A Primal Dual Smoothing Framework for Max-Structured Nonconvex Optimization

Renbo Zhao

Operations Research Center, Massachusetts Institute of Technology

Full video at http://web.mit.edu/renboz/www/talks.html

INFORMS Annual Meeting
Nov 2020

Problem Statement

Consider the following nonconvex nonsmooth optimization problem:

$$q^* \triangleq \min_{x \in \mathcal{X} \subseteq \mathbb{X}} \left\{ q(x) \triangleq f(x) + r(x) \right\}, \quad f(x) \triangleq \max_{y \in \mathcal{Y} \subseteq \mathbb{Y}} \Phi(x, y) - g(y), \qquad \texttt{(P)}$$

Problem Statement

Consider the following nonconvex nonsmooth optimization problem:

$$q^* \triangleq \min_{x\in\mathcal{X}\subseteq\mathbb{X}} \left\{q(x) \triangleq f(x) + r(x)\right\}, \quad f(x) \triangleq \max_{y\in\mathcal{Y}\subseteq\mathbb{Y}} \Phi(x,y) - g(y), \qquad \texttt{(P)}$$

- $\mathbb{X}$ and $\mathbb{Y}$ are finite-dimensional real normed spaces.

Problem Statement

Consider the following nonconvex nonsmooth optimization problem:

$$q^* \triangleq \min_{x \in \mathcal{X} \subseteq \mathbb{X}} \left\{ q(x) \triangleq f(x) + r(x) \right\}, \quad f(x) \triangleq \max_{y \in \mathcal{Y} \subseteq \mathbb{Y}} \Phi(x,y) - g(y), \qquad \texttt{(P)}$$

- $\mathbb{X}$ and $\mathbb{Y}$ are finite-dimensional real normed spaces.
- $\mathcal{X}$ and $\mathcal{Y}$ are nonempty, closed and convex sets, and $\mathcal{Y}$ is bounded.

Problem Statement

Consider the following nonconvex nonsmooth optimization problem:

$$q^* \triangleq \min_{x\in\mathcal{X}\subseteq\mathbb{X}} \left\{q(x) \triangleq f(x) + r(x)\right\}, \quad f(x) \triangleq \max_{y\in\mathcal{Y}\subseteq\mathbb{Y}} \Phi(x,y) - g(y), \qquad \text{(P)}$$

- $\mathbb{X}$ and $\mathbb{Y}$ are finite-dimensional real normed spaces.
- $\mathcal{X}$ and $\mathcal{Y}$ are nonempty, closed and convex sets, and $\mathcal{Y}$ is bounded.
- q is bounded below, i.e., $q^* > -\infty$.

Problem Statement

Consider the following nonconvex nonsmooth optimization problem:

$$q^* \triangleq \min_{x\in\mathcal{X}\subseteq\mathbb{X}} \left\{q(x) \triangleq f(x) + r(x)\right\}, \quad f(x) \triangleq \max_{y\in\mathcal{Y}\subseteq\mathbb{Y}} \Phi(x,y) - g(y), \qquad \texttt{(P)}$$

- ▷ $\mathbb{X}$ and $\mathbb{Y}$ are finite-dimensional real normed spaces.
- ▷ $\mathcal{X}$ and $\mathcal{Y}$ are nonempty, closed and convex sets, and $\mathcal{Y}$ is bounded.
- ▷ q is bounded below, i.e., $q^* > -\infty$.
- ▷ $r : \mathbb{X} \to \overline{\mathbb{R}} \triangleq \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{Y} \to \overline{\mathbb{R}}$ are closed, convex and proper.

Problem Statement

Consider the following nonconvex nonsmooth optimization problem:

$$q^* \triangleq \min_{x \in \mathcal{X} \subseteq \mathbb{X}} \big\{ q(x) \triangleq f(x) + r(x) \big\}, \quad f(x) \triangleq \max_{y \in \mathcal{Y} \subseteq \mathbb{Y}} \Phi(x, y) - g(y), \qquad \texttt{(P)}$$

- ▷ $\mathbb{X}$ and $\mathbb{Y}$ are finite-dimensional real normed spaces.
- ▷ $\mathcal{X}$ and $\mathcal{Y}$ are nonempty, closed and convex sets, and $\mathcal{Y}$ is bounded.
- ▷ q is bounded below, i.e., $q^* > -\infty$.
- ▷ $r : \mathbb{X} \to \overline{\mathbb{R}} \triangleq \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{Y} \to \overline{\mathbb{R}}$ are closed, convex and proper.
- ▷ r and g are M_r- and M_g-Lipschitz on $\mathcal{X}$ and $\mathcal{Y}$, respectively, with easily computable Bregman proximal projections.

Assumptions on Φ

The function $\Phi : \mathbb{X} \times \mathbb{Y} \to [-\infty, +\infty]$ satisfies the following assumptions.

Assumptions on Φ

The function $\Phi : \mathbb{X} \times \mathbb{Y} \to [-\infty, +\infty]$ satisfies the following assumptions.

- For any $x \in \mathcal{X}$, $\Phi(x, \cdot)$ is concave on $\mathcal{Y}$.

Assumptions on Φ

The function $\Phi : \mathbb{X} \times \mathbb{Y} \to [-\infty, +\infty]$ satisfies the following assumptions.

- For any $x \in \mathcal{X}$, $\Phi(x, \cdot)$ is concave on $\mathcal{Y}$.
- For any $y \in \mathcal{Y}$, $\Phi(\cdot, y)$ is γ-weakly convex on $\mathcal{X}$ for some $\gamma \in (0, L_{xx}]$:

$$-(\gamma/2)\left\|x' - x\right\|^2 \leq \Phi(x', y) - \Phi(x, y) - \langle \nabla_x \Phi(x, y), x' - x\rangle, \quad \forall\, x, x' \in \mathcal{X}.$$

Assumptions on Φ

The function $\Phi : \mathbb{X} \times \mathbb{Y} \to [-\infty, +\infty]$ satisfies the following assumptions.

- For any $x \in \mathcal{X}$, $\Phi(x, \cdot)$ is concave on $\mathcal{Y}$.
- For any $y \in \mathcal{Y}$, $\Phi(\cdot, y)$ is γ-weakly convex on $\mathcal{X}$ for some $\gamma \in (0, L_{xx}]$:

  $$-(\gamma/2) \|x' - x\|^2 \leq \Phi(x', y) - \Phi(x, y) - \langle \nabla_x \Phi(x, y), x' - x \rangle, \quad \forall\, x, x' \in \mathcal{X}.$$

- $\Phi(\cdot, \cdot)$ is jointly continuous on $\mathcal{X} \times \mathcal{Y}$.

Assumptions on Φ

The function $\Phi : \mathbb{X} \times \mathbb{Y} \to [-\infty, +\infty]$ satisfies the following assumptions.

- ▷ For any $x \in \mathcal{X}$, $\Phi(x, \cdot)$ is concave on $\mathcal{Y}$.
- ▷ For any $y \in \mathcal{Y}$, $\Phi(\cdot, y)$ is γ-weakly convex on $\mathcal{X}$ for some $\gamma \in (0, L_{xx}]$:

$$-(\gamma/2)\left\|x' - x\right\|^2 \le \Phi(x', y) - \Phi(x, y) - \langle \nabla_x \Phi(x, y), x' - x \rangle, \quad \forall\, x, x' \in \mathcal{X}.$$

- ▷ $\Phi(\cdot, \cdot)$ is jointly continuous on $\mathcal{X} \times \mathcal{Y}$.
- ▷ $\Phi(\cdot, \cdot)$ is differentiable on $\mathcal{X} \times \mathcal{Y}$, and for any $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$:

$$\begin{aligned}
\|\nabla_x \Phi(x, y) - \nabla_x \Phi(x', y)\|_* &\le L_{xx}\|x - x'\|, \\
\|\nabla_x \Phi(x, y) - \nabla_x \Phi(x, y')\|_* &\le L_{xy}\|y - y'\|, \\
\|\nabla_y \Phi(x, y) - \nabla_y \Phi(x', y)\|_* &\le L_{xy}\|x - x'\|, \\
\|\nabla_y \Phi(x, y) - \nabla_y \Phi(x, y')\|_* &\le L_{yy}\|y - y'\|.
\end{aligned}$$

Application: Distributionally Robust Optimization

$$\min_{x\in\mathcal{X}} \max_{p\in\mathcal{P}} \mathbb{E}_{\xi\sim p}[\ell(x,\xi)] + r(x), \quad \mathbb{E}_{\xi\sim p}[\ell(x,\xi)] = \sum_{i=1}^n p_i \ell(x,\xi_i).$$

Application: Distributionally Robust Optimization

$$\min_{x\in\mathcal{X}} \max_{p\in\mathcal{P}} \mathbb{E}_{\xi\sim p}[\ell(x,\xi)] + r(x), \quad \mathbb{E}_{\xi\sim p}[\ell(x,\xi)] = \sum_{i=1}^{n} p_i \ell(x,\xi_i).$$

- Let $(\Xi, \mathcal{B}, \bar{p})$ be a probability space, where $\Xi \triangleq \{\xi_1, \ldots, \xi_n\}$.

Application: Distributionally Robust Optimization

$$\min_{x \in \mathcal{X}} \max_{p \in \mathcal{P}} \mathbb{E}_{\xi \sim p}[\ell(x, \xi)] + r(x), \quad \mathbb{E}_{\xi \sim p}[\ell(x, \xi)] = \sum_{i=1}^{n} p_i \ell(x, \xi_i).$$

- Let $(\Xi, \mathcal{B}, \bar{p})$ be a probability space, where $\Xi \triangleq \{\xi_1, \ldots, \xi_n\}$.
- Let $\ell : \mathbb{X} \times \Xi \to \mathbb{R}$ be a loss function such that $\ell(x, \xi)$ returns the loss of decision $x \in \mathcal{X}$ given the (random) parameter $\xi \in \Xi$.

Application: Distributionally Robust Optimization

$$\min_{x\in\mathcal{X}} \max_{p\in\mathcal{P}} \mathbb{E}_{\xi\sim p}[\ell(x,\xi)] + r(x), \quad \mathbb{E}_{\xi\sim p}[\ell(x,\xi)] = \textstyle\sum_{i=1}^n p_i \ell(x,\xi_i).$$

- ▷ Let $(\Xi, \mathcal{B}, \bar{p})$ be a probability space, where $\Xi \triangleq \{\xi_1, \ldots, \xi_n\}$.
- ▷ Let $\ell : \mathbb{X} \times \Xi \to \mathbb{R}$ be a loss function such that $\ell(x,\xi)$ returns the loss of decision $x \in \mathcal{X}$ given the (random) parameter $\xi \in \Xi$.
- ▷ Let $\ell(\cdot,\xi)$ be $L(\xi)$-smooth on $\mathcal{X}$, i.e., $\ell(\cdot,\xi)$ is differentiable with $L(\xi)$-Lipschitz gradient on $\mathcal{X}$.

Application: Distributionally Robust Optimization

$$\min_{x\in\mathcal{X}} \max_{p\in\mathcal{P}} \mathbb{E}_{\xi\sim p}[\ell(x,\xi)] + r(x), \quad \mathbb{E}_{\xi\sim p}[\ell(x,\xi)] = \sum_{i=1}^n p_i \ell(x,\xi_i).$$

- Let $(\Xi, \mathcal{B}, \bar{p})$ be a probability space, where $\Xi \triangleq \{\xi_1, \ldots, \xi_n\}$.
- Let $\ell : \mathbb{X} \times \Xi \to \mathbb{R}$ be a loss function such that $\ell(x,\xi)$ returns the loss of decision $x \in \mathcal{X}$ given the (random) parameter $\xi \in \Xi$.
- Let $\ell(\cdot,\xi)$ be $L(\xi)$-smooth on $\mathcal{X}$, i.e., $\ell(\cdot,\xi)$ is differentiable with $L(\xi)$-Lipschitz gradient on $\mathcal{X}$.
- Let $\mathcal{P}$ denotes the uncertainty set that contains $\bar{p}$ as a nominal distribution, e.g., $\mathcal{P} \triangleq \{p \in \Delta_n : d_{\mathrm{TV}}(p,\bar{p}) \le \alpha_{\mathcal{X}}\}$.

Application: Distributionally Robust Optimization

$$\min_{x\in\mathcal{X}}\max_{p\in\mathcal{P}} \mathbb{E}_{\xi\sim p}[\ell(x,\xi)] + r(x), \quad \mathbb{E}_{\xi\sim p}[\ell(x,\xi)] = \textstyle\sum_{i=1}^n p_i \ell(x,\xi_i).$$

- Let $(\Xi, \mathcal{B}, \bar{p})$ be a probability space, where $\Xi \triangleq \{\xi_1, \ldots, \xi_n\}$.
- Let $\ell : \mathbb{X} \times \Xi \to \mathbb{R}$ be a loss function such that $\ell(x,\xi)$ returns the loss of decision $x \in \mathcal{X}$ given the (random) parameter $\xi \in \Xi$.
- Let $\ell(\cdot,\xi)$ be $L(\xi)$-smooth on $\mathcal{X}$, i.e., $\ell(\cdot,\xi)$ is differentiable with $L(\xi)$-Lipschitz gradient on $\mathcal{X}$.
- Let $\mathcal{P}$ denotes the uncertainty set that contains $\bar{p}$ as a nominal distribution, e.g., $\mathcal{P} \triangleq \{p \in \Delta_n : d_{\mathrm{TV}}(p,\bar{p}) \le \alpha_{\mathcal{X}}\}$.
- $r : \mathbb{X} \to \overline{\mathbb{R}}$ is a regularizer, e.g., $\|\cdot\|_1$.

Other applications

Other applications

- Generative adversarial training with "simple" discriminator

Other applications

- Generative adversarial training with "simple" discriminator
- Dual problem of composite optimization

Other applications

- Generative adversarial training with "simple" discriminator
- Dual problem of composite optimization
- Minimizing the largest eigenvalue of factorized matrices

Non-Euclidean Geometry

Non-Euclidean Geometry

- Let $\mathcal{U} \subseteq \mathbb{U}$ be nonempty, closed and convex, where $\mathbb{U}$ is a finite-dimensional real normed space.

Non-Euclidean Geometry

- Let $\mathcal{U} \subseteq \mathbb{U}$ be nonempty, closed and convex, where $\mathbb{U}$ is a finite-dimensional real normed space.
- We call $h_{\mathcal{U}}$ a *distance generating function* (DGF) on $\mathcal{U}$ if

Non-Euclidean Geometry

- Let $\mathcal{U} \subseteq \mathbb{U}$ be nonempty, closed and convex, where $\mathbb{U}$ is a finite-dimensional real normed space.
- We call $h_{\mathcal{U}}$ a *distance generating function* (DGF) on $\mathcal{U}$ if
  - it is essentially smooth, i.e., cont. differentiable on $\mathsf{int}\,\mathsf{dom}\, h_{\mathcal{U}} \neq \emptyset$, and for any $u_k \to u \in \mathsf{bd}\,\mathcal{U}$, $\|\nabla h_{\mathcal{U}}(u_k)\|_* \to +\infty$,

Non-Euclidean Geometry

- Let $\mathcal{U} \subseteq \mathbb{U}$ be nonempty, closed and convex, where $\mathbb{U}$ is a finite-dimensional real normed space.
- We call $h_{\mathcal{U}}$ a *distance generating function* (DGF) on $\mathcal{U}$ if
  - it is essentially smooth, i.e., cont. differentiable on $\mathsf{int\,dom}\, h_{\mathcal{U}} \neq \emptyset$, and for any $u_k \to u \in \mathsf{bd}\,\mathcal{U}$, $\|\nabla h_{\mathcal{U}}(u_k)\|_* \to +\infty$,
  - it is continuous and 1-s.c. on $\mathcal{U}$,

Non-Euclidean Geometry

- Let $\mathcal{U} \subseteq \mathbb{U}$ be nonempty, closed and convex, where $\mathbb{U}$ is a finite-dimensional real normed space.

- We call $h_{\mathcal{U}}$ a *distance generating function* (DGF) on $\mathcal{U}$ if
  - it is essentially smooth, i.e., cont. differentiable on $\mathsf{int\,dom}\, h_{\mathcal{U}} \neq \emptyset$, and for any $u_k \to u \in \mathsf{bd}\,\mathcal{U}$, $\|\nabla h_{\mathcal{U}}(u_k)\|_* \to +\infty$,
  - it is continuous and 1-s.c. on $\mathcal{U}$,
  - it generates the *Bregman distance*
  $$D_{h_{\mathcal{U}}}(u, u') \triangleq h_{\mathcal{U}}(u) - h_{\mathcal{U}}(u') - \langle \nabla h_{\mathcal{U}}(u'), u - u' \rangle$$
  that satisfies $D_{h_{\mathcal{U}}}(u, u') \geq (1/2)\|u - u'\|^2$.

Non-Euclidean Geometry

- Let $\mathcal{U} \subseteq \mathbb{U}$ be nonempty, closed and convex, where $\mathbb{U}$ is a finite-dimensional real normed space.

- We call $h_{\mathcal{U}}$ a *distance generating function* (DGF) on $\mathcal{U}$ if
  - it is essentially smooth, i.e., cont. differentiable on $\mathsf{int\,dom}\, h_{\mathcal{U}} \neq \emptyset$, and for any $u_k \to u \in \mathsf{bd}\,\mathcal{U}$, $\|\nabla h_{\mathcal{U}}(u_k)\|_* \to +\infty$,
  - it is continuous and 1-s.c. on $\mathcal{U}$,
  - it generates the *Bregman distance*

$$D_{h_{\mathcal{U}}}(u, u') \triangleq h_{\mathcal{U}}(u) - h_{\mathcal{U}}(u') - \langle \nabla h_{\mathcal{U}}(u'), u - u' \rangle$$

  that satisfies $D_{h_{\mathcal{U}}}(u, u') \geq (1/2) \|u - u'\|^2$.

- Example: $\mathbb{U} = (\mathbb{R}^n, \|\cdot\|_1)$, $\mathcal{U} = \Delta_n \triangleq \{u \in \mathbb{R}^n_+ : \sum_{i=1}^n u_i = 1\}$, $h_{\mathcal{U}} = \sum_{i=1}^n u_i \log u_i$, $D_{h_{\mathcal{U}}}(u, u') \geq (1/2) \|u - u'\|_1^2$.

Bregman Proximal Projection (BPP)

Let $u' \in \mathcal{U}^o$, $u^* \in \mathbb{U}^*$ and $\varphi : \mathbb{U} \to \overline{\mathbb{R}}$ be CCP.

Bregman Proximal Projection (BPP)

Let $u' \in \mathcal{U}^o$, $u^* \in \mathbb{U}^*$ and $\varphi : \mathbb{U} \to \overline{\mathbb{R}}$ be CCP.

$$\boxed{u' \mapsto u^+ \triangleq \arg\min_{u \in \mathcal{U}} \;\; \varphi(u) + \langle u^*, u \rangle + \lambda^{-1} D_{h_{\mathcal{U}}}(u, u')} \qquad \texttt{(BPP)}$$

Bregman Proximal Projection (BPP)

Let $u' \in \mathcal{U}^o$, $u^* \in \mathbb{U}^*$ and $\varphi : \mathbb{U} \to \overline{\mathbb{R}}$ be CCP.

$$\boxed{u' \mapsto u^+ \triangleq \arg\min_{u \in \mathcal{U}} \;\; \varphi(u) + \langle u^*, u \rangle + \lambda^{-1} D_{h_{\mathcal{U}}}(u, u')} \qquad \texttt{(BPP)}$$

▷ We say φ has an easily computable proximal operator if there exists a DGF $h_{\mathcal{U}}$ on $\mathcal{U}$ such that (BPP) has a (unique) *easily computable* solution.

Bregman Proximal Projection (BPP)

Let $u' \in \mathcal{U}^o$, $u^* \in \mathbb{U}^*$ and $\varphi : \mathbb{U} \to \overline{\mathbb{R}}$ be CCP.

$$\boxed{u' \mapsto u^+ \triangleq \arg\min_{u \in \mathcal{U}} \;\varphi(u) + \langle u^*, u\rangle + \lambda^{-1} D_{h_{\mathcal{U}}}(u, u')} \tag{BPP}$$

- We say φ has an easily computable proximal operator if there exists a DGF $h_{\mathcal{U}}$ on $\mathcal{U}$ such that (BPP) has a (unique) *easily computable* solution.
- If $\mathbb{U}$ is a Hilbert space, then (BPP) becomes

$$u' \mapsto u^+ \triangleq \mathsf{prox}_{\lambda\varphi}(u' - \lambda u^*).$$

Near-stationary point

Near-stationary point

- Let $\omega_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$ be a DGF on $\mathcal{X}$. Let ω be twice differentiable on $\mathcal{X}'$ and $\beta_{\mathcal{X}}$-smooth on $\mathcal{X}$, i.e., $\sup_{x \in \mathcal{X}} \|\nabla^2 \omega_{\mathcal{X}}(x)\| \leq \beta_{\mathcal{X}}$.

Near-stationary point

- Let $\omega_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$ be a DGF on $\mathcal{X}$. Let ω be twice differentiable on $\mathcal{X}'$ and $\beta_{\mathcal{X}}$-smooth on $\mathcal{X}$, i.e., $\sup_{x \in \mathcal{X}} \|\nabla^2 \omega_{\mathcal{X}}(x)\| \leq \beta_{\mathcal{X}}$.
- $x \in \mathcal{X}$ an ε-near-stationary point of (P) if for any $\lambda > 0$,

$$\|x - \mathsf{prox}(q, x, \lambda)\| \leq \varepsilon\lambda/\beta_{\mathcal{X}},$$

$$\mathsf{prox}(q, x, \lambda) \triangleq \arg\min_{x' \in \mathcal{X}} q(x') + \lambda^{-1} D_{\omega_{\mathcal{X}}}(x', x).$$

Near-stationary point

- Let $\omega_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$ be a DGF on $\mathcal{X}$. Let ω be twice differentiable on $\mathcal{X}'$ and $\beta_{\mathcal{X}}$-smooth on $\mathcal{X}$, i.e., $\sup_{x\in\mathcal{X}} \|\nabla^2\omega_{\mathcal{X}}(x)\| \le \beta_{\mathcal{X}}$.

- $x \in \mathcal{X}$ an ε-near-stationary point of (P) if for any $\lambda > 0$,

$$\|x - \mathsf{prox}(q, x, \lambda)\| \le \varepsilon\lambda/\beta_{\mathcal{X}},$$
$$\mathsf{prox}(q, x, \lambda) \triangleq \arg\min_{x'\in\mathcal{X}} q(x') + \lambda^{-1} D_{\omega_{\mathcal{X}}}(x', x).$$

- Note that $\|x - \mathsf{prox}(q, x, \lambda)\| \le \varepsilon\lambda/\beta_{\mathcal{X}} \Rightarrow \mathsf{dist}\big(0, \partial q\big(\mathsf{prox}(q, x, \lambda)\big)\big) \le \varepsilon$. In other words, $\mathsf{prox}(q, x, \lambda)$ is an approximate stationary point of (P), and x is $O(\varepsilon)$-close to $\mathsf{prox}(q, x, \lambda)$.

Near-stationary point

- Let $\omega_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$ be a DGF on $\mathcal{X}$. Let ω be twice differentiable on $\mathcal{X}'$ and $\beta_{\mathcal{X}}$-smooth on $\mathcal{X}$, i.e., $\sup_{x\in\mathcal{X}} \|\nabla^2\omega_{\mathcal{X}}(x)\| \le \beta_{\mathcal{X}}$.

- $x \in \mathcal{X}$ an ε-near-stationary point of (P) if for any $\lambda > 0$,

$$\|x - \mathsf{prox}(q, x, \lambda)\| \le \varepsilon\lambda/\beta_{\mathcal{X}},$$
$$\mathsf{prox}(q, x, \lambda) \triangleq \arg\min_{x'\in\mathcal{X}} q(x') + \lambda^{-1} D_{\omega_{\mathcal{X}}}(x', x).$$

- Note that $\|x - \mathsf{prox}(q, x, \lambda)\| \le \varepsilon\lambda/\beta_{\mathcal{X}} \Rightarrow \mathsf{dist}\left(0, \partial q\big(\mathsf{prox}(q, x, \lambda)\big)\right) \le \varepsilon$. In other words, $\mathsf{prox}(q, x, \lambda)$ is an approximate stationary point of (P), and x is $O(\varepsilon)$-close to $\mathsf{prox}(q, x, \lambda)$.

- We refer to solving (P) as finding an ε-near-stationary point of (P).

First-Order Oracles

First-Order Oracles

- There exist a primal first-order oracle $\mathscr{O}^{\mathrm{P}}$ and a dual first-order oracle $\mathscr{O}^{\mathrm{D}}$ that take in any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and returns $\nabla_x \Phi(x, y)$ and $\nabla_y \Phi(x, y)$, respectively.

First-Order Oracles

- There exist a primal first-order oracle $\mathscr{O}^{\mathrm{P}}$ and a dual first-order oracle $\mathscr{O}^{\mathrm{D}}$ that take in any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and returns $\nabla_x \Phi(x, y)$ and $\nabla_y \Phi(x, y)$, respectively.

- We use the *primal* and *dual* oracle complexities required by a certain algorithm to obtain an ε-near-stationary point to measure its performance.

Main Contribution

Main Contribution

- Propose a primal dual smoothing framework for solving (P) that unifies two approaches, i.e., dual-then-primal and primal-then-dual smoothing.

Main Contribution

- ▷ Propose a primal dual smoothing framework for solving (P) that unifies two approaches, i.e., dual-then-primal and primal-then-dual smoothing.
  - It solves (P) in its full generality, and improves the best-known complexity (Theku. et al., 2019) even in the restricted setting, i.e., $f \equiv 0$, $r \equiv 0$ and both $\mathbb{X}$ and $\mathbb{Y}$ are Euclidean .

Main Contribution

- Propose a primal dual smoothing framework for solving (P) that unifies two approaches, i.e., dual-then-primal and primal-then-dual smoothing.
  - It solves (P) in its full generality, and improves the best-known complexity (Theku. et al., 2019) even in the restricted setting, i.e., $f \equiv 0$, $r \equiv 0$ and both $\mathbb{X}$ and $\mathbb{Y}$ are Euclidean .
- As the cornerstone of our framework, we propose an efficient method for solving a class of convex-concave saddle-point problems with primal strong convexity, with significantly improved dual complexity.

Main Contribution

- Propose a primal dual smoothing framework for solving (P) that unifies two approaches, i.e., dual-then-primal and primal-then-dual smoothing.
  - It solves (P) in its full generality, and improves the best-known complexity (Theku. et al., 2019) even in the restricted setting, i.e., $f \equiv 0$, $r \equiv 0$ and both $\mathbb{X}$ and $\mathbb{Y}$ are Euclidean .

- As the cornerstone of our framework, we propose an efficient method for solving a class of convex-concave saddle-point problems with primal strong convexity, with significantly improved dual complexity.
  - In this method, we develop the first *non-Euclidean inexact* accelerated proximal gradient (APG) method for strongly convex composite optimization.

Comparison with Theku. et al. (2019)

$f \equiv 0$, $r \equiv 0$ and both $\mathbb{X}$ and $\mathbb{Y}$ are Euclidean

Algorithms	Primal Oracle Comp.
Theku. et al.	$O\Big((L_{xx} + L_{xy} + L_{yy})^2 \varepsilon^{-3} \log^2(\varepsilon^{-1})\Big)$
Our method	$O\Big(\sqrt{\gamma(L_{xx} + \gamma)}\Big(\sqrt{L_{yy}\gamma} + L_{xy}\Big)\varepsilon^{-3} \log^2(\varepsilon^{-1})\Big)$

Algorithms	Dual Oracle Comp.
Theku. et al.	$O\Big((L_{xx} + L_{xy} + L_{yy})^2 \varepsilon^{-3} \log^2(\varepsilon^{-1})\Big)$
Our method	$O\Big(\gamma\Big(\sqrt{L_{yy}\gamma} + L_{xy}\Big)\varepsilon^{-3} \log(\varepsilon^{-1})\Big)$

Fréchet sub-differential and derivative

Fréchet sub-differential and derivative

- Define the Fréchet subdifferential of f at $x \in \mathsf{dom}\, f$, denoted by $\partial f(x)$, as

$$\partial f(x) \triangleq \left\{ x^* \in \mathbb{X}^* : \liminf_{h \to 0} \frac{f(x+h) - f(x) - \langle x^*, h \rangle}{\|h\|} \geq 0 \right\}.$$

In other words, $x^* \in \partial f(x) \Leftrightarrow f(x+h) \geq f(x) + \langle x^*, h \rangle + o(\|h\|)$.

Fréchet sub-differential and derivative

- Define the Fréchet subdifferential of f at $x \in \mathsf{dom}\, f$, denoted by $\partial f(x)$, as

$$\partial f(x) \triangleq \left\{ x^* \in \mathbb{X}^* : \liminf_{h \to 0} \frac{f(x+h) - f(x) - \langle x^*, h \rangle}{\|h\|} \geq 0 \right\}.$$

  In other words, $x^* \in \partial f(x) \Leftrightarrow f(x+h) \geq f(x) + \langle x^*, h \rangle + o(\|h\|)$.

- When f is convex, ∂f becomes the convex sub-differential.

Fréchet sub-differential and derivative

- Define the Fréchet subdifferential of f at $x \in \mathsf{dom}\, f$, denoted by $\partial f(x)$, as

$$\partial f(x) \triangleq \left\{ x^* \in \mathbb{X}^* : \liminf_{h \to 0} \frac{f(x+h) - f(x) - \langle x^*, h \rangle}{\|h\|} \geq 0 \right\}.$$

  In other words, $x^* \in \partial f(x) \Leftrightarrow f(x+h) \geq f(x) + \langle x^*, h \rangle + o(\|h\|)$.

- When f is convex, ∂f becomes the convex sub-differential.
- Define the Fréchet derivative of f (or simply, gradient) at x, denoted by $\nabla f(x)$, as the unique element in $\mathbb{X}^*$ that satisfies

$$\lim_{h \to 0} \frac{f(x+h) - f(x) - \langle \nabla f(x), h \rangle}{\|h\|} = 0.$$

  In other words, $f(x+h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|)$.

Smoothing

Smoothing

Define the *dually smoothed* f, with dual smoothing parameter $\rho > 0$, as

$$f_\rho(x) = \max_{y \in \mathcal{Y}} \left[\phi_\rho^{\mathrm{D}}(x,y) \triangleq \Phi(x,y) - g(y) - \rho\omega_{\mathcal{Y}}(y) \right], \tag{DS}$$

where $\omega_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ is the DGF on $\mathcal{Y}$.

Lemma 1

- $\nabla f_\rho(x) = \nabla_x \Phi(x, y_\rho^*(x))$.
- ∇f_ρ *is* L_ρ*-Lipschitz on* $\mathcal{X}$*, where* $L_\rho \triangleq L_{xx} + L_{xy}^2/\rho$.

Lemma 2

Both of the functions f *and* f_ρ *are* γ*-weakly convex on* $\mathcal{X}$*.*

Primal Dual Smoothing Framework

For any $\rho, \lambda > 0$, $x' \in \mathcal{X}$ and $x \in \mathcal{X}^o$, we define

$$\begin{aligned}
Q^\lambda(x'; x) &\triangleq q(x') + \lambda^{-1} D_{\omega_\mathcal{X}}(x'; x), \\
q^\lambda(x) &\triangleq \inf_{x' \in \mathcal{X}} Q^\lambda(x'; x), && (\lambda\text{-Moreau env. of } q) \\
\mathsf{prox}(q, x, \lambda) &\triangleq \arg\min_{x' \in \mathcal{X}} Q^\lambda(x'; x),
\end{aligned}$$

$$\begin{aligned}
q_\rho(x) &\triangleq f_\rho(x) + r(x), && (\rho\text{-dually smoothed } q) \\
Q^\lambda_\rho(x'; x) &\triangleq q_\rho(x') + \lambda^{-1} D_{\omega_\mathcal{X}}(x'; x), \\
q^\lambda_\rho(x) &\triangleq \inf_{x' \in \mathcal{X}} Q^\lambda_\rho(x'; x), && (\lambda\text{-Moreau env. of } q_\rho) \\
\mathsf{prox}(q_\rho, x, \lambda) &\triangleq \arg\min_{x' \in \mathcal{X}} Q^\lambda_\rho(x'; x).
\end{aligned}$$

Primal Dual Smoothing Framework

For any $\rho, \lambda > 0$, $x' \in \mathcal{X}$ and $x \in \mathcal{X}^o$, we define

$$\begin{aligned} Q^\lambda(x';x) &\triangleq q(x') + \lambda^{-1} D_{\omega_\mathcal{X}}(x';x), \\ q^\lambda(x) &\triangleq \inf_{x' \in \mathcal{X}} Q^\lambda(x';x), && (\lambda\text{-Moreau env. of } q) \\ \mathsf{prox}(q,x,\lambda) &\triangleq \arg\min_{x' \in \mathcal{X}} Q^\lambda(x';x), \end{aligned}$$

$$\begin{aligned} q_\rho(x) &\triangleq f_\rho(x) + r(x), && (\rho\text{-dually smoothed } q) \\ Q^\lambda_\rho(x';x) &\triangleq q_\rho(x') + \lambda^{-1} D_{\omega_\mathcal{X}}(x';x), \\ q^\lambda_\rho(x) &\triangleq \inf_{x' \in \mathcal{X}} Q^\lambda_\rho(x';x), && (\lambda\text{-Moreau env. of } q_\rho) \\ \mathsf{prox}(q_\rho,x,\lambda) &\triangleq \arg\min_{x' \in \mathcal{X}} Q^\lambda_\rho(x';x). \end{aligned}$$

Primal Dual Smoothing Framework

Primal Dual Smoothing Framework

- **Input**: Accuracy parameter $\eta > 0$, smoothing parameters $\lambda = 1/(2\gamma)$, $\rho = \eta/(4\Omega_{\mathcal{Y}}(\omega_{\mathcal{Y}}))$

Primal Dual Smoothing Framework

- **Input**: Accuracy parameter $\eta > 0$, smoothing parameters $\lambda = 1/(2\gamma)$, $\rho = \eta/(4\Omega_{\mathcal{Y}}(\omega_{\mathcal{Y}}))$
- **Init**: $k = 0$, $x_1 \in \mathcal{X}^o$

Primal Dual Smoothing Framework

- **Input**: Accuracy parameter $\eta > 0$, smoothing parameters $\lambda = 1/(2\gamma)$, $\rho = \eta/(4\Omega_{\mathcal{Y}}(\omega_{\mathcal{Y}}))$
- **Init**: $k = 0$, $x_1 \in \mathcal{X}^o$
- **Repeat**

Primal Dual Smoothing Framework

- **Input**: Accuracy parameter $\eta > 0$, smoothing parameters $\lambda = 1/(2\gamma)$, $\rho = \eta/(4\Omega_{\mathcal{Y}}(\omega_{\mathcal{Y}}))$
- **Init**: $k = 0$, $x_1 \in \mathcal{X}^o$
- **Repeat**
  - $k := k + 1$.

Primal Dual Smoothing Framework

- **Input**: Accuracy parameter $\eta > 0$, smoothing parameters $\lambda = 1/(2\gamma)$, $\rho = \eta/(4\Omega_{\mathcal{Y}}(\omega_{\mathcal{Y}}))$
- **Init**: $k = 0$, $x_1 \in \mathcal{X}^o$
- **Repeat**
  - $k := k + 1$.
  - Find $x_{k+1} \in \mathcal{X}^o$ such that $Q_\rho^\lambda(x_{k+1}; x_k) \le q_\rho^\lambda(x_k) + \eta$.

Primal Dual Smoothing Framework

- **Input**: Accuracy parameter $\eta > 0$, smoothing parameters $\lambda = 1/(2\gamma)$, $\rho = \eta/(4\Omega_{\mathcal{Y}}(\omega_{\mathcal{Y}}))$
- **Init**: $k = 0$, $x_1 \in \mathcal{X}^o$
- **Repeat**
  - $k := k + 1$.
  - Find $x_{k+1} \in \mathcal{X}^o$ such that $Q_\rho^\lambda(x_{k+1}; x_k) \leq q_\rho^\lambda(x_k) + \eta$.
- **Until:**

$$\|x_{k+1} - x_k\| \leq 4\sqrt{\lambda\eta}.$$

Primal Dual Smoothing Framework

- **Input**: Accuracy parameter $\eta > 0$, smoothing parameters $\lambda = 1/(2\gamma)$, $\rho = \eta/(4\Omega_{\mathcal{Y}}(\omega_{\mathcal{Y}}))$
- **Init**: $k = 0$, $x_1 \in \mathcal{X}^o$
- **Repeat**
  - $k := k + 1$.
  - Find $x_{k+1} \in \mathcal{X}^o$ such that $Q_\rho^\lambda(x_{k+1}; x_k) \leq q_\rho^\lambda(x_k) + \eta$.
- **Until:**
$$\|x_{k+1} - x_k\| \leq 4\sqrt{\lambda\eta}.$$
- **Output**: x_k

Two approaches

Two approaches

- (Dual-then-primal) Perform dual smoothing on q to obtain q_ρ, and then apply proximal point method (PPM) on q_ρ.

Two approaches

- (Dual-then-primal) Perform dual smoothing on q to obtain q_ρ, and then apply proximal point method (PPM) on q_ρ.
- (Primal-then-dual) Apply PPM directly on q, and then in solving the sub-problem, perform dual smoothing.

Two approaches

- (Dual-then-primal) Perform dual smoothing on q to obtain q_ρ, and then apply proximal point method (PPM) on q_ρ.
- (Primal-then-dual) Apply PPM directly on q, and then in solving the sub-problem, perform dual smoothing.
- Different analyses, but the same result.

Two approaches

- (Dual-then-primal) Perform dual smoothing on q to obtain q_ρ, and then apply proximal point method (PPM) on q_ρ.
- (Primal-then-dual) Apply PPM directly on q, and then in solving the sub-problem, perform dual smoothing.
- Different analyses, but the same result.

Theorem 3

Let K denote the terminating iteration. For any $\varepsilon > 0$, if we set the accuracy parameter $\eta = \varepsilon^2\lambda/(64\beta_{\mathcal{X}}^2)$, then $\|x_K - \mathsf{prox}(q, x_K, \lambda)\| \le \varepsilon\lambda/\beta_{\mathcal{X}}$, i.e., x_K is an ε-near stationary point of (P).

Two approaches

- (Dual-then-primal) Perform dual smoothing on q to obtain q_ρ, and then apply proximal point method (PPM) on q_ρ.
- (Primal-then-dual) Apply PPM directly on q, and then in solving the sub-problem, perform dual smoothing.
- Different analyses, but the same result.

Theorem 3

Let K denote the terminating iteration. For any $\varepsilon > 0$, if we set the accuracy parameter $\eta = \varepsilon^2\lambda/(64\beta_{\mathcal{X}}^2)$, then $\|x_K - \mathsf{prox}(q, x_K, \lambda)\| \le \varepsilon\lambda/\beta_{\mathcal{X}}$, i.e., x_K is an ε-near stationary point of (P).

Theorem 4

The method terminates with no more than $\bar{K} \triangleq \lceil 2(q(x_1) - q^*)/(13\eta) \rceil$ iterations.

Two approaches

- (Dual-then-primal) Perform dual smoothing on q to obtain q_ρ, and then apply proximal point method (PPM) on q_ρ.
- (Primal-then-dual) Apply PPM directly on q, and then in solving the sub-problem, perform dual smoothing.
- Different analyses, but the same result.

Theorem 3

Let K denote the terminating iteration. For any $\varepsilon > 0$, if we set the accuracy parameter $\eta = \varepsilon^2\lambda/(64\beta_{\mathcal{X}}^2)$, then $\|x_K - \mathsf{prox}(q, x_K, \lambda)\| \leq \varepsilon\lambda/\beta_{\mathcal{X}}$, i.e., x_K is an ε-near stationary point of (P).

Theorem 4

The method terminates with no more than $\bar{K} \triangleq \lceil 2(q(x_1) - q^*)/(13\eta) \rceil$ iterations.

Proof sketch: if $\|x_{k+1} - x_k\| > 4\sqrt{\lambda\eta}$, then $q(x_{k+1}) \leq q(x_k) - (13/2)\eta$.

Solving sub-problem

The sub-problem is indeed a convex-concave saddle-point problem, i.e.,

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\; r(x) + \lambda^{-1} D_{\omega_{\mathcal{X}}}(x; x_k) + \Phi(x, y) - g(y) - \rho\omega_{\mathcal{Y}}(y),$$

where $\lambda = 1/(2\gamma)$, $\rho = \eta/(4\Omega_{\mathcal{Y}}(\omega_{\mathcal{Y}}))$.

Solving sub-problem

The sub-problem is indeed a convex-concave saddle-point problem, i.e.,

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\, r(x) + \lambda^{-1} D_{\omega_{\mathcal{X}}}(x; x_k) + \Phi(x, y) - g(y) - \rho\omega_{\mathcal{Y}}(y),$$

where $\lambda = 1/(2\gamma)$, $\rho = \eta/(4\Omega_{\mathcal{Y}}(\omega_{\mathcal{Y}}))$.

- ▷ Develop an efficient method to obtain $(x, y) \in \mathcal{X} \times \mathcal{Y}$ such that the duality gap falls below η.

Solving sub-problem

The sub-problem is indeed a convex-concave saddle-point problem, i.e.,

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\, r(x) + \lambda^{-1} D_{\omega_{\mathcal{X}}}(x; x_k) + \Phi(x, y) - g(y) - \rho\omega_{\mathcal{Y}}(y),$$

where $\lambda = 1/(2\gamma)$, $\rho = \eta/(4\Omega_{\mathcal{Y}}(\omega_{\mathcal{Y}}))$.

- ▷ Develop an efficient method to obtain $(x, y) \in \mathcal{X} \times \mathcal{Y}$ such that the duality gap falls below η.
- ▷ Based on a newly developed *non-Euclidean inexact* accelerated proximal gradient (APG) method for strongly convex composite optimization.

Solving sub-problem

The sub-problem is indeed a convex-concave saddle-point problem, i.e.,

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\, r(x)+\lambda^{-1}D_{\omega_{\mathcal{X}}}(x;x_k)+\Phi(x,y)-g(y)-\rho\omega_{\mathcal{Y}}(y),$$

where $\lambda = 1/(2\gamma)$, $\rho = \eta/(4\Omega_{\mathcal{Y}}(\omega_{\mathcal{Y}}))$.

- ▷ Develop an efficient method to obtain $(x, y) \in \mathcal{X} \times \mathcal{Y}$ such that the duality gap falls below η.
- ▷ Based on a newly developed *non-Euclidean inexact* accelerated proximal gradient (APG) method for strongly convex composite optimization.
- ▷ Apply this method to the dual function, to find a dual point with dual gap $\leq \eta/2$, and solve for a primal point with primal gap $\leq \eta/2$.

Solving sub-problem

The sub-problem is indeed a convex-concave saddle-point problem, i.e.,

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\ r(x)+\lambda^{-1}D_{\omega_{\mathcal{X}}}(x;x_k)+\Phi(x,y)-g(y)-\rho\omega_{\mathcal{Y}}(y),$$

where $\lambda = 1/(2\gamma)$, $\rho = \eta/(4\Omega_{\mathcal{Y}}(\omega_{\mathcal{Y}}))$.

- ▷ Develop an efficient method to obtain $(x, y) \in \mathcal{X} \times \mathcal{Y}$ such that the duality gap falls below η.
- ▷ Based on a newly developed *non-Euclidean inexact* accelerated proximal gradient (APG) method for strongly convex composite optimization.
- ▷ Apply this method to the dual function, to find a dual point with dual gap $\leq \eta/2$, and solve for a primal point with primal gap $\leq \eta/2$.
- ▷ This is conceptually simple, but with relatively complicated details (hence omitted).

Comparison with other methods

Algorithms	Primal Oracle Comp.	Dual Oracle Comp.
Restart	$O(\varepsilon^{-1})$	$O(\varepsilon^{-1})$
EGT-type	$O(\varepsilon^{-1/2}\log(\varepsilon^{-1}))$	$O(\varepsilon^{-1}\log(\varepsilon^{-1}))$
Our method	$O(\varepsilon^{-1/2}\log^2(\varepsilon^{-1}))$	$O(\varepsilon^{-1/2}\log(\varepsilon^{-1}))$

Overall Oracle Complexities

Based on the oracle complexities of our sub-problem solver, we can obtain the overall complexities of the smoothing framework.

Overall Oracle Complexities

Based on the oracle complexities of our sub-problem solver, we can obtain the overall complexities of the smoothing framework.

Theorem 5

For any $\varepsilon > 0$, *choose* $\eta = \varepsilon^2\lambda/(18\beta_{\mathcal{X}}^2)$. *Then it takes no more than*

$$O\big(\sqrt{\gamma(L_{xx}+\gamma)}\big(\sqrt{L_{yy}\gamma}+L_{xy}\big)\varepsilon^{-3}\log^2(\varepsilon^{-1})\big)$$

primal oracle calls and

$$O\big(\gamma\big(\sqrt{L_{yy}\gamma}+L_{xy}\big)\varepsilon^{-3}\log(\varepsilon^{-1})\big)$$

dual oracle calls to find an ε*-near-stationary point of* (P).

Thank you!

https://arxiv.org/abs/2003.04375