Some Recent Advances in Distributionally Robust Optimization

Anthony Man-Cho So

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong (CUHK)

School of Management and Engineering
Nanjing University

August 28, 2020

Outline

Wasserstein Distributionally Robust Risk Minimization

Optimistic Likelihood Estimation

Wasserstein Distributionally Robust Risk Minimization

Consider the distributionally robust risk minimization problem

$$\inf_{\beta} \sup_{\mathbb{Q}\in\mathcal{M}_\epsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{(x,y)\sim\mathbb{Q}}[\ell(f_\beta(x),y)] \qquad (*)$$

with

$$\mathcal{M}_\epsilon(\widehat{\mathbb{P}}_N) := \{\mathbb{Q} : W(\mathbb{Q},\widehat{\mathbb{P}}_N) \le \epsilon\}.$$

Wasserstein Distributionally Robust Risk Minimization

Consider the distributionally robust risk minimization problem

$$\inf_{\beta} \sup_{\mathbb{Q} \in \mathcal{M}_\epsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{(x,y)\sim\mathbb{Q}}[\ell(f_\beta(x), y)] \qquad (*)$$

with

$$\mathcal{M}_\epsilon(\widehat{\mathbb{P}}_N) := \{\mathbb{Q} : W(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \epsilon\}.$$

- ✓ strong connection to regularization techniques in machine learning
- ✓ good generalization properties and confidence interval guarantees under minimal assumptions
- ✓ in most cases, Problem (*) admits an equivalent, efficiently solvable convex reformulation via duality for the inner sup [Shafieezadeh-Abadeh et al., 2019]

Distributionally Robust Logistic Regression

Recall the setting:

- $x \in \mathbb{R}^n$ and $y \in \{-1, +1\}$;
- $\ell(u, v) = \log(1 + \exp(-uv))$ and $f_\beta(x) = \beta^T x$;
- $d(z, z') = \|x - x'\| + \kappa|y - y'|$ with $\kappa > 0$ and $\|\cdot\|$ being a generic norm on $\mathbb{R}^n$ (recall $z = (x, y)$).

Theorem [Shafieezadeh-Abadeh et al., 2019]

Problem (*) is equivalent to

$$\begin{array}{lll} \inf\limits_{\beta,\, s,\, \lambda} & \lambda\epsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i & \\ \text{subject to} & \ell(\beta^T \hat{x}_i, \hat{y}_i) \le s_i, \quad \forall i, & \\ & \ell(\beta^T \hat{x}_i, -\hat{y}_i) - \lambda\kappa \le s_i, \quad \forall i, & \\ & \|\beta\|_* \le \lambda. & \end{array} \tag{DRLR}$$

Here, $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Wasserstein Distributionally Robust Risk Minimization

- Most existing works solve the convex reformulations of the Wasserstein distributionally robust risk minimizaton problem (*) using standard solvers.
  - does not scale well with problem size

Wasserstein Distributionally Robust Risk Minimization

- Most existing works solve the convex reformulations of the Wasserstein distributionally robust risk minimizaton problem (*) using standard solvers.
  - does not scale well with problem size
- Question: Can we develop practically efficient methods with provable guarantees to solve Problem (*)?

Wasserstein Distributionally Robust Risk Minimization

- Most existing works solve the convex reformulations of the Wasserstein distributionally robust risk minimizaton problem (*) using standard solvers.
  - does not scale well with problem size
- Question: Can we develop practically efficient methods with provable guarantees to solve Problem (*)?
- More generally, the development of fast numerical methods for solving distributionally robust optimization problems is still in its infancy stage.
  - progress on this front will help realize the benefits of the distributionally robust optimization approach

First-Order Algorithmic Framework for (DRLR)

Recall that

$$
\begin{aligned}
\inf_{\beta,\, s,\, \lambda} \quad & \lambda\epsilon + \frac{1}{N}\sum_{i=1}^{N} s_i \\
\text{subject to} \quad & \ell(\beta^T\hat{x}_i, \hat{y}_i) \le s_i, \quad \forall i, \\
& \ell(\beta^T\hat{x}_i, -\hat{y}_i) - \lambda\kappa \le s_i, \quad \forall i, \\
& \|\beta\|_* \le \lambda.
\end{aligned}
\tag{DRLR}
$$

First-Order Algorithmic Framework for (DRLR)

Recall that

$$
\begin{aligned}
\inf_{\beta,\, s,\, \lambda} \quad & \lambda\epsilon + \frac{1}{N}\sum_{i=1}^{N} s_i \\
\text{subject to} \quad & \ell(\beta^T \hat{x}_i, \hat{y}_i) \le s_i, \quad \forall i, \\
& \ell(\beta^T \hat{x}_i, -\hat{y}_i) - \lambda\kappa \le s_i, \quad \forall i, \\
& \|\beta\|_* \le \lambda.
\end{aligned}
\tag{DRLR}
$$

- By considering the KKT conditions of Problem (DRLR), one can establish an upper bound λ^U on the optimal λ^*.

First-Order Algorithmic Framework for (DRLR)

Recall that

$$\begin{array}{ll} \inf\limits_{\beta,\, s,\, \lambda} & \lambda\epsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{subject to} & \ell(\beta^T \hat{x}_i, \hat{y}_i) \le s_i, \quad \forall i, \\ & \ell(\beta^T \hat{x}_i, -\hat{y}_i) - \lambda\kappa \le s_i, \quad \forall i, \\ & \|\beta\|_* \le \lambda. \end{array} \tag{DRLR}$$

- By considering the KKT conditions of Problem (DRLR), one can establish an upper bound λ^U on the optimal λ^*.
- This suggests the following strategy for solving (DRLR):
  - initialize λ to a value in $[0, \lambda^U]$
  - solve the resulting problem for β
  - perform an one-dimensional search to update λ
  - repeat

First-Order Algorithmic Framework for (DRLR)

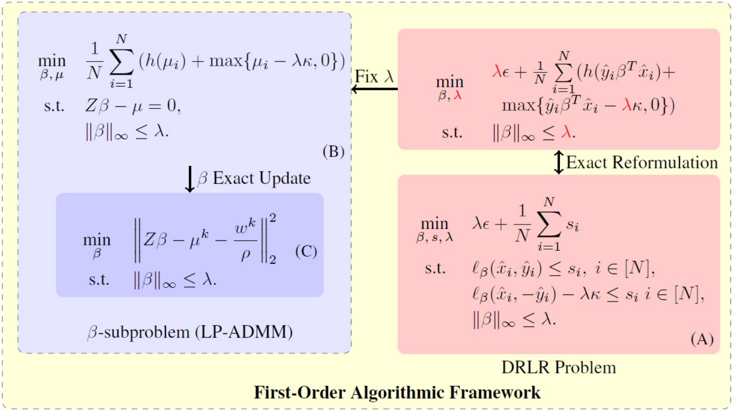

Figure: Proposed Algorithmic Framework with ℓ_1-induced Transport Cost

First-Order Algorithmic Framework for (DRLR)

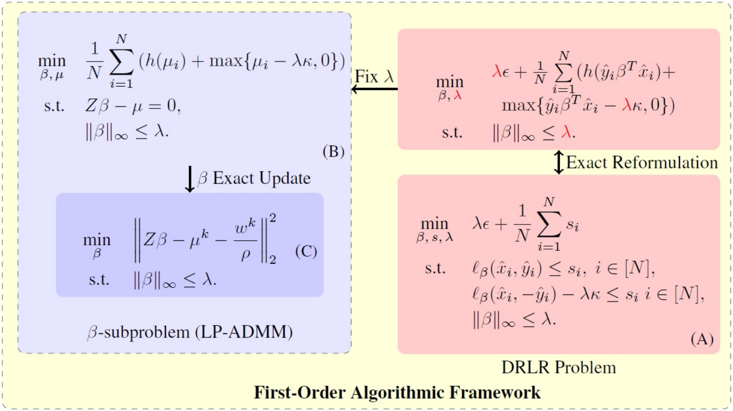

Figure: Proposed Algorithmic Framework with ℓ_1-induced Transport Cost

- We developed a linearized proximal ADMM (LP-ADMM) to solve Problem (B) and established its sublinear convergence. Li et al.: A First-Order Algorithmic Framework for Wasserstein Distributionally Robust Logistic Regression. NeurIPS 2019.

Numerical Results

Table: Comparison of CPU times of YALMIP (solver used in [Shafieezadeh-Abadeh et al., 2015]) and LP-ADMM on UCI adult datasets from LIBSVM [Chang and Lin, 2011]

Dataset	Data Statistics		CPU Time (s)		Ratio
	Samples	Features	YALMIP	LP-ADMM	
a1a	1605	123	25.63	**2.93**	**9**
a2a	2265	123	39.20	**3.53**	**11**
a3a	3185	123	57.79	**4.26**	**14**
a4a	4781	123	105.32	**4.56**	**23**
a5a	6414	123	155.42	**4.39**	**35**
a6a	11220	123	413.65	**4.68**	**88**
a7a	16100	123	738.12	**5.41**	**137**
a8a	22696	123	1396.45	**5.81**	**240**
a9a	32561	123	2993.30	**7.08**	**423**

Neural Network Training

- Consider now a more general setting:

$$\inf_{\beta} \sup_{\mathbb{Q} \in \mathcal{M}_\epsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{(x,y)\sim\mathbb{Q}}[\ell_\beta(x,y)], \qquad (*)$$

  where ℓ may not even be convex.
- This arises, e.g., in the adversarial training of neural networks, where the goal is to protect against adversarial perturbations in the training data set.

Neural Network Training

- Consider now a more general setting:

$$\inf_{\beta} \sup_{\mathbb{Q} \in \mathcal{M}_\epsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{(x,y)\sim\mathbb{Q}}[\ell_\beta(x,y)], \tag{*}$$

  where ℓ may not even be convex.
- This arises, e.g., in the adversarial training of neural networks, where the goal is to protect against adversarial perturbations in the training data set.
- One idea to tackle (*) is to consider its Lagrangian relaxation:

$$\inf_{\beta} \left\{ \sup_{\mathbb{Q}} \left(\mathbb{E}_{\mathbb{Q}}[\ell_\beta(x,y)] - \gamma W(\mathbb{Q}, \widehat{\mathbb{P}}_N) \right) \right\}. \tag{LR}$$

Neural Network Training

- Such a formulation has been explored in
  Sinha et al.: Certifying Some Distributional Robustness with Principled Adversarial Training. arXiv, 2017.
- The authors proposed to tackle (LR) using stochastic gradient descent and established some interesting theoretical results.

Outline

Wasserstein Distributionally Robust Risk Minimization

Optimistic Likelihood Estimation

Optimistic Likelihood Estimation

- Consider a set of i.i.d. data points $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M]$ with $\boldsymbol{x}_i \in \mathbb{R}^n$. The data points are generated from one of several Gaussian distributions $\mathbb{P}_1, \ldots, \mathbb{P}_C$.
- We are interested in determining the distribution $\mathbb{P}_{c^\star}$ such that $\boldsymbol{X}$ has the highest likelihood across $\{\mathbb{P}_c\}_{c=1}^C$. The likelihood function is given by

$$\ell(\boldsymbol{X}, \mathbb{P}_c) = -\frac{1}{M} \sum_{i=1}^{M} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_c) - \log\det \boldsymbol{\Sigma}_c,$$

  where $(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ are the mean and covariance of $\mathbb{P}_c$. In particular,

$$c^\star \in \operatorname*{arg\,max}_{c \in \{1,\ldots,C\}} \ell(\boldsymbol{X}, \mathbb{P}_c).$$

Optimistic Likelihood Estimation

- Usually, we only have estimates of $(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ from the training data, which results in an estimated distribution $\widehat{\mathbb{P}}_c$.
- To guard against misspecification of the distribution, we consider replacing the likelihood function ℓ by the following *optimistic likelihood*:

$$\ell_{\mathsf{DR}}(\boldsymbol{X}, c) = \max_{\mathbb{P} \in \mathcal{P}_c} \ell(\boldsymbol{X}, \mathbb{P}),$$

  where
  - $\mathcal{P}_c = \{\mathbb{P} \in \mathcal{M} : \varphi(\widehat{\mathbb{P}}_c, \mathbb{P}) \leq \rho_c\}$;
  - $\mathcal{M}$: set of non-degenerate Gaussian distributions on $\mathbb{R}^n$;
  - φ: dissimilarity measure satisfying $\varphi(\mathbb{P}, \mathbb{P}) = 0$ for all $\mathbb{P} \in \mathcal{M}$;
  - $\rho_c > 0$: radius of the uncertainty set $\mathcal{P}_c$.

Optimistic Likelihood Estimation

- Usually, we only have estimates of $(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ from the training data, which results in an estimated distribution $\widehat{\mathbb{P}}_c$.
- To guard against misspecification of the distribution, we consider replacing the likelihood function ℓ by the following *optimistic likelihood*:

$$\ell_{\mathsf{DR}}(\boldsymbol{X}, c) = \max_{\mathbb{P} \in \mathcal{P}_c} \ell(\boldsymbol{X}, \mathbb{P}),$$

  where
  - $\mathcal{P}_c = \{\mathbb{P} \in \mathcal{M} : \varphi(\widehat{\mathbb{P}}_c, \mathbb{P}) \leq \rho_c\}$;
  - $\mathcal{M}$: set of non-degenerate Gaussian distributions on $\mathbb{R}^n$;
  - φ: dissimilarity measure satisfying $\varphi(\mathbb{P}, \mathbb{P}) = 0$ for all $\mathbb{P} \in \mathcal{M}$;
  - $\rho_c > 0$: radius of the uncertainty set $\mathcal{P}_c$.
- Now, we are interested in the distributionally robust optimization problem

$$c^{\star}_{\mathsf{DR}} \in \arg \max_{c \in \{1,\ldots,C\}} \ell_{\mathsf{DR}}(\boldsymbol{X}, c).$$

Optimistic Likelihood Estimation

- Consider the scenario where the mean $\hat{\boldsymbol{\mu}}$ is fixed. Then, the space of non-degenerate Gaussian distributions can be parametrized by $\mathbb{S}^n_{++}$. This is a manifold.
- Various dissimilarity measures induce different Riemannian metrics on $\mathbb{S}^n_{++}$.
  - Wasserstein φ_W
  - Fisher-Rao φ_{FR}: Given Gaussian distributions $\mathcal{N}(\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_2)$, the FR distance is defined as

$$\varphi(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = \frac{1}{\sqrt{2}} \| \log(\boldsymbol{\Sigma}_2^{-1/2} \boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2^{-1/2}) \|_F.$$

Optimistic Likelihood Estimation

- Using the FR distance, optimistic likelihood estimation reduces to a geodesically convex optimization problem on $\mathbb{S}^n_{++}$, for which efficient algorithms can be developed.
  Nguyen et al.: Calculating Optimistic Likelihoods Using (Geodesically) Convex Optimization. NeurIPS 2019.

References I

Chang, C.-C. and Lin, C.-J. (2011).
LIBSVM: A Library for Support Vector Machines.
ACM Transactions on Intelligent Systems and Technology (TIST), 2(3):Article 27.

Shafieezadeh-Abadeh, S., Kuhn, D., and Mohajerin Esfahani, P. (2019).
Regularization via Mass Transportation.
Journal of Machine Learning Research, 20(103):1–68.

Shafieezadeh-Abadeh, S., Mohajerin Esfahani, P., and Kuhn, D. (2015).
Distributionally Robust Logistic Regression.
In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28: Proceedings of the 2015 Conference*, pages 1576–1584.