Optimisation distributionnellement robuste : régularisations et applications en learning

Jérôme Malick

CNRS, LJK, Université Grenoble Alpes
jerome.malick@univ-grenoble-alpes.fr

Mots-clés : *optimisation sous incertitude, data-driven optimization, distance de Wasserstein, régularisation/pénalisation, approximation, machine learning.*

DRO : contexte. Les problèmes de décision en présence d'incertitude apparaissent en nombre dans les applications, notamment, en recherche opérationnelle, en économie, et en intelligence artificielle. Lorsque l'incertitude provient d'erreurs de mesures, d'anomalies, ou de modèles inadéquats, l'optimisation distributionnellement robuste (distributionnally robust optimisation, DRO) offre une approche «data-driven» prometteuse. L'idée est de chercher des décisions optimales robustes qui minimisent l'espérance de perte à rapport à la pire distribution dans un voisinage (appelé «ambiguity set») d'une distribution nominale (par exemple, la distribution empirique sur les données).

Tout comme en optimisation robuste, la question centrale en DRO est le choix de cet ambiguity set. Un équilibre doit être trouvé :

— **modélisation** : le voisinage doit être suffisamment riche pour contenir les distributions pertinentes pour la prise de décision, mais, en même temps, il ne doit pas inclure trop de distributions, au risque de produire des décisions trop conservatives ;

— **résolution numérique** : l'incorporation des propriétés de l'application pratique doit se faire en veillant à ce que le problème résultant soit tractable numériquement.

Dans cette présentation, je propose revenir sur ces idées et de discuter de résultats récents (obtenus à l'occasion de la thèse de Yassine Laguel et du stage de M2 de Waiss Azizian) qui viennent d'enrichir la palette d'outils et de modèles disponibles en DRO.

DRO : Wasserstein et plus. Il a été proposé une multitude de formulations DRO en fonction du choix de l'ambiguity set (faisant apparaitre des mesures de risque, comme la CVaR [3]). La distance de Wasserstein, objet clé du transport optimal, est souvent appréciée pour sa richesse de modélisation et ses propriétés statistiques [4]. Cependant, le DRO avec des voisinages définis par Wasserstein (WDRO), dans un contexte data-driven, a quelques inconvénients, sur les deux aspects pratiques (modélisation et résolution numérique).

Pour pallier à ces difficultés, nous proposons dans [2] une étude complète de versions régularisées de WDRO, notamment inspirées par la divergence de Sinkhorn, populaire dans les applications du transport optimal en learning [5]. Nous avons ainsi obtenu :

— de jolis **résultats de dualité généraux** pour des problèmes de WDRO doublement régularisés (par l'ajout d'une fonction convexe dans les contraintes et une dans l'objectif) ;

— des **résultats d'approximation originaux** quantifiant l'erreur d'approximation

$$O\Big(d\,(\varepsilon+\Lambda\delta)\log\left(1/(\epsilon+\Lambda\delta)\right)\Big)$$

où ε et δ sont les deux paramètres de régularisation, d est la dimension du problème, et Λ un majorant explicite de la solution duale optimale.

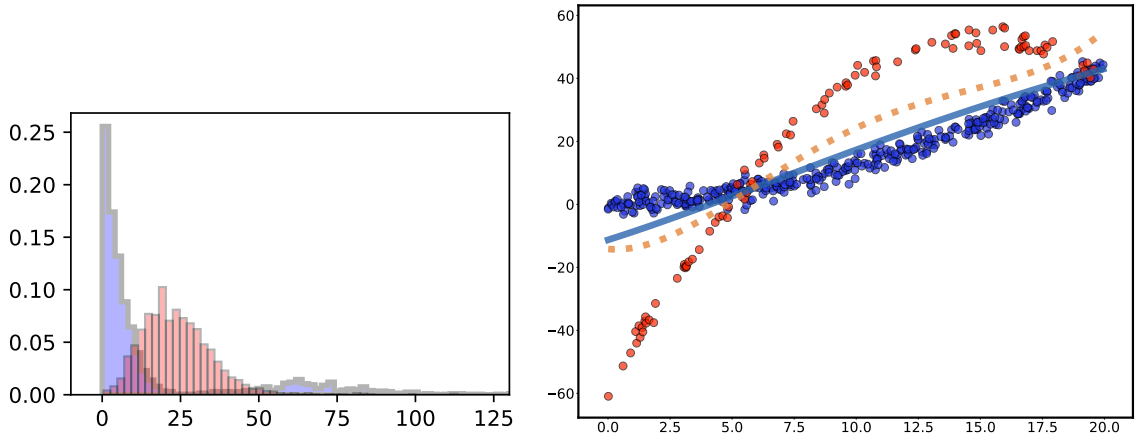

FIG. 1 – Illustrations de l'effet de DRO en learning sur deux problèmes synthétiques. À gauche, comparaison de l'histogramme des pertes pour deux modèles (par moindres carrés DRO vs. moindres carrés classiques) : on voit un «reshaping» de l'histogramme des pertes avec le modèle appris par DRO (en orange) à rapport au classique (en violet). À droite, comparaison de régressions (DRO vs. classique) dans le cas en présence d'un sous-groupe minoritaire (points rouges à coté des points bleus) : on voit que la régression DRO (en pointillés oranges) capte et respecte mieux le comportement du groupe minoritaire (points rouges) par apport à la régression classique (en bleu). Nous expliciterons ces expériences dans la présentation.

DRO à l'oeuvre en learning. L'hypothèse classique en machine learning, qui pose que les données de test ont la même distribution que les données d'entrainement, est mise à mal par des applications mettant en jeu des données hétérogènes [1] ou des questions d'équité [6]. C'est tout naturellement que ces applications font apparaitre des problèmes de DRO (voir le livre [7]) et notamment de WDRO (voir l'article de review [4]).

Dans cette présentation, nous illustrerons ces idées sur des exemples «jouets» (cf Figure 1) ainsi que sur une application industrielle issue des problématiques de Google, déjà esquissée à la fin de [8]. Nous comparerons ces résultats à des approches alternatives stochastiques/robustes.

Références

[1] Yassine Laguel, Jérôme Malick, and Zaid Harchaoui. Optimization for Superquantile-based Supervised Learning. *30th Workshop on Machine Learning for Signal Processing*, 2020.

[2] Azizian Waiss, Franck Iutzeler, and Jérôme Malick. Regularization for Wasserstein distributionnally robust optimization. *Soumis*, 2021.

[3] Terry Rockafellar, Johannes Royset, Sofia Miranda. Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. *European Journal of Operational Research*, 2014

[4] Daniel Kuhn *et al.* Wasserstein distributionally robust optimization : Theory and applications in machine learning. *Operations Research & Management Science in the Age of Analytics*, INFORMS, 2019

[5] Jean Feydy *et al.* Interpolating between optimal transport and MMD using Sinkhorn divergences. *AiStats (Conference on Artificial Intelligence and Statistics)*, 2019.

[6] Julien Ferry. DRO pour améliorer la généralisation de l'équité en apprentissage. *Exposé à la session «RO/apprentissage» à la ROADEF*, 2020.

[7] Ruidi Chen and Ioannis Ch Paschalidis. *Distributionally Robust Learning.* Foundations and Trends in Optimization, Now Publishers Inc., 2020.

[8] Jérôme Malick. Intelligence Opérationnelle. *Tutoriel au congrès de la ROADEF*, Montpellier, 2019.