} := \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$$

is the empirical distribution associated with the training dataset.

Overfitting and Regularization

- A well-known issue with ERM is overfitting.

Overfitting and Regularization

- A well-known issue with ERM is overfitting.
- A standard approach to deal with this is regularization:

$$\min_{\beta} \left\{ \mathbb{E}_{X \sim \hat{\mathbb{P}}}[\ell(f_\beta(X))] + R(f_\beta) \right\}.$$

Overfitting and Regularization

- A well-known issue with ERM is overfitting.
- A standard approach to deal with this is regularization:

$$\min_{\beta} \left\{ \mathbb{E}_{X \sim \hat{\mathbb{P}}}[\ell(f_\beta(X))] + R(f_\beta) \right\}.$$

- Distributionally robust optimization (DRO) — a fresh and principled perspective on regularization [Shafieezadeh-Abadeh et al.(2019), Gao et al. (2022)].

Optimal Transport-based DRO Formulation

- We consider minimizing the worst-case expected loss

$$\inf_{\beta} \sup_{\mathbb{Q} \in B_\rho(\hat{\mathbb{P}})} \mathbb{E}_{X \sim \mathbb{Q}}[\ell(f_\beta(X))], \qquad \text{(OT-DRO)}$$

where $B_\rho(\hat{\mathbb{P}})$, the so-called ambiguity set, is defined as

$$B_\rho(\hat{\mathbb{P}}) = \{\mathbb{Q} : D(\mathbb{Q}, \hat{\mathbb{P}}) \leq \rho\}.$$

Here $D(\mathbb{Q}, \hat{\mathbb{P}})$ is the optimal transport distance between $\mathbb{Q}$ and $\hat{\mathbb{P}}$ with the quadratic cost.

Optimal Transport-based DRO Formulation

- We consider minimizing the worst-case expected loss

$$\inf_{\beta} \sup_{\mathbb{Q} \in B_\rho(\hat{\mathbb{P}})} \mathbb{E}_{X \sim \mathbb{Q}}[\ell(f_\beta(X))], \qquad \text{(OT-DRO)}$$

  where $B_\rho(\hat{\mathbb{P}})$, the so-called ambiguity set, is defined as

$$B_\rho(\hat{\mathbb{P}}) = \{\mathbb{Q} : D(\mathbb{Q}, \hat{\mathbb{P}}) \leq \rho\}.$$

  Here $D(\mathbb{Q}, \hat{\mathbb{P}})$ is the optimal transport distance between $\mathbb{Q}$ and $\hat{\mathbb{P}}$ with the quadratic cost.

- The average size perturbation among all empirical data is less than a given budget.

Main Focus

Individual Perturbation

Impose additional **martingale constraints!**

Main Focus

Individual Perturbation

Impose additional **martingale constraints!**

$$B_\rho^M(\hat{\mathbb{P}}) = \{\mathbb{Q} : D(\mathbb{Q}, \hat{\mathbb{P}}) \le \rho\} \cap \boxed{\{\mathsf{Q}{:}\mathsf{E}_{\mathbb{Q}|\hat{\mathbb{P}}}[\bar{X}|X] = X\}}$$

Motivation Question

Why the martingale constraint makes sense as a regularization techinique?

Motivation Question

Why the martingale constraint makes sense as a regularization techinique?

Martingale Constraints

- Combat the overconveratіveness issue of the OT-DRO;

Martingale Constraints

- Combat the overconveratíveness issue of the OT-DRO;
- The conditional expectation of the additive perturbation for each data point (individually) equals to zero.

Martingale Constraints

- Combat the overconveratíveness issue of the OT-DRO;
- The conditional expectation of the additive perturbation for each data point (individually) equals to zero.
- $\mathbb{E}[\bar{X}|X] = X \iff$ The distribution of $\bar{X}$ dominate X in convex order [Strassen et al.(1965)].

Martingale Constraints

- Combat the overconverativeness issue of the OT-DRO;
- The conditional expectation of the additive perturbation for each data point (individually) equals to zero.
- $\mathbb{E}[\bar{X}|X] = X \iff$ The distribution of $\bar{X}$ dominate X in convex order [Strassen et al.(1965)].
- The adversary $\bar{X}$ will have high dispersion than empirical data in non-parametric sense.

Martingale Constraints

- Combat the overconverativeness issue of the OT-DRO;
- The conditional expectation of the additive perturbation for each data point (individually) equals to zero.
- $\mathbb{E}[\bar{X}|X] = X \iff$ The distribution of $\bar{X}$ dominate X in convex order [Strassen et al.(1965)].
- The adversary $\bar{X}$ will have high dispersion than empirical data in non-parametric sense.
- Well-motivated in robust mathematical finance, e.g., martingale optimal transport $\cdots$

Outline

A Motivation Example: Linear Regression

$$\inf_{\beta} \sup_{\mathbb{Q} \in B_\rho^M(\hat{\mathbb{P}})} \mathbb{E}_{X \sim \mathbb{Q}}[\ell(f_\beta(X))], \qquad \text{(Exact Martingale DRO)}$$

- Exact martingale based ambiguity set:

$$B_\rho^M(\hat{\mathbb{P}}) = \{\mathbb{Q} : D(\mathbb{Q}, \hat{\mathbb{P}}) \leq \rho\} \cap \{\mathbb{Q} : \mathbb{E}_{\mathbb{Q}|\hat{\mathbb{P}}}[\bar{X} | X] = X\}$$

A Motivation Example: Linear Regression

$$\inf_{\beta} \sup_{\mathbb{Q} \in B_\rho^M(\hat{\mathbb{P}})} \mathbb{E}_{X \sim \mathbb{Q}}[\ell(f_\beta(X))], \qquad \text{(Exact Martingale DRO)}$$

- Exact martingale based ambiguity set:

$$B_\rho^M(\hat{\mathbb{P}}) = \{\mathbb{Q} : D(\mathbb{Q}, \hat{\mathbb{P}}) \le \rho\} \cap \{\mathbb{Q} : \mathbb{E}_{\mathbb{Q}|\hat{\mathbb{P}}}[\bar{X}|X] = X\}$$

- Family of linear functions $X \to f_\beta(X) := \beta^T X$;

A Motivation Example: Linear Regression

$$\inf_{\beta} \sup_{\mathbb{Q} \in B_\rho^M(\hat{\mathbb{P}})} \mathbb{E}_{X \sim \mathbb{Q}}[\ell(f_\beta(X))], \qquad \text{(Exact Martingale DRO)}$$

- Exact martingale based ambiguity set:

$$B_\rho^M(\hat{\mathbb{P}}) = \{\mathbb{Q} : D(\mathbb{Q}, \hat{\mathbb{P}}) \le \rho\} \cap \{\mathbb{Q} : \mathbb{E}_{\mathbb{Q}|\hat{\mathbb{P}}}[\bar{X}|X] = X\}$$

- Family of linear functions $X \to f_\beta(X) := \beta^T X$;
- $\ell(\cdot) = \|\cdot\|^2$ is a quadratic loss;

A Motivation Example: Linear Regression

$$\inf_{\beta} \sup_{\mathbb{Q} \in B_\rho^M(\hat{\mathbb{P}})} \mathbb{E}_{X \sim \mathbb{Q}}[\ell(f_\beta(X))], \qquad \text{(Exact Martingale DRO)}$$

- Exact martingale based ambiguity set:

$$B_\rho^M(\hat{\mathbb{P}}) = \{\mathbb{Q} : D(\mathbb{Q}, \hat{\mathbb{P}}) \leq \rho\} \cap \{\mathbb{Q} : \mathbb{E}_{\mathbb{Q}|\hat{\mathbb{P}}}[\bar{X}|X] = X\}$$

- Family of linear functions $X \to f_\beta(X) := \beta^T X$;
- $\ell(\cdot) = \|\cdot\|^2$ is a quadratic loss;

A Motivation Example: Linear Regression

$$\inf_{\beta} \sup_{\mathbb{Q} \in B_\rho^M(\hat{\mathbb{P}})} \mathbb{E}_{X \sim \mathbb{Q}}[\ell(f_\beta(X))], \qquad \text{(Exact Martingale DRO)}$$

- Exact martingale based ambiguity set:

$$B_\rho^M(\hat{\mathbb{P}}) = \{\mathbb{Q} : D(\mathbb{Q}, \hat{\mathbb{P}}) \le \rho\} \cap \{\mathbb{Q} : \mathbb{E}_{\mathbb{Q}|\hat{\mathbb{P}}}[\bar{X}|X] = X\}$$

- Family of linear functions $X \to f_\beta(X) := \beta^T X$;
- $\ell(\cdot) = \|\cdot\|^2$ is a quadratic loss;

Theorem

The exact Martingale DRO model is exactly equivalent to ridge regression with Tikhonov regularization, i.e.,

$$\min_{\beta} \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\beta^T X)] + \rho\|\beta\|_2^2.$$

A Motivating Example: Linear Regression

- Exact Martingle DRO is equivalent to ridge regression,

$$\min_{\beta} \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\beta^T X)] + \rho\|\beta\|_2^2 \qquad \text{(Exact Martingale DRO)}$$

- The conventional OT-DRO is equivalent to the square-root regression problem [Blanchet et al. (2019)], i.e.,

$$\min_{\beta} \left(\sqrt{\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\beta^T X)]} + \sqrt{\rho}\|\beta\|_2\right)^2 \qquad \text{(OT-DRO)}$$

A Motivating Example: Linear Regression

- Exact Martingle DRO is equivalent to ridge regression,

$$\min_{\beta} \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\beta^T X)] + \rho\|\beta\|_2^2 \qquad \text{(Exact Martingale DRO)}$$

- The conventional OT-DRO is equivalent to the square-root regression problem [Blanchet et al. (2019)], i.e.,

$$\min_{\beta} \left(\sqrt{\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\beta^T X)]} + \sqrt{\rho}\|\beta\|_2\right)^2 \qquad \text{(OT-DRO)}$$

Introducing an additional power in norm regularization

$$\Longleftrightarrow$$

Adding martingale constraints in the perturbations

Interpolation?

Can we interpolate between the OT-DRO and Martingale DRO models, and produce new regularization techniques?

Interpolation?

Can we interpolate between the OT-DRO and Martingale DRO models, and produce new regularization techniques?

Outline

Perturbed Martingale DRO

We focus on

$$\inf_{\beta} \sup_{\mathbb{Q} \in B_{\rho}^{M,\epsilon}(\hat{\mathbb{P}})} \mathbb{E}_{X \sim \mathbb{Q}}[\ell(f_{\beta}(X))], \qquad \text{(Martingale DRO)}$$

where $B_{\rho}^{M,\epsilon}(\hat{\mathbb{P}})$ is perturbed martingale based ambiguity set, i.e.,

$$B_{\rho}^{M,\epsilon}(\hat{\mathbb{P}}) = \{\mathbb{Q} : D(\mathbb{Q}, \hat{\mathbb{P}}) \le \rho\} \cap \boxed{\left\{\mathbb{Q} : \left\| \mathbb{E}_{\mathbb{Q}|\hat{\mathbb{P}}}[\bar{X}|X] - X \right\|_2 \le \epsilon \right\}}.$$

Perturbed Martingale DRO

We focus on

$$\inf_{\beta} \sup_{\mathbb{Q} \in B_{\rho}^{M,\epsilon}(\hat{\mathbb{P}})} \mathbb{E}_{X \sim \mathbb{Q}}[\ell(f_\beta(X))], \qquad \text{(Martingale DRO)}$$

where $B_{\rho}^{M,\epsilon}(\hat{\mathbb{P}})$ is perturbed martingale based ambiguity set, i.e.,

$$B_{\rho}^{M,\epsilon}(\hat{\mathbb{P}}) = \{\mathbb{Q} : D(\mathbb{Q}, \hat{\mathbb{P}}) \le \rho\} \cap \boxed{\left\{\mathbb{Q} : \left\| \mathbb{E}_{\mathbb{Q}|\hat{\mathbb{P}}}[\bar{X}|X] - X \right\|_2 \le \epsilon \right\}}.$$

- When ϵ is small (cf. $\epsilon^2 \le \rho$), Martingale DRO will reduce to the well-known Jacobian/input gradient regularization.

Perturbed Martingale DRO

We focus on

$$\inf_{\beta} \sup_{\mathbb{Q} \in B_{\rho}^{M,\epsilon}(\hat{\mathbb{P}})} \mathbb{E}_{X \sim \mathbb{Q}}[\ell(f_\beta(X))], \qquad \text{(Martingale DRO)}$$

where $B_{\rho}^{M,\epsilon}(\hat{\mathbb{P}})$ is perturbed martingale based ambiguity set, i.e.,

$$B_{\rho}^{M,\epsilon}(\hat{\mathbb{P}}) = \{\mathbb{Q} : D(\mathbb{Q}, \hat{\mathbb{P}}) \leq \rho\} \cap \boxed{\left\{\mathbb{Q} : \left\|\mathbb{E}_{\mathbb{Q}|\hat{\mathbb{P}}}[\bar{X}|X] - X\right\|_2 \leq \epsilon\right\}}.$$

- When ϵ is large (cf. $\epsilon^2 \geq N\rho$), Martingale DRO will reduce to the conventional OT-DRO.

A New Principled Adversarial Training Procedure

- By strong duality theorem developed in our paper, Martingale DRO is identical to

$$\inf_{\lambda \geq 0, \alpha, \beta} \lambda\rho + \frac{\epsilon}{N} \sum_{i=1}^{N} \|\alpha_i\| + \frac{1}{N} \sum_{i=1}^{N} \sup_{\Delta_i} \left[\ell(f_\beta(X_i + \Delta_i)) - \alpha_i^\top \Delta_i - \lambda \|\Delta_i\|^2 \right].$$

[1] https://github.com/duchi-lab/certifiable-distributional-robustness

A New Principled Adversarial Training Procedure

- By strong duality theorem developed in our paper, Martingale DRO is identical to

$$\inf_{\lambda \geq 0, \alpha, \beta} \lambda\rho + \frac{\epsilon}{N} \sum_{i=1}^{N} \|\alpha_i\| + \frac{1}{N} \sum_{i=1}^{N} \sup_{\Delta_i} \left[\ell(f_\beta(X_i + \Delta_i)) - \alpha_i^\top \Delta_i - \lambda \|\Delta_i\|^2 \right].$$

- Regarding the dual variable λ as a constant [Sinha et al. (2018)], we have

$$\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \max_{\|\Delta_i\| \leq \epsilon} \left[\ell(f_\beta(X_i + \Delta_i)) - \lambda \|\Delta_i\|^2 \right].$$

[1] https://github.com/duchi-lab/certifiable-distributional-robustness

A New Principled Adversarial Training Procedure

- By strong duality theorem developed in our paper, Martingale DRO is identical to

$$\inf_{\lambda \geq 0, \alpha, \beta} \lambda\rho + \frac{\epsilon}{N}\sum_{i=1}^{N}\|\alpha_i\| + \frac{1}{N}\sum_{i=1}^{N}\sup_{\Delta_i}\left[\ell(f_\beta(X_i+\Delta_i)) - \alpha_i^\top \Delta_i - \lambda\|\Delta_i\|^2\right].$$

- Regarding the dual variable λ as a constant [Sinha et al. (2018)], we have

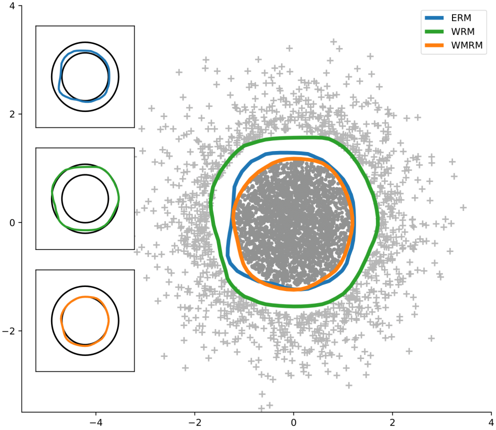
$$\min_{\beta} \frac{1}{N}\sum_{i=1}^{N}\max_{\|\Delta_i\|\leq\epsilon}\left[\ell(f_\beta(X_i+\Delta_i)) - \lambda\|\Delta_i\|^2\right].$$

- Can be addressed by SGD efficiently (only change 3 lines of *Pytorch* code)! [1]

[1] https://github.com/duchi-lab/certifiable-distributional-robustness

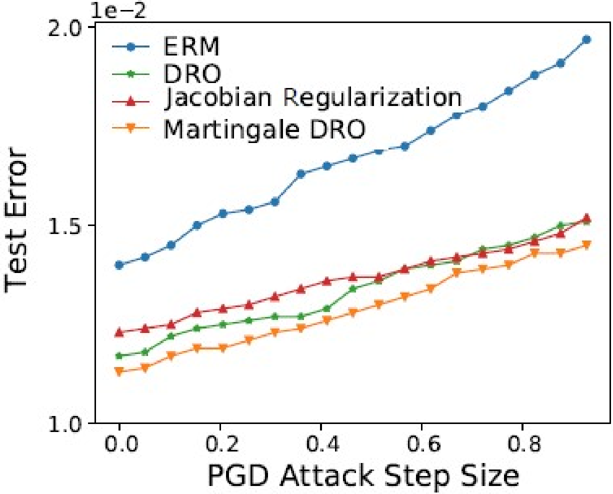
Outline

Toy Example for Binary Classification



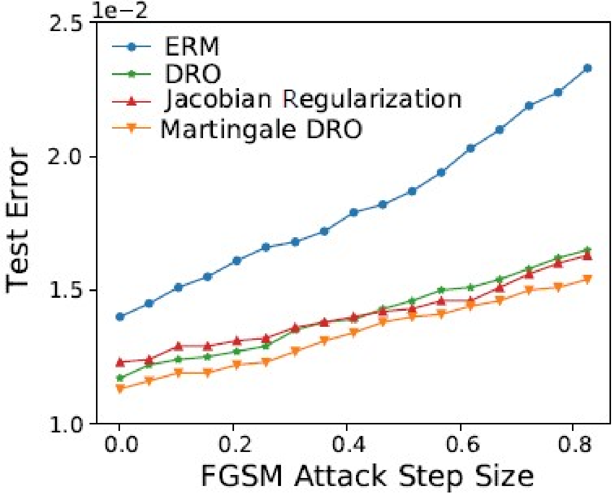
Deep Neural Network for Adversarial Training

MINIST Dataset:



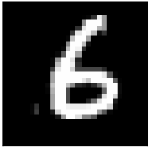
Deep Neural Network for Adversarial Training

MINIST Dataset:

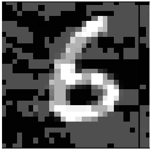


Deep Neural Network for Adversarial Training

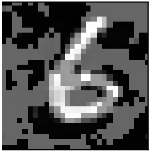
The largest DRO perturbation such that each model makes correct prediction:



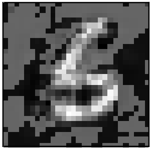
(a) Original



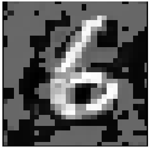
(b) ERM



(c) Jacobian Regularization



(d) DRO



(e) Martingale DRO

Summary

- Tikhonov regularization is distributionally robust in a non-paramteric sense when exact martingale constraints are imposed to the conventional DRO model.

Summary

- Tikhonov regularization is distributionally robust in a non-paramteric sense when exact martingale constraints are imposed to the conventional DRO model.

- The interpolation between the conventional OT-DRO and the exact martingale DRO models (Perturbed Martingale DRO) can result in a novel and effective set of regularizer techniques.

Reference

Jiajin Li, Sirui Lin, Jose Blanchet, Viet Anh Nguyen "Tikhonov Regularization is Optimal Transport Robust under Martingale Constraints." Accepted by NeurIPS 2022.

Thank you! Questions?