Frank-Wolfe in Probability Space

Connection with Distributionlly Robust Optimization

Jiajin Li

Department of Management Science and Engineering
Stanford University

Erice, May 2022

Joint work with Carson Kent, Jose Blanchet and Peter Glynn.

Our Focus

Main Story: We propose Frank-Wolfe-type algorithms to well-address

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} J(\mu) \tag{1}$$

where $\mathcal{P}_2(\mathbb{R}^d)$ is the space of probability measures on $\mathbb{R}^d$ with a finite second moment.

Motivation I

Problem (1) is rich and gives rise to a wide range of contemporary applications (Chu, Blanchet and Gylnn, 2019).

- Trivial Embedding: every optimization problem can be written as

$$\min_{\theta \in \mathbb{R}^d} f(\theta) = \min_{\mu \in \mathcal{P}(\mathbb{R}^d)} J(\mu),$$

where $J(\mu) = \int f(\theta)\mu(d\theta)$ and the optimal solution is supported on the set of optimizers.

Motivation I

Problem (1) is rich and gives rise to a wide range of contemporary applications (Chu, Blanchet and Gylnn, 2019).

- Trivial Embedding: every optimization problem can be written as

$$\min_{\theta \in \mathbb{R}^d} f(\theta) = \min_{\mu \in \mathcal{P}(\mathbb{R}^d)} J(\mu).$$

where $J(\mu) = \int f(\theta)\mu(d\theta)$ and the optimal solution is supported on the set of optimizers.

Motivation II

- Barycenter Problems: Let $\lambda_i > 0$ be a set of weights,

$$\min_{\mu} \sum_{i=1}^{m} \lambda_i D(\mu, \mu_i)$$

where $D(\mu, \mu_i)$ is a discrepancy between μ and μ_i.

- Generative Adversarial Network (Goodfellow et al. 2014): It takes the form of

$$J(\mu) = D(\mu, \mu_n) + R(\mu)$$

for a suitable discrepancy (e.g., Wasserstein, f-divergence) $D(\cdot, \cdot)$ and a regularization term $R(\cdot)$.

Motivation II

- Barycenter Problems: Let $\lambda_i > 0$ be a set of weights,

$$\min_{\mu} \sum_{i=1}^{m} \lambda_i D(\mu, \mu_i)$$

where $D(\mu, \mu_i)$ is a discrepancy between μ and μ_i.

- Generative Adversarial Network (Goodfellow et al. 2014): It takes the form of

$$J(\mu) = D(\mu, \mu_n) + R(\mu)$$

for a suitable discrepancy (e.g., Wasserstein, f-divergence) $D(\cdot, \cdot)$ and a regularization term $R(\cdot)$.

Motivation III

- Variational Inference:

$$J(\mu) = \mathrm{KL}(\mu \| \mu_n).$$

- Mean-Field Games: Population risk for two-layer neural network (Mei, Montanari and Nguyen, 2018).
- Reinforcement Learning ...

Motivation III

- Variational Inference:

$$J(\mu) = \mathrm{KL}(\mu \| \mu_n).$$

- Mean-Field Games: Population risk for two-layer neural network (Mei, Montanari and Nguyen, 2018).

- Reinforcement Learning ...

Motivation III

- Variational Inference:

$$J(\mu) = \mathrm{KL}(\mu \| \mu_n).$$

- Mean-Field Games: Population risk for two-layer neural network (Mei, Montanari and Nguyen, 2018).
- Reinforcement Learning ...

Frank-Wolfe from $\mathbb{R}^d$ to $\mathcal{P}_2(\mathbb{R}^d)$

Frank-Wolfe in $\mathbb{R}^d$

- Assuming that $\mu \in \mathbb{R}^d$ and $J(\cdot)$ is differentiable, we can iteratively solve

$$\min_{\mu \in D} \quad \underbrace{\nabla J(\mu_0)^T(\mu - \mu_0)}_{\text{Directional Derivative } J'(\mu_0;\, \mu - \mu_0)}$$

  where D is a compact convex set.

- Extended to probability space, a natural way is to invoke Gateaux derivative (i.e, influence function) to act as an analogy.

Frank-Wolfe in $\mathbb{R}^d$

- Assuming that $\mu \in \mathbb{R}^d$ and $J(\cdot)$ is differentiable, we can iteratively solve

$$\min_{\mu \in D} \quad \underbrace{\nabla J(\mu_0)^T(\mu - \mu_0)}_{\text{Directional Derivative } J'(\mu_0;\, \mu - \mu_0)}$$

  where D is a compact convex set.

- Extended to probability space, a natural way is to invoke Gateaux derivative (i.e, influence function) to act as an analogy.

Frank-Wolfe in $\mathcal{P}_2(\mathbb{R}^d)$

Gateaux derivative

For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $t \in [0,1]$ and $(1-t)\mu + t\nu \in \mathcal{P}_2(\mathbb{R}^d)$

$$\lim_{t \to 0} \frac{J(\mu + t(\nu - \mu)) - J(\mu)}{t} = \int DJ_\mu(x)\nu(dx) - \int DJ_\mu(x)\mu(dx)$$
$$:= \langle DJ_\mu, \nu - \mu \rangle.$$

It results in the vanilla Frank-Wolfe extension:

$$\inf_{\mu \in \mathcal{P}_2(\mathcal{R}^d)} \langle DJ_{\mu_0}, \mu - \mu_0 \rangle. \tag{2}$$

Modified Frank-Wolfe in $\mathcal{P}_2(\mathbb{R}^d)$

- The vanilla Frank-Wolfe (2) may be not well-defined, when the distribution do not have compact support (i.e., DJ_{μ_0} may be unbounded). It motivates us to conduct a natural modification:

$$\inf_{\mu \in \mathcal{P}_2(\mathcal{R}^d) \cap \{\mu : W(\mu,\mu_0) \leqslant \delta\}} \langle DJ_{\mu_0}, \mu - \mu_0 \rangle. \tag{3}$$

- 2-Wasserstein Distance: Let $\Pi(\mu, \nu)$ be the class of joint distributions π of random variables (X, Y) such that

$$W^2(\mu, \nu) := \min_{\pi} \left\{ \mathbb{E}_\pi \left[\|X - Y\|_2^2 \right] : \pi \in \Pi(\mu, \nu), \pi_X = \mu, \pi_Y = \nu \right\}.$$

Modified Frank-Wolfe in $\mathcal{P}_2(\mathbb{R}^d)$

- The vanilla Frank-Wolfe (2) may be not well-defined, when the distribution do not have compact support (i.e., DJ_{μ_0} may be unbounded). It motivates us to conduct a natural modification:

$$\inf_{\mu \in \mathcal{P}_2(\mathcal{R}^d) \cap \{\mu : W(\mu, \mu_0) \leqslant \delta\}} \langle DJ_{\mu_0}, \mu - \mu_0 \rangle. \tag{3}$$

- 2-Wasserstein Distance: Let $\Pi(\mu, \nu)$ be the class of joint distributions π of random variables (X, Y) such that

$$W^2(\mu, \nu) := \min_{\pi} \left\{ \mathbb{E}_\pi \left[\|X - Y\|_2^2 \right] : \pi \in \Pi(\mu, \nu), \pi_X = \mu, \pi_Y = \nu \right\}.$$

Connection with Distributionally Robust Optimization (DRO)

Modified FW Step and DRO

- By the strong duality theorem developed in (Blanchet and Murthy 2019), we have

$$\begin{aligned}&\inf_{\mu:W(\mu,\mu_0)\leqslant\delta}\int DJ_{\mu_0}(x)\mu(dx)\\&=\max_{\lambda\geqslant 0}\left(E_{\mu_0}\left[\inf_y\left[DJ_{\mu_0}(y)+\lambda\|X-y\|_2^2\right]\right]+\frac{\lambda\delta^2}{2}\right).\end{aligned}\tag{4}$$

- Given $X\sim\mu_0$ (i.e., empirical distribution) and λ is fixed, $Y=\arg\min_y[DJ_{\mu_0}(y)+\lambda\|X-y\|_2^2]$ can be computed in a parallel fashion over all particles.

Modified FW Step and DRO

- By the strong duality theorem developed in (Blanchet and Murthy 2019), we have

$$\inf_{\mu: W(\mu,\mu_0)\leqslant\delta} \int DJ_{\mu_0}(x)\mu(dx) = \max_{\lambda\geqslant 0}\left(E_{\mu_0}\left[\inf_y \left[DJ_{\mu_0}(y) + \lambda\|X - y\|_2^2\right]\right] + \frac{\lambda\delta^2}{2}\right). \tag{5}$$

- Our choice of δ will make sure the inner optimization is strongly convex — accelerated gradient descent with linear convergence rate when we assume the L-smoothness of DJ_{μ_0}.

Modified FW Step and DRO

- By the strong duality theorem developed in (Blanchet and Murthy 2019), we have

$$\inf_{\mu: W(\mu,\mu_0)\leqslant\delta} \int DJ_{\mu_0}(x)\mu(dx) \\ = \max_{\lambda\geqslant 0}\left(E_{\mu_0}\left[\inf_y \left[DJ_{\mu_0}(y)+\lambda\|X-y\|_2^2\right]\right]+\frac{\lambda\delta^2}{2}\right). \tag{6}$$

- Uniform strategy for the dual variable $\lambda > 0$ or bisection method.

Modified FW: Advantages in Probability Space

- It avoids a fixed finite dimensional parameterization in favor of sampling based approximations.
- It has strong connections with Wasserstein distributionally robust optimization (DRO).
- It suggests a parallelizable particle based algorithm.

Modified FW: Advantages in Probability Space

- It avoids a fixed finite dimensional parameterization in favor of sampling based approximations.
- It has strong connections with Wasserstein distributionally robust optimization (DRO).
- It suggests a parallelizable particle based algorithm.

Modified FW: Advantages in Probability Space

- It avoids a fixed finite dimensional parameterization in favor of sampling based approximations.
- It has strong connections with Wasserstein distributionally robust optimization (DRO).
- It suggests a parallelizable particle based algorithm.

Convergence Analysis

General Descent Lemma in $\mathbb{R}^d$

- A notion of good first-order approximation, for some $\alpha > 0$,

$$J(\mu) = J\left(\mu_0\right) + \nabla J\left(\mu_0\right)^T\left(\mu - \mu_0\right) + O\left(\left\|\mu - \mu_0\right\|^{1+\alpha}\right) \quad (7)$$

- When $\alpha = 1$, (7) reduce to the standard L-smooth condition.

How to make sense the general descent lemma in probability space?

General Descent Lemma in $\mathcal{P}_2(\mathbb{R}^d)$

$$J(\mu) = J(\mu_0) + \boxed{\langle DJ_{\mu_0}, \mu - \mu_0 \rangle} + O\left(\|\mu - \mu_0\|^{1+\alpha} \right)$$

- **Planer Geometry** (i.e., Gateaux derivative)
- Wasserstein Geometry

General Descent Lemma in $\mathcal{P}_2(\mathbb{R}^d)$

$$J(\mu) = J(\mu_0) + \langle DJ_{\mu_0}, \mu - \mu_0 \rangle + O\left(\boxed{W^{1+\alpha}(\mu, \mu_0)}\right)$$

- Planer Geometry (i.e., Gateaux derivative)
- Wasserstein Geometry

How to connect the planer and Wasserstein geometry?

General Descent Lemma in $\mathcal{P}_2(\mathbb{R}^d)$

$$J(\mu) = J(\mu_0) + \langle DJ_{\mu_0}, \mu - \mu_0 \rangle + O\left(\boxed{W^{1+\alpha}(\mu, \mu_0)}\right)$$

- Planer Geometry (i.e., Gateaux derivative)
- Wasserstein Geometry

How to connect the planer and Wasserstein geometry?

Wasserstein Differentiability

- A framework allows us to relate the planer geometry and Wasserstein geometry (Luigi, Gigli and Savaré, 2005). That is, the Wasserstein derivative is given by $F_\mu(\cdot)$ satisfying

$$\langle DJ_\mu, \nu - \mu \rangle = \int F_\mu(x)^T (y - x) \pi^*(dx, dy)$$

where π^* is the optimal coupling between μ and ν.

$$J(\mu) = J(\mu_0) + \int F_{\mu_0}(x)^T (y - x) \pi^*(dx, dy) + O\left(W^{1+\alpha}(\mu, \mu_0)\right)$$

Wasserstein Differentiability

- A framework allows us to relate the planer geometry and Wasserstein geometry (Luigi, Gigli and Savaré, 2005). That is, the Wasserstein derivative is given by $F_\mu(\cdot)$ satisfying

$$\langle DJ_\mu, \nu - \mu \rangle = \int F_\mu(x)^T (y - x) \pi^*(dx, dy)$$

where π^* is the optimal coupling between μ and ν.

$$J(\mu) = J(\mu_0) + \int F_{\mu_0}(x)^T (y - x) \pi^*(dx, dy) + O\left(W^{1+\alpha}(\mu, \mu_0)\right)$$

Technical Assumptions

- A1) Suppose that $J(\cdot)$ is α-Wasserstein smooth in the sense that for $\alpha \in (0, 1]$

$$J(\mu) = J(\mu_0) + \int F_{\mu_0}(x)^T (y - x)\pi^*(dx, dy) + O\left(W^{1+\alpha}(\mu, \mu_0)\right).$$

- A2) Assume that $DJ_\mu(\cdot)$ is L-smooth.

  Here, $F_{\mu_0} = \nabla DJ_{\mu_0}$ (i.e., connect Wasserstein derivative and Gateaux derivative).

- A3) Assume that $J(\cdot)$ satisfies a PŁ inequality of the form

$$\tau \cdot \left(J(\mu) - \inf_\mu J(\mu)\right)^\theta \leqslant \|\nabla DJ_\mu(x)\|_{L_2(\mu)}.$$

Technical Assumptions

- A1) Suppose that $J(\cdot)$ is α-Wasserstein smooth in the sense that for $\alpha \in (0, 1]$

$$J(\mu) = J(\mu_0) + \int F_{\mu_0}(x)^T (y - x)\pi^*(dx, dy) + O\left(W^{1+\alpha}(\mu, \mu_0)\right).$$

- A2) Assume that $DJ_\mu(\cdot)$ is L-smooth.

  Here, $F_{\mu_0} = \nabla DJ_{\mu_0}$ (i.e., connect Wasserstein derivative and Gateaux derivative).

- A3) Assume that $J(\cdot)$ satisfies a PŁ inequality of the form

$$\tau \cdot \left(J(\mu) - \inf_\mu J(\mu)\right)^\theta \leq \|\nabla DJ_\mu(x)\|_{L_2(\mu)}.$$

Technical Assumptions

- A1) Suppose that $J(\cdot)$ is α-Wasserstein smooth in the sense that for $\alpha \in (0, 1]$

$$J(\mu) = J(\mu_0) + \int F_{\mu_0}(x)^T (y - x)\pi^*(dx, dy) + O\left(W^{1+\alpha}(\mu, \mu_0)\right).$$

- A2) Assume that $DJ_\mu(\cdot)$ is L-smooth.

  Here, $F_{\mu_0} = \nabla DJ_{\mu_0}$ (i.e., connect Wasserstein derivative and Gateaux derivative).

- A3) Assume that $J(\cdot)$ satisfies a PŁ inequality of the form

$$\tau \cdot \left(J(\mu) - \inf_{\mu} J(\mu)\right)^{\theta} \leqslant \|\nabla DJ_\mu(x)\|_{L_2(\mu)}.$$

Performance Guarantee

Theorem

Under A1) to A3) choose at the I-th iterate, as long as

$$\left\| \nabla DJ_{\mu_{i-1}}(X) \right\|_{L_2(\mu_{i-1})} > \varepsilon^{\theta}$$

let

$$\delta_i = O\left(\min\left\{1/L, \left\| \nabla DJ_{\mu_{i-1}}(X) \right\|_{L_2(\mu_{i-1})}^{1/\alpha}\right\}\right).$$

Then at most $\tilde{O}\left(\varepsilon^{-((1+\alpha)\theta/\alpha-1)^{+}/\alpha}\right)$ *iterations result in* ε *error in value function with a sample complexity of order* $\tilde{O}\left(\varepsilon^{-2(1+\alpha)/\alpha}\right)$ *of the initial distribution* μ_0.

Performance Guarantee

Theorem

At most $\tilde{O}\left(\varepsilon^{-((1+\alpha)\theta/\alpha-1)^+/\alpha}\right)$ *iterations result in* ε *error in value function with a sample complexity of order* $\tilde{O}\left(\varepsilon^{-2(1+\alpha)/\alpha}\right)$ *of the initial distribution* μ_0.

If $J(\cdot)$ is strongly convex and smooth, this recovers $\tilde{O}(1)$ complexity. If $J(\cdot)$ is convex, this recovers $\tilde{O}\left(\varepsilon^{-1}\right)$ complexity. Both of which are canonical results in finite dimensions.

Numerical Results

Gaussian Deconvolution

- Observation $Y_i = X_i + Z_i$ where Z_i is Gaussian Noise and you want to recover the distribution μ (i.e., $X_i \sim \mu$).
- $J(\mu) = D_{\sigma^2}(\mu, \mu_N)$ where D_{σ^2} is so-called entropic regularization of the 2-Wasserstein distance and μ_N is the empirical measure of Y (Rigollet and Weed, 2019).

$$\inf_{\pi_X=\mu, \pi_Y=\mu_N} \frac{1}{2} \int \|x - y\|_2^2 \pi(dx, dy) + \sigma^2 D_{KL}\left(\pi \| \mu \times \mu_N\right). \quad (8)$$

Gaussian Deconvolution — 2D Case

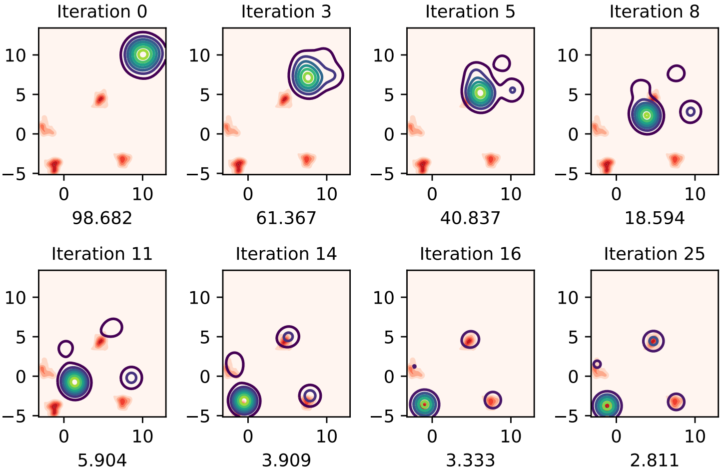

Figure: Gaussian Deconvolution 2D

Gaussian Deconvolution — High Dimension

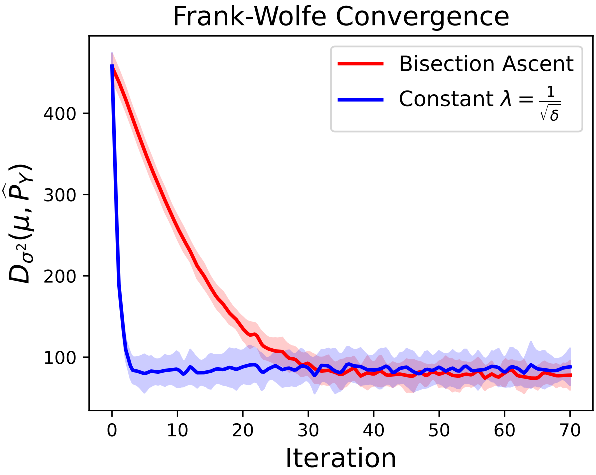

Figure: High-dimensional Gaussian deconvolution for $d = 64$.

Maxmimum Mean Discrepancy

- Let H be a reproducing Kernel Hilbert space and define

$$J(\mu) = \sup_{\|f\|_H \leqslant 1} E_\mu f(X) - E_{\mu_n} f(X). \tag{9}$$

- Student-Teacher neural network to parameterize f.
- We compare our method with MMD flow and Kernel Sobolev descent in term of gradient evaluations and same sample size.

Maxmimum Mean Discrepancy

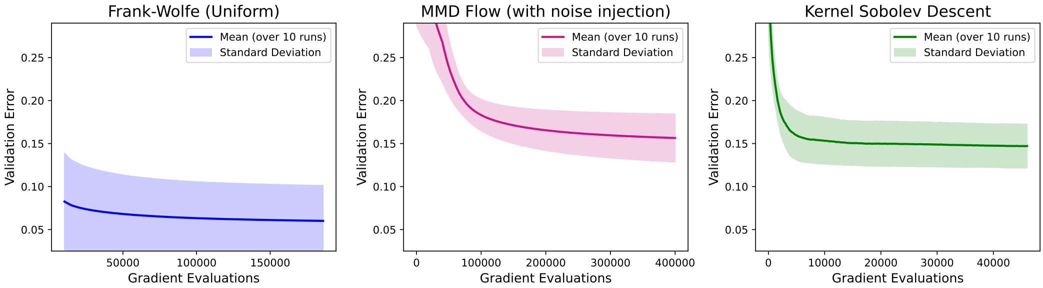

Figure: Student-Teacher Network; The left one is the result for our Frank-Wolfe method with the uniform strategy $\lambda = \frac{0.05}{\delta}$ and the step size δ is 0.5. The number of particle is 200.

Take Home Message

- Optimization over probabilities is a powerful concept connecting to many areas, including deep learning, variational inference, deconvolution, etc.
- We presented a modified Frank-Wolfe method which uses both planar geometry (i.e., computation) and Wasserstein geometry (i.e., convergence analysis).
- We obtain results of independent interest for solving worst-case expectations for Wasserstein DRO.

Take Home Message

- Optimization over probabilities is a powerful concept connecting to many areas, including deep learning, variational inference, deconvolution, etc.
- We presented a modified Frank-Wolfe method which uses both planar geometry (i.e., computation) and Wasserstein geometry (i.e., convergence analysis).
- We obtain results of independent interest for solving worst-case expectations for Wasserstein DRO.

Take Home Message

- Optimization over probabilities is a powerful concept connecting to many areas, including deep learning, variational inference, deconvolution, etc.
- We presented a modified Frank-Wolfe method which uses both planar geometry (i.e., computation) and Wasserstein geometry (i.e., convergence analysis).
- We obtain results of independent interest for solving worst-case expectations for Wasserstein DRO.

Reference

- Carson Kent, Jiajin Li, Jose Blanchet and Peter Glynn. Modified Frank Wolfe in Probability Space. **NeurIPS 2021**.

Thank you for listening! Q&A?

Jiajin Li

jiajinli@stanford.edu

https://gerrili1996.github.io/