Optimization in the Space of Measures: ML Using Optimal Transport

2019 Program Review

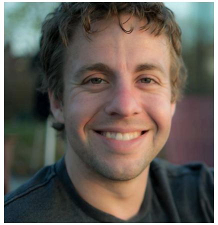
Carson Kent
Stanford University
crkent@stanford.edu

About this talk

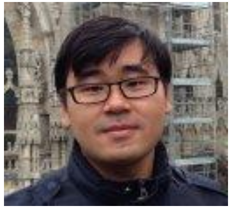
- Introduction to optimal transport
  - Formulation and applications
  - Recent computational advances
  - Hope for the future
- Robust stochastic optimization
  - Approaches using OT
  - Duality and computation
- The roadmap from here

Joint work with:

Jose Blanchet



Aaron Sidford



Ruodu Wang

And others!

Optimal transport according to Monge (1781)

66 MÉMOIRES DE L'ACADÉMIE ROYALE

MÉMOIRE

SUR LA

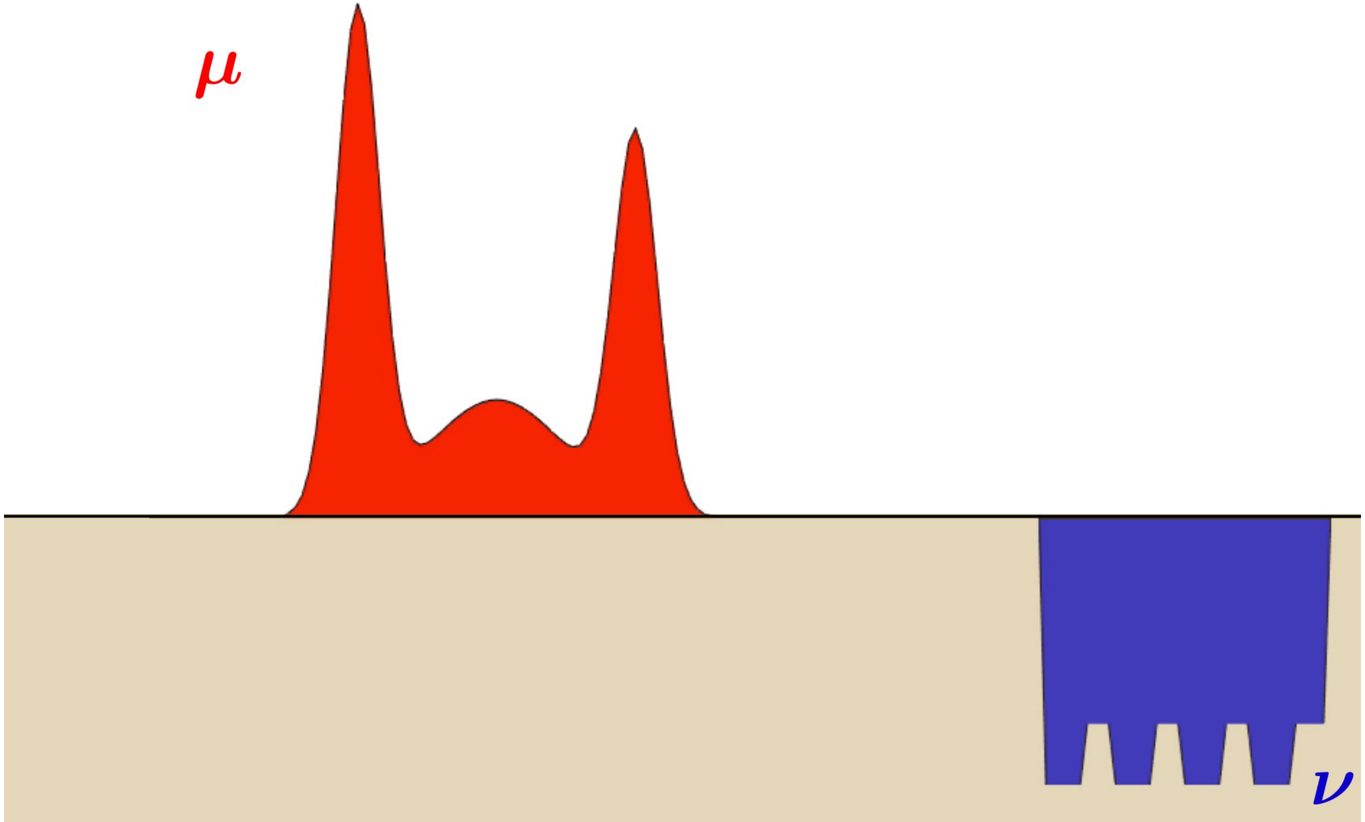
THÉORIE DES DÉBLAIS

ET DES REMBLAIS.

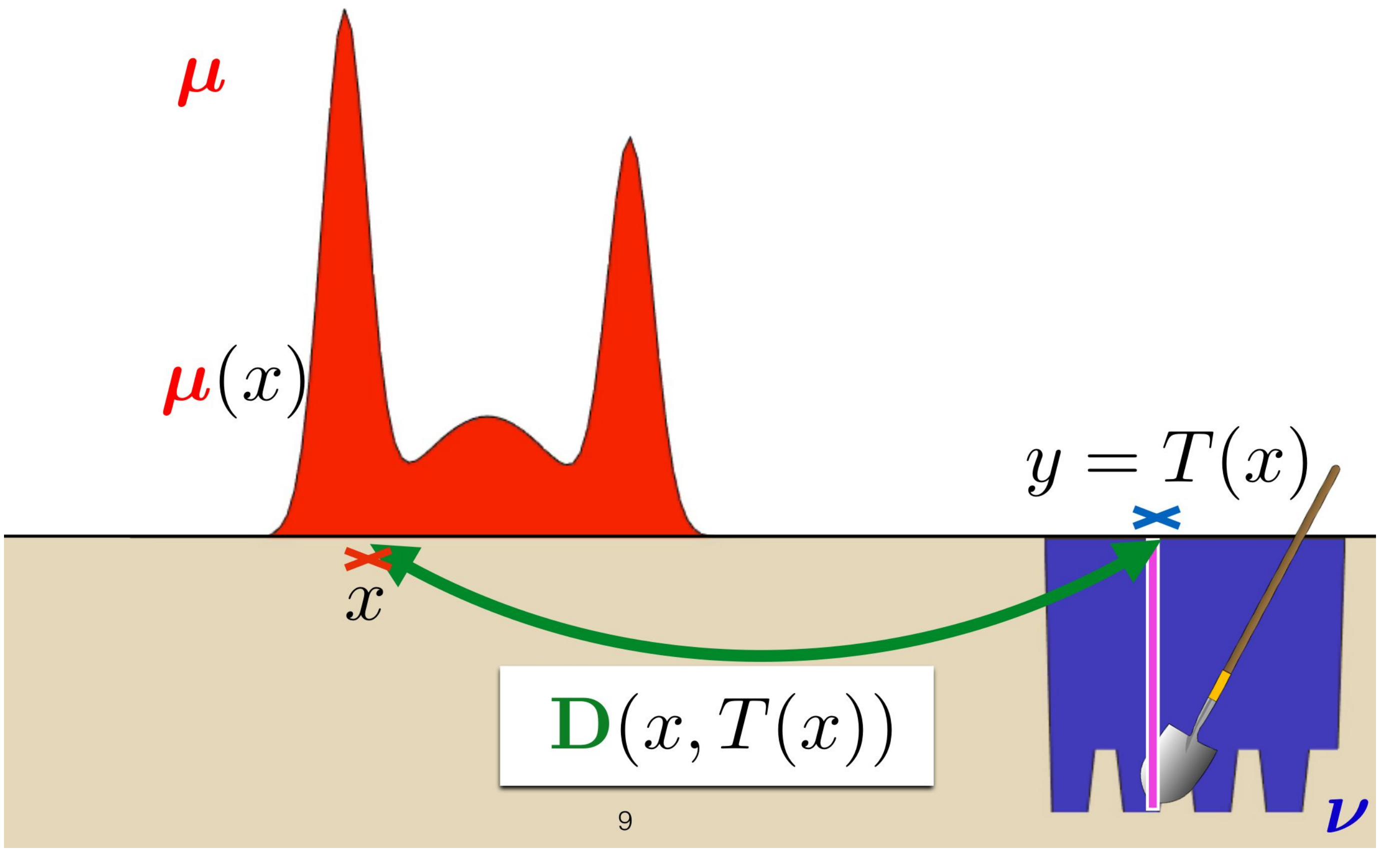
When one has to bring earth from one place to another…

LORSQU'ON doit tranſporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit tranſporter, & le nom de *Remblai* à l'eſpace qu'elles doivent occuper après le tranſport.

In pictures...



In pictures...



Mathematically Speaking

- Two probability distributions: μ, ν
- A sensible transportation cost: $c(X, Y)$
- Compute

$$T^* = \arg\min_{T_{\#}\mu=\nu} \int c(X, T(X))\, d\mu$$

where $T_{\#}\mu = \nu$ means $T(X) \sim \nu$

Kantorovich’s Refinement

- Finding the map $T(X)$ could be an ill-posed problem!
- Better to relax the problem and compute

$$\pi^* = \arg\min_{\pi \in \Pi(\mu,\nu)} \int c(X,Y)\, d\pi$$

where $\Pi(\mu, \nu)$ is the set of all joint distributions over (X, Y) with marginals μ, ν

- An infinite dimensional LP! With an elegant dual!

Rich mathematical structure

- Dual problem (dual variables are continuous, bounded functions)

$$\max_{\phi(X)+\psi(Y)\leq c(X,Y)} \int \phi \, d\mu + \int \psi \, d\nu$$

- When $c(X,Y) = d(X,Y)^p$ for some distance function $d(X,Y)$ we get a notion of distance between distributions-- namely the Wasserstein distance!
- When $p = 1$ the dual become particularly elegant. Can you say Wasserstein GAN?

$$\max_{\phi \in \mathrm{Lip}_1} \int \phi \, (d\mu - d\,\nu)$$

A toy example

- When ν and μ are one dimensional and the cost is "pretty nice," say

$$c(X, Y) = |X - Y|$$

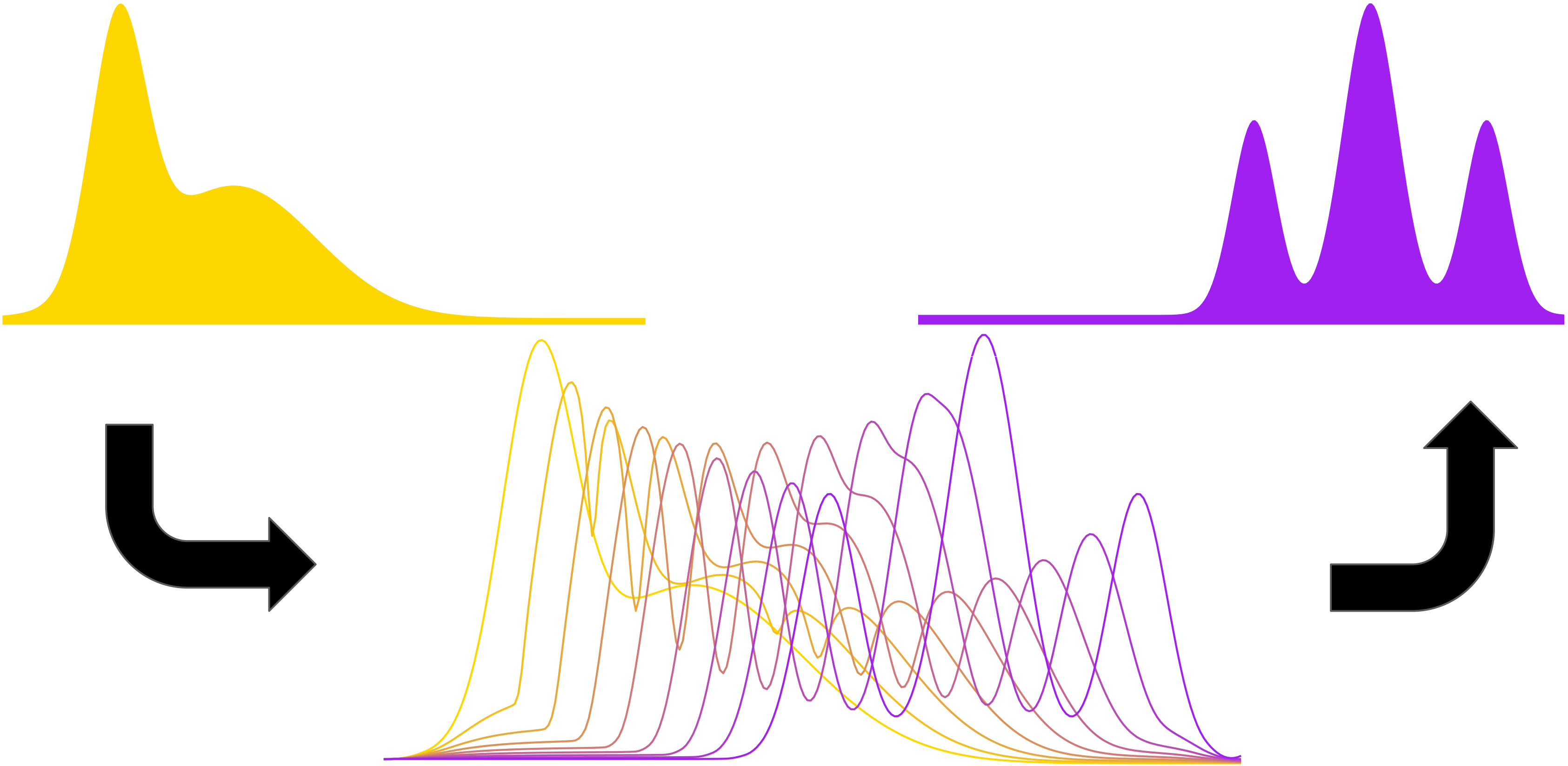
the transport is quite natural

$$T(X) = G^{-1}\left(F\left(X\right)\right)$$

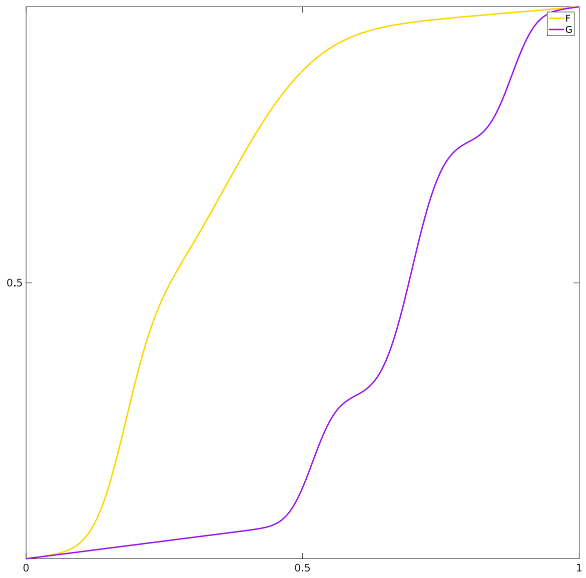
where F, G are the cumulative distribution functions for μ, ν respectively.

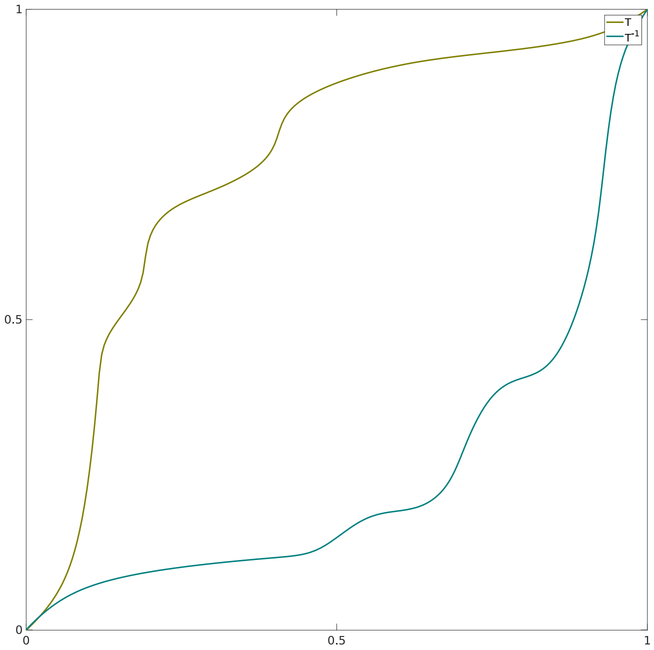
- Intuitively consistent, matches quantiles with "no crossings."
- Highly dependent on the ordering of the real line. Does not generalize to higher dimensions!

Getting the picture...



Getting the picture...



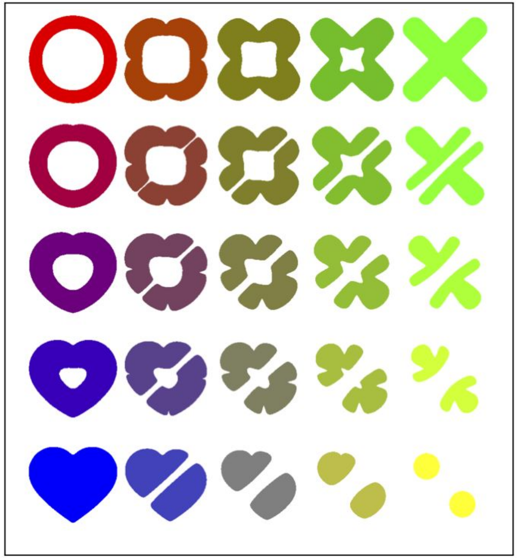


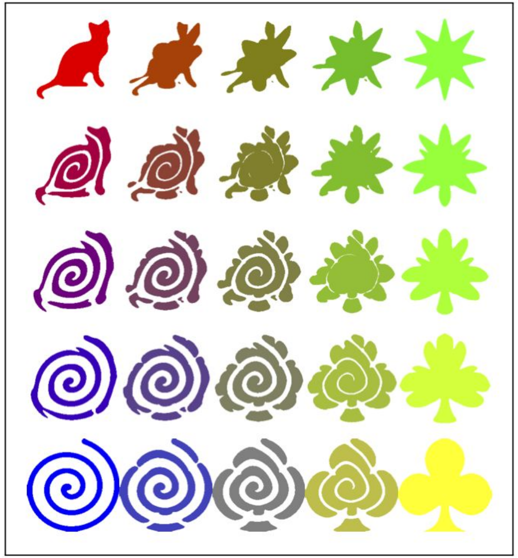
Lots and lots of applications!

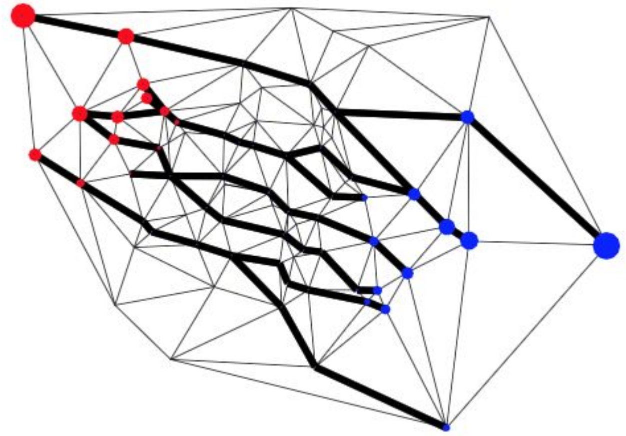
Plus 5 or 6 Nobel prizes and Fields medals

- Assignment and routing
- Contrast equalization and texture synthesis
- Image matching, image fusion, and shape registration
- Market design, robust derivative pricing and risk aggregation
- Embeddings, feature aggregation, and dimensionality reduction
- Music transcription and record restoration
- Drug screening, protein folding, and cancer detection
- Sampling and Bayesian inference
- Robust stochastic optimization*

Photogenic applications







Photogenic applications

Computational Optimal Transport

- Beyond 1-dimension, highly non-trivial to compute either primal

$$\pi^* = \arg\min_{\pi \in \Pi(\mu,\nu)} \int c(X,Y)\, d\pi \qquad \psi^*, \phi^* = \arg\max_{\phi(X)+\psi(Y) \le c(X,Y)} \int \phi\, d\mu + \int \psi\, d\nu$$

- In infinite dimensions one must discretize

$$\nu = \sum_{i \in [N]} \delta_{y_i} \qquad \mu = \sum_{i \in [N]} \delta_{x_i}$$

- Typically, discretization appears in the marginals. Empirical margins (sum of point masses) are assumed.

Computational Optimal Transport

- Under marginal discretization the infinite dimensional LP becomes a finite dimensional!

$$\min_{X \in \mathcal{U}(r,c)} \langle C, X \rangle \qquad \mathcal{U}(r,c) := \{X \in \mathbb{R}_+^{n \times n} : X\mathbf{1} = r, X^T\mathbf{1} = c\}$$

- So the problem is solved? Plug and chug for our favorite LP solver?
- Computational scale will hit you in the face as the curse of dimensionality kicks in.
- X is on the order $O\left(N^2\right)$. In theory we would like $N \sim \frac{1}{\epsilon^n}$, in practice, take as much data as you can get
- At best, the fastest LP solver will get you $O\left(N^{2.5} \log \frac{1}{\epsilon}\right)$. In practice, count on doing worse with such a black box approach

Computational Optimal Transport

- Significantly more structure than your average LP. Constraints provide special structure!
- Key insight along these lines was by Cuturi: regularize with entropy $H(X)$

$$\min_{X \in \mathcal{U}(r,c)} \langle C, X \rangle - \eta H(X)$$

- Taking the dual + alternating minimization = an elegant and practical algorithm (arguably the most popular method for computing OT)
- Recently, it was shown to be almost linear

$$O\left(\frac{\|C\|_\infty N^2}{\epsilon^2}\right)$$

- The dependence on the tolerance is still punishing!

Computational Optimal Transport

- Can we do better?

“There's Plenty of Room at the Bottom”

- Key idea: exploit connections to packing LPs and matrix scaling

$$\max_{x \in \mathbb{R}_+^{N^2}} \{d^T x : Ax \leq b\}$$

- Intuition: apply highly specialized first and second order methods.
- Accelerated coordinate descent (packing LP) and box-constrained Newton method (matrix scaling)

Computational Optimal Transport

- Beats previous complexities (attains best known) and offers scalable parallel depth $$O\left(\frac{\|C\|_\infty N^2}{\epsilon}\right) \text{ in work} \qquad O\left(\frac{1}{\epsilon}\right) \text{ in depth}$$
- Even practical, serial implementations are competitive with Sinkhorn! Additionally, the $O\left(1/\epsilon\right)$ provides greater numerical stability
- Parallel discovery with Kent Quanrud.
- Co-authors recently created a direct, fully first order algorithm with the same parallel depth!
- Same performance attained by two, largely orthogonal methods. Coincidence?

Lower bounds

- Theme: lower bounds particularly in the linear work regime are hard!
- We show that a method which does less than $O\left(N^2/\epsilon\right)$ work would give a algorithm $o\left(m^{2.5}\right)$ for *maximum cardinality bipartite matching*
- Means further computational complexity would be highly surprising!
- Only known algorithms which achieve this running time use fast matrix multiplication. No flow-based algorithms!
- Pseudo-complexity reduction, however, since no formal hardness or information theoretic lower bound.

Further progress

- Sorry Mr. Feynman! There's no more room here! ….Or is there?
- Many costs are highly structured, sub-linear performance of Sinkhorn can still observed in practice. Think 2-norm squared cost!
- Recent work, showing that exploitation of low-rank cost matrix leads to fast, sublinear iterations for Sinkhorn!
- Lesson for computational scientists: when a lower bound hits you in the face, make further assumptions!
- Continued progress is also quite plausible! Most “nice” transports are sparse!

Part 2: How a Practicum Can Inspire You

Robust Optimization in a Nutshell

- Consider a linear program of the form

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & c^T x \\ \text{subject to} \quad & Ax \leq b \end{aligned}$$

in practice we really don't know . Typically we have an estimate and some bounds

- Really, we'd like to compute

$$\begin{aligned} \min_{x} \max_{c \in \mathcal{U}} \quad & c^T x \\ \text{subject to} \quad & Ax \leq b \end{aligned}$$

- Carson's practicum 2017 at ANL was based around robust optimization for nonlinear problems.

Themes of Robust Optimization

- Robust problems might appear to be a complex animal. Formulation is bi-level, necessitating advanced techniques (e.g. mirror-prox, alternating min/maximization)
- Key trick: duality!

$$\begin{aligned} \min_{x,\lambda} \quad & z^T x \\ \text{subject to} \quad & W^T x + \lambda = 0 \\ & Ax \leq b \\ & \lambda \geq 0 \end{aligned}$$

- Robustness of a solution at any level is not computable. If you require your problem will easily become NP-hard!

Wasserstein Robust Optimization

- Consider the robust “infinite dimensional” LP

$$\max_{D_c(\mu,\nu)\leq\delta} \int f\, d\mu$$

where denotes a ball in optimal transport distance.

- Similarly we can use duality to rewrite the problem as

$$\inf_{\lambda\geq 0} \lambda\delta + \int \left(\sup_{Y\in\mathbb{R}^n} f(Y) - \lambda c(X,Y) \right) d\nu$$

- Surprisingly, our infinite dimensional problem has now become a finite dimensional one!
- Lesson: take the dual!

Distributionally Robust Optimization

- Our recent results:
    - Under smoothness assumptions on we precisely quantify the level of robustness for which the problem is computationally solvable
    - We extend duality to the multi-marginal case, i.e. the intersection of multiple balls in optimal transport distance
- Why is this important?
- Allows us to give an actual computational analogue of “Frank-Wolfe in infinite dimensions.” Algorithms with exact parameters as opposed to heuristics are also nice.
- The multi-marginal case allows us to perform stochastic optimization which is robust to violations of independence! Key application: risk aggregation!

Future directions

- If we can do Frank-Wolfe, why not gradient flows in infinite dimensions?
- With multiple-margins, how about robust reinforcement learning?
- How about a general framework for statistical estimators which accurately handle violations of independence?
- Big idea: it’s a wild-west out there in optimal transport and robust optimization!

Acknowledgements

- Many, many thanks to the CSGF for funding this work and incredible support over the past 4 years!
- Many thanks to Lindsey, Lisa, and all the staff at Krell!
- Thanks so much to YOU for making being part of this cohort such wonderful and unique experience!