An Inexact Primal Dual Smoothing Framework for Large-Scale Non-Bilinear Saddle Point Problems

L. T. K. Hien[1], **Renbo Zhao**[2], William B. Haskell[3]

[1]Department of Mathematics and OR, University of Mons, Belgium

[2]Operations Research Center, MIT, MA

[3]Krannert School of Management, Purdue University, IN

INFORMS Annual Meeting
Seattle, WA, Oct. 2019

Problem Setup

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\Lambda} \ \{S(x,\lambda) := f(x) + g(x) + \Phi(x,\lambda) - h(\lambda)\} \qquad \text{(SPP)}$$

Problem Setup

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\Lambda} \ \{S(x,\lambda){:=}f(x)+g(x)+\Phi(x,\lambda)-h(\lambda)\} \qquad \text{(SPP)}$$

- $(\mathbb{E}_1, \|\cdot\|_{\mathbb{E}_1})$ and $(\mathbb{E}_2, \|\cdot\|_{\mathbb{E}_2})$ are finite-dimensional real normed spaces, with dual spaces $(\mathbb{E}_1^*, \|\cdot\|_{\mathbb{E}_1^*})$ and $(\mathbb{E}_2^*, \|\cdot\|_{\mathbb{E}_2^*})$.

Problem Setup

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\Lambda} \; \{S(x,\lambda) := f(x) + g(x) + \Phi(x,\lambda) - h(\lambda)\} \qquad \text{(SPP)}$$

- ▷ $(\mathbb{E}_1, \|\cdot\|_{\mathbb{E}_1})$ and $(\mathbb{E}_2, \|\cdot\|_{\mathbb{E}_2})$ are finite-dimensional real normed spaces, with dual spaces $(\mathbb{E}_1^*, \|\cdot\|_{\mathbb{E}_1^*})$ and $(\mathbb{E}_2^*, \|\cdot\|_{\mathbb{E}_2^*})$.
- ▷ $\mathcal{X} \subseteq \mathbb{E}_1$ and $\Lambda \subseteq \mathbb{E}_2$ are nonempty, convex and compact.

Problem Setup

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\Lambda} \left\{S(x,\lambda):=f(x)+g(x)+\Phi(x,\lambda)-h(\lambda)\right\} \qquad \text{(SPP)}$$

- $(\mathbb{E}_1, \|\cdot\|_{\mathbb{E}_1})$ and $(\mathbb{E}_2, \|\cdot\|_{\mathbb{E}_2})$ are finite-dimensional real normed spaces, with dual spaces $(\mathbb{E}_1^*, \|\cdot\|_{\mathbb{E}_1^*})$ and $(\mathbb{E}_2^*, \|\cdot\|_{\mathbb{E}_2^*})$.
- $\mathcal{X} \subseteq \mathbb{E}_1$ and $\Lambda \subseteq \mathbb{E}_2$ are nonempty, convex and compact.
- f, g and h are convex, closed and proper (CCP) functions.

Problem Setup

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\Lambda} \; \{S(x,\lambda) := f(x) + g(x) + \Phi(x,\lambda) - h(\lambda)\} \qquad \text{(SPP)}$$

- $(\mathbb{E}_1, \|\cdot\|_{\mathbb{E}_1})$ and $(\mathbb{E}_2, \|\cdot\|_{\mathbb{E}_2})$ are finite-dimensional real normed spaces, with dual spaces $(\mathbb{E}_1^*, \|\cdot\|_{\mathbb{E}_1^*})$ and $(\mathbb{E}_2^*, \|\cdot\|_{\mathbb{E}_2^*})$.
- $\mathcal{X} \subseteq \mathbb{E}_1$ and $\Lambda \subseteq \mathbb{E}_2$ are nonempty, convex and compact.
- f, g and h are convex, closed and proper (CCP) functions.
- f is μ-strongly convex (s.c.) and L-smooth on $\mathcal{X}$ ($\mu > 0$), i.e., $\frac{\mu}{2}\|x - x'\|^2_{\mathbb{E}_1} \le f(x) - f(x') - \langle \nabla f(x'), x - x'\rangle \le \frac{L}{2}\|x - x'\|^2_{\mathbb{E}_1}, \forall\, x, x' \in X.$

Problem Setup

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\Lambda} \{S(x,\lambda){:=}f(x)+g(x)+\Phi(x,\lambda)-h(\lambda)\} \qquad \text{(SPP)}$$

- $(\mathbb{E}_1, \|\cdot\|_{\mathbb{E}_1})$ and $(\mathbb{E}_2, \|\cdot\|_{\mathbb{E}_2})$ are finite-dimensional real normed spaces, with dual spaces $(\mathbb{E}_1^*, \|\cdot\|_{\mathbb{E}_1^*})$ and $(\mathbb{E}_2^*, \|\cdot\|_{\mathbb{E}_2^*})$.
- $\mathcal{X} \subseteq \mathbb{E}_1$ and $\Lambda \subseteq \mathbb{E}_2$ are nonempty, convex and compact.
- f, g and h are convex, closed and proper (CCP) functions.
- f is μ-strongly convex (s.c.) and L-smooth on $\mathcal{X}$ ($\mu > 0$), i.e., $\frac{\mu}{2}\|x-x'\|_{\mathbb{E}_1}^2 \le f(x)-f(x')-\langle \nabla f(x'), x-x'\rangle \le \frac{L}{2}\|x-x'\|_{\mathbb{E}_1}^2, \forall\, x, x' \in X.$
- g and h have easily computable proximal operators.

Problem Setup (Cont'd)

$$\Phi(x, \lambda) := \frac{1}{n} \sum_{i=1}^{n} \Phi_i(x, \lambda)$$

For any $i \in [n]$:

Problem Setup (Cont'd)

$$\Phi(x, \lambda) := \frac{1}{n} \sum_{i=1}^{n} \Phi_i(x, \lambda)$$

For any $i \in [n]$:

- $\Phi_i(\cdot, \lambda)$ is convex and $\Phi_i(x, \cdot)$ is concave for any $(x, \lambda) \in \mathcal{X} \times \Lambda$.

Problem Setup (Cont'd)

$$\Phi(x,\lambda):=\frac{1}{n}\sum_{i=1}^{n}\Phi_i(x,\lambda)$$

For any $i \in [n]$:

- $\Phi_i(\cdot,\lambda)$ is convex and $\Phi_i(x,\cdot)$ is concave for any $(x,\lambda) \in \mathcal{X} \times \Lambda$.

- Φ_i is $(L^i_{xx},\ L^i_{x\lambda},\ L^i_{\lambda\lambda})$-smooth, i.e., for any $x, x' \in \mathcal{X}$ and $\lambda, \lambda' \in \Lambda$,

$$\|\nabla_x \Phi_i(x,\lambda) - \nabla_x \Phi_i(x',\lambda)\|_{\mathbb{E}_1^*} \le L^i_{xx}\,\|x - x'\|_{\mathbb{E}_1}\,,$$
$$\|\nabla_x \Phi_i(x,\lambda) - \nabla_x \Phi_i(x,\lambda')\|_{\mathbb{E}_1^*} \le L^i_{x\lambda}\,\|\lambda - \lambda'\|_{\mathbb{E}_2}\,,$$
$$\|\nabla_\lambda \Phi_i(x,\lambda) - \nabla_\lambda \Phi_i(x',\lambda)\|_{\mathbb{E}_2^*} \le L^i_{x\lambda}\,\|x - x'\|_{\mathbb{E}_1}\,,$$
$$\|\nabla_\lambda \Phi_i(x,\lambda) - \nabla_\lambda \Phi_i(x,\lambda')\|_{\mathbb{E}_2^*} \le L^i_{\lambda\lambda}\,\|\lambda - \lambda'\|_{\mathbb{E}_2}\,.$$

Problem Setup (Cont'd)

$$\Phi(x,\lambda):=\frac{1}{n}\sum_{i=1}^{n}\Phi_i(x,\lambda)$$

For any $i \in [n]$:

- $\Phi_i(\cdot,\lambda)$ is convex and $\Phi_i(x,\cdot)$ is concave for any $(x,\lambda) \in \mathcal{X} \times \Lambda$.
- Φ_i is $(L^i_{xx},\, L^i_{x\lambda},\, L^i_{\lambda\lambda})$-smooth, i.e., for any $x, x' \in \mathcal{X}$ and $\lambda, \lambda' \in \Lambda$,

$$\begin{aligned}
\|\nabla_x\Phi_i(x,\lambda) - \nabla_x\Phi_i(x',\lambda)\|_{\mathbb{E}_1^*} &\le L^i_{xx}\,\|x - x'\|_{\mathbb{E}_1}\,, \\
\|\nabla_x\Phi_i(x,\lambda) - \nabla_x\Phi_i(x,\lambda')\|_{\mathbb{E}_1^*} &\le L^i_{x\lambda}\,\|\lambda - \lambda'\|_{\mathbb{E}_2}\,, \\
\|\nabla_\lambda\Phi_i(x,\lambda) - \nabla_\lambda\Phi_i(x',\lambda)\|_{\mathbb{E}_2^*} &\le L^i_{x\lambda}\,\|x - x'\|_{\mathbb{E}_1}\,, \\
\|\nabla_\lambda\Phi_i(x,\lambda) - \nabla_\lambda\Phi_i(x,\lambda')\|_{\mathbb{E}_2^*} &\le L^i_{\lambda\lambda}\,\|\lambda - \lambda'\|_{\mathbb{E}_2}\,.
\end{aligned}$$

- Φ is $(L_{xx},\, L_{x\lambda},\, L_{\lambda\lambda})$-smooth, where $L_{xx} \le (1/n)\sum_{i=1}^{n} L^i_{xx}$, $L_{\lambda x} \le (1/n)\sum_{i=1}^{n} L^i_{x\lambda}$, $L_{\lambda\lambda} \le (1/n)\sum_{i=1}^{n} L^i_{\lambda\lambda}$.

Application I: Constrained Optimization

$$\min_{x\in\mathcal{X}} f(x) + r(x) \quad \text{s.t.} \quad g_i(x) \le 0, \forall\, i \in [n]$$

Application I: Constrained Optimization

$$\min_{x \in \mathcal{X}} f(x) + r(x) \quad \text{s.t.} \quad g_i(x) \le 0, \forall\, i \in [n]$$

- f is μ-strongly convex (s.c.) and L-smooth on $\mathcal{X}$.

Application I: Constrained Optimization

$$\min_{x\in\mathcal{X}} f(x) + r(x) \quad \text{s. t.} \quad g_i(x) \leq 0, \forall\, i \in [n]$$

- ▷ f is μ-strongly convex (s.c.) and L-smooth on $\mathcal{X}$.
- ▷ r is CCP with an easily computable proximal operator.

Application I: Constrained Optimization

$$\min_{x\in\mathcal{X}} f(x) + r(x) \quad \text{s. t.} \quad g_i(x) \le 0, \forall\, i \in [n]$$

- f is μ-strongly convex (s.c.) and L-smooth on $\mathcal{X}$.
- r is CCP with an easily computable proximal operator.
- For each $i \in [n]$, g_i is convex and α_i-smooth on $\mathcal{X}$.

Application I: Constrained Optimization

$$\min_{x \in \mathcal{X}} f(x) + r(x) \quad \text{s. t.} \quad g_i(x) \leq 0, \forall\, i \in [n]$$

- f is μ-strongly convex (s.c.) and L-smooth on $\mathcal{X}$.
- r is CCP with an easily computable proximal operator.
- For each $i \in [n]$, g_i is convex and α_i-smooth on $\mathcal{X}$.
- Slater condition holds $\Rightarrow$ no duality gap and an optimal primal-dual pair (x^*, λ^*) exists.

Application I: Constrained Optimization

$$\min_{x\in\mathcal{X}} f(x) + r(x) \quad \text{s.t.} \quad g_i(x) \leq 0, \forall\, i \in [n]$$

- f is μ-strongly convex (s.c.) and L-smooth on $\mathcal{X}$.
- r is CCP with an easily computable proximal operator.
- For each $i \in [n]$, g_i is convex and α_i-smooth on $\mathcal{X}$.
- Slater condition holds $\Rightarrow$ no duality gap and an optimal primal-dual pair (x^*, λ^*) exists.

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\mathbb{R}^n_+} \{S(x,\lambda) = f(x) + r(x) + (1/n)\textstyle\sum_{i=1}^{n} n\lambda_i g_i(x)\} \qquad \text{(Lag)}$$

Application I: Constrained Optimization

$$\min_{x \in \mathcal{X}} f(x) + r(x) \quad \text{s.t.} \quad g_i(x) \le 0, \forall\, i \in [n]$$

- f is μ-strongly convex (s.c.) and L-smooth on $\mathcal{X}$.
- r is CCP with an easily computable proximal operator.
- For each $i \in [n]$, g_i is convex and α_i-smooth on $\mathcal{X}$.
- Slater condition holds $\Rightarrow$ no duality gap and an optimal primal-dual pair (x^*, λ^*) exists.

$$\min_{x \in \mathcal{X}} \max_{\lambda \in \mathbb{R}^n_+} \{S(x, \lambda) = f(x) + r(x) + (1/n)\textstyle\sum_{i=1}^{n} n\lambda_i g_i(x)\} \qquad \text{(Lag)}$$

- Any saddle point of (Lag) is an optimal primal-dual pair.

Application I: Constrained Optimization

$$\min_{x \in \mathcal{X}} f(x) + r(x) \quad \text{s. t.} \quad g_i(x) \leq 0, \forall\, i \in [n]$$

- f is μ-strongly convex (s.c.) and L-smooth on $\mathcal{X}$.
- r is CCP with an easily computable proximal operator.
- For each $i \in [n]$, g_i is convex and α_i-smooth on $\mathcal{X}$.
- Slater condition holds $\Rightarrow$ no duality gap and an optimal primal-dual pair (x^*, λ^*) exists.

$$\min_{x \in \mathcal{X}} \max_{\lambda \in \mathbb{R}^n_+} \{S(x, \lambda) = f(x) + r(x) + (1/n) \textstyle\sum_{i=1}^n n \lambda_i g_i(x)\} \qquad \text{(Lag)}$$

- Any saddle point of (Lag) is an optimal primal-dual pair.
- $\mathbb{R}^n_+$ is unbounded: allowed since different convergence criteria (other than duality gap) is used.

Application II: Maximum Margin Clustering

Application II: Maximum Margin Clustering

- Set $\mathcal{D}$ of m objects consisting of two unknown disjoint subsets $\mathcal{D}_1 \cup \mathcal{D}_2 = \mathcal{D}$

Application II: Maximum Margin Clustering

- ▷ Set $\mathcal{D}$ of m objects consisting of two unknown disjoint subsets $\mathcal{D}_1 \cup \mathcal{D}_2 = \mathcal{D}$
- ▷ n noisy samples $\{\mathcal{S}_i\}_{i=1}^n$ of $\mathcal{D}$ consisting of m objects with kernel matrices $\{K_i\}_{i=1}^n \subseteq \mathbb{S}_+^m$

Application II: Maximum Margin Clustering

- Set $\mathcal{D}$ of m objects consisting of two unknown disjoint subsets $\mathcal{D}_1 \cup \mathcal{D}_2 = \mathcal{D}$
- n noisy samples $\{\mathcal{S}_i\}_{i=1}^n$ of $\mathcal{D}$ consisting of m objects with kernel matrices $\{K_i\}_{i=1}^n \subseteq \mathbb{S}_+^m$
- Find a label kernel matrix $M \in \mathbb{S}_+^m$ that assigns $z \in \mathcal{D}$ to $\mathcal{D}_1$ or $\mathcal{D}_2$

Application II: Maximum Margin Clustering

- Set $\mathcal{D}$ of m objects consisting of two unknown disjoint subsets $\mathcal{D}_1 \cup \mathcal{D}_2 = \mathcal{D}$
- n noisy samples $\{\mathcal{S}_i\}_{i=1}^n$ of $\mathcal{D}$ consisting of m objects with kernel matrices $\{K_i\}_{i=1}^n \subseteq \mathbb{S}_+^m$
- Find a label kernel matrix $M \in \mathbb{S}_+^m$ that assigns $z \in \mathcal{D}$ to $\mathcal{D}_1$ or $\mathcal{D}_2$

$$\min_{M \in \mathsf{M}} \max_{\lambda_i \in \Lambda, \forall i \in [n]} \frac{1}{n} \sum_{i=1}^{n} \left[-\langle \lambda_i \lambda_i^T, \, K_i \circ M + \alpha \, I \rangle + 2 \, \lambda_i^T e \right],$$

Application II: Maximum Margin Clustering

- ▷ Set $\mathcal{D}$ of m objects consisting of two unknown disjoint subsets $\mathcal{D}_1 \cup \mathcal{D}_2 = \mathcal{D}$
- ▷ n noisy samples $\{\mathcal{S}_i\}_{i=1}^n$ of $\mathcal{D}$ consisting of m objects with kernel matrices $\{K_i\}_{i=1}^n \subseteq \mathbb{S}_+^m$
- ▷ Find a label kernel matrix $M \in \mathbb{S}_+^m$ that assigns $z \in \mathcal{D}$ to $\mathcal{D}_1$ or $\mathcal{D}_2$

$$\min_{M \in \mathsf{M}} \max_{\lambda_i \in \Lambda, \forall i \in [n]} \frac{1}{n} \sum_{i=1}^n \left[-\langle \lambda_i \lambda_i^T, K_i \circ M + \alpha I \rangle + 2 \lambda_i^T e \right],$$

where

$$\mathsf{M} := \left\{ M \in \mathbb{S}_+^m : \operatorname{diag}(M) = e, |e^T M| \le l \right\}$$

$$\Lambda := \{\lambda \in \mathbb{R}^m : 0 \le \lambda_i \le C, \forall i \in [m]\}$$

$$l, C : \text{finite constants}$$

Application III: Distributionally Robust Optimization

Application III: Distributionally Robust Optimization

- Convex and compact decision set $\mathcal{X} \subseteq \mathbb{R}^d$

Application III: Distributionally Robust Optimization

- ▷ Convex and compact decision set $\mathcal{X} \subseteq \mathbb{R}^d$
- ▷ Random variable ξ with realizations $\Xi := \{\xi_1, \ldots, \xi_n\}$

Application III: Distributionally Robust Optimization

- Convex and compact decision set $\mathcal{X} \subseteq \mathbb{R}^d$
- Random variable ξ with realizations $\Xi := \{\xi_1, \ldots, \xi_n\}$
- Convex and compact distributional uncertainty set $\mathcal{P}$, each element P is a probability distribution with support Ξ

Application III: Distributionally Robust Optimization

- ▷ Convex and compact decision set $\mathcal{X} \subseteq \mathbb{R}^d$
- ▷ Random variable ξ with realizations $\Xi := \{\xi_1, \ldots, \xi_n\}$
- ▷ Convex and compact distributional uncertainty set $\mathcal{P}$, each element P is a probability distribution with support Ξ
- ▷ For each $i \in [n]$, μ_i-strongly convex loss function $f(\cdot, \xi_i)$

Application III: Distributionally Robust Optimization

- ▷ Convex and compact decision set $\mathcal{X} \subseteq \mathbb{R}^d$
- ▷ Random variable ξ with realizations $\Xi := \{\xi_1, \ldots, \xi_n\}$
- ▷ Convex and compact distributional uncertainty set $\mathcal{P}$, each element P is a probability distribution with support Ξ
- ▷ For each $i \in [n]$, μ_i-strongly convex loss function $f(\cdot, \xi_i)$

$$\min_{x \in \mathcal{X}} \max_{P \in \mathcal{P}} \sum_{i=1}^{n} p_i f(x, \xi_i).$$

Background (SPP)

Background (SPP)

- Sion's minimax theorem ensures (SPP) has at least one saddle point $(x^*, \lambda^*) \in X \times \Lambda$, i.e.,

$$S(x^*, \lambda) \leq S(x^*, \lambda^*) \leq S(x, \lambda^*), \quad \forall\, (x, \lambda) \in X \times \Lambda.$$

Background (SPP)

- Sion's minimax theorem ensures (SPP) has at least one saddle point $(x^*, \lambda^*) \in X \times \Lambda$, i.e.,

$$S(x^*, \lambda) \le S(x^*, \lambda^*) \le S(x, \lambda^*), \quad \forall\, (x, \lambda) \in X \times \Lambda.$$

- For any $(x, \lambda) \in X \times \Lambda$, define

$$\widehat{\psi}^{\mathrm{P}}(x) := \max_{\lambda \in \Lambda} \Phi(x, \lambda) - H(\lambda)$$

$$\psi^{\mathrm{P}}(x) := \max_{\lambda \in \Lambda} S(x, \lambda) = f(x) + g(x) + \widehat{\psi}^{\mathrm{P}}(x) \quad \text{(Primal func.)}$$

$$\widehat{\psi}^{\mathrm{D}}(\lambda) := \min_{x \in \mathcal{X}} f(x) + g(x) + \Phi(x, \lambda)$$

$$\psi^{\mathrm{D}}(\lambda) := \min_{x \in \mathcal{X}} S(x, \lambda) = \widehat{\psi}^{\mathrm{D}}(\lambda) - h(\lambda). \quad \text{(Dual func.)}$$

$$\Delta(x, \lambda) := \psi^{\mathrm{P}}(x) - \psi^{\mathrm{D}}(\lambda). \quad \text{(Duality gap)}$$

Background (SPP)

- ▷ Sion's minimax theorem ensures (SPP) has at least one saddle point $(x^*, \lambda^*) \in X \times \Lambda$, i.e.,

$$S(x^*, \lambda) \leq S(x^*, \lambda^*) \leq S(x, \lambda^*), \quad \forall\, (x, \lambda) \in X \times \Lambda.$$

- ▷ For any $(x, \lambda) \in X \times \Lambda$, define

$$\widehat{\psi}^{\mathrm{P}}(x) := \max_{\lambda \in \Lambda} \Phi(x, \lambda) - H(\lambda)$$

$$\psi^{\mathrm{P}}(x) := \max_{\lambda \in \Lambda} S(x, \lambda) = f(x) + g(x) + \widehat{\psi}^{\mathrm{P}}(x) \quad \text{(Primal func.)}$$

$$\widehat{\psi}^{\mathrm{D}}(\lambda) := \min_{x \in \mathcal{X}} f(x) + g(x) + \Phi(x, \lambda)$$

$$\psi^{\mathrm{D}}(\lambda) := \min_{x \in \mathcal{X}} S(x, \lambda) = \widehat{\psi}^{\mathrm{D}}(\lambda) - h(\lambda). \quad \text{(Dual func.)}$$

$$\Delta(x, \lambda) := \psi^{\mathrm{P}}(x) - \psi^{\mathrm{D}}(\lambda). \quad \text{(Duality gap)}$$

- ▷ The saddle point (x^*, λ^*) exists $\Rightarrow \psi^{\mathrm{P}}(x^*) = \Phi(x^*, \lambda^*) = \psi^{\mathrm{D}}(\lambda^*)$.

Background (Bregman Proximal Projection)

The (non-smooth) function h have easily computable proximal operators.

Background (Bregman Proximal Projection)

The (non-smooth) function h have easily computable proximal operators.

$$\Updownarrow$$

There exists a Legendre function $\omega : \mathbb{E}_2 \to \overline{\mathbb{R}}$ that is 1-s.c. and continuous on Λ, such that for any $\xi \in \mathbb{E}_2^*$ and $\tau > 0$,

$$\min_{\lambda \in \Lambda} h(\lambda) + \langle \xi, \lambda \rangle + \tau^{-1}\omega(\lambda) \qquad \text{(BPP)}$$

has easily computable solution.

Background (Bregman Proximal Projection)

The (non-smooth) function h have easily computable proximal operators.

$$\Updownarrow$$

There exists a Legendre function $\omega : \mathbb{E}_2 \to \overline{\mathbb{R}}$ that is 1-s.c. and continuous on Λ, such that for any $\xi \in \mathbb{E}_2^*$ and $\tau > 0$,

$$\min_{\lambda \in \Lambda} h(\lambda) + \langle \xi, \lambda \rangle + \tau^{-1} \omega(\lambda) \qquad \text{(BPP)}$$

has easily computable solution.

The same applies to the (non-smooth) function g.

Background (Smoothing)

For any $(x, \lambda) \in X \times \Lambda$, define

$$\begin{aligned}
S_\rho(x,\lambda) &:= S(x,\lambda) - \rho\omega(\lambda) && \text{(Regularized saddle func.)} \\
\widehat{\psi}^{\mathrm{P}}_\rho(x) &:= \max_{\lambda\in\Lambda} \Phi(x,\lambda) - h(\lambda) - \rho\omega(\lambda) && \\
\psi^{\mathrm{P}}_\rho(x) &:= \max_{\lambda\in\Lambda} S_\rho(x,\lambda) && \\
&= f(x) + g(x) + \widehat{\psi}^{\mathrm{P}}_\rho(x) && \text{(Smoothed primal func.)} \\
\Delta_\rho(x,\lambda) &:= \psi^{\mathrm{P}}_\rho(x) - \psi^{\mathrm{D}}(\lambda) && \text{(Smoothed duality gap)}
\end{aligned}$$

Background (Smoothing)

For any $(x, \lambda) \in X \times \Lambda$, define

$$S_\rho(x,\lambda) := S(x,\lambda) - \rho\omega(\lambda) \qquad \text{(Regularized saddle func.)}$$

$$\widehat{\psi}^{\mathrm{P}}_\rho(x) := \max_{\lambda\in\Lambda} \Phi(x,\lambda) - h(\lambda) - \rho\omega(\lambda)$$

$$\psi^{\mathrm{P}}_\rho(x) := \max_{\lambda\in\Lambda} S_\rho(x,\lambda)$$

$$= f(x) + g(x) + \widehat{\psi}^{\mathrm{P}}_\rho(x) \qquad \text{(Smoothed primal func.)}$$

$$\Delta_\rho(x,\lambda) := \psi^{\mathrm{P}}_\rho(x) - \psi^{\mathrm{D}}(\lambda) \qquad \text{(Smoothed duality gap)}$$

Recall $\widehat{\psi}^{\mathrm{D}}(\lambda) := \min_{x\in\mathcal{X}} f(x) + g(x) + \Phi(x,\lambda)$ and f is μ-s.c. on $\mathcal{X}$.
Define $x^*(\lambda) := \arg\min_{x\in\mathcal{X}} f(x) + g(x) + \Phi(x,\lambda)$.

Background (Smoothing)

For any $(x, \lambda) \in X \times \Lambda$, define

$$S_\rho(x,\lambda) := S(x,\lambda) - \rho\omega(\lambda) \qquad \text{(Regularized saddle func.)}$$

$$\widehat{\psi}_\rho^{\mathrm{P}}(x) := \max_{\lambda\in\Lambda} \Phi(x,\lambda) - h(\lambda) - \rho\omega(\lambda)$$

$$\psi_\rho^{\mathrm{P}}(x) := \max_{\lambda\in\Lambda} S_\rho(x,\lambda)$$

$$= f(x) + g(x) + \widehat{\psi}_\rho^{\mathrm{P}}(x) \qquad \text{(Smoothed primal func.)}$$

$$\Delta_\rho(x,\lambda) := \psi_\rho^{\mathrm{P}}(x) - \psi^{\mathrm{D}}(\lambda) \qquad \text{(Smoothed duality gap)}$$

Recall $\widehat{\psi}^{\mathrm{D}}(\lambda) := \min_{x\in\mathcal{X}} f(x) + g(x) + \Phi(x,\lambda)$ and f is μ-s.c. on $\mathcal{X}$.
Define $x^*(\lambda) := \arg\min_{x\in\mathcal{X}} f(x) + g(x) + \Phi(x,\lambda)$.

Lemma 1 (Smoothness of $\widehat{\psi}^{\mathrm{D}}$)

The function $\widehat{\psi}^{\mathrm{D}}$ is differentiable on $\mathbb{E}_2$ and $\nabla\widehat{\psi}^{\mathrm{D}}(\lambda) = \nabla_\lambda\Phi(x^*(\lambda),\lambda)$, for any $\lambda \in \mathbb{E}_2$. In addition, $\nabla\widehat{\psi}^{\mathrm{D}}$ is L_{D}-Lipschitz on $\mathbb{E}_2$, where

$$L_{\mathrm{D}} := L_{\lambda\lambda} + 2L_{\lambda x}^2/\mu$$

Deterministic Smoothing Framework (DSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences; $\{\tau_k\}_{k\geq 0}$: interpolation sequence; N_1, N_2: deterministic first-order solvers.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

Deterministic Smoothing Framework (DSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences; $\{\tau_k\}_{k\geq 0}$: interpolation sequence; N_1, N_2: deterministic first-order solvers.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

- Use N_1 to find $\tilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that

$$\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \tilde{\lambda}_{\rho_k,\eta_k}(x^k)) \leq \eta_k. \tag{DS1}$$

Deterministic Smoothing Framework (DSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences; $\{\tau_k\}_{k\geq 0}$: interpolation sequence; N_1, N_2: deterministic first-order solvers.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

- Use N_1 to find $\tilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that

$$\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \tilde{\lambda}_{\rho_k,\eta_k}(x^k)) \leq \eta_k. \tag{DS1}$$

- $\hat{\lambda}^k := \tau_k \lambda^k + (1-\tau_k)\tilde{\lambda}_{\rho_k,\eta_k}(x^k)$.

Deterministic Smoothing Framework (DSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences; $\{\tau_k\}_{k\geq 0}$: interpolation sequence; N_1, N_2: deterministic first-order solvers.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

- Use N_1 to find $\tilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that

$$\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \tilde{\lambda}_{\rho_k,\eta_k}(x^k)) \leq \eta_k. \tag{DS1}$$

- $\hat{\lambda}^k := \tau_k \lambda^k + (1-\tau_k)\tilde{\lambda}_{\rho_k,\eta_k}(x^k)$.
- Use N_2 to find $\tilde{x}_{\gamma_k}(\hat{\lambda}^k) \in \mathcal{X}$ such that

$$S(\tilde{x}_{\gamma_k}(\hat{\lambda}^k), \hat{\lambda}^k) - \psi^{\mathrm{D}}(\hat{\lambda}^k) \leq \gamma_k. \tag{PS}$$

Deterministic Smoothing Framework (DSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences; $\{\tau_k\}_{k\geq 0}$: interpolation sequence; N_1, N_2: deterministic first-order solvers.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

- Use N_1 to find $\tilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that

$$\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \tilde{\lambda}_{\rho_k,\eta_k}(x^k)) \leq \eta_k. \tag{DS1}$$

- $\hat{\lambda}^k := \tau_k \lambda^k + (1-\tau_k)\tilde{\lambda}_{\rho_k,\eta_k}(x^k)$.
- Use N_2 to find $\tilde{x}_{\gamma_k}(\hat{\lambda}^k) \in \mathcal{X}$ such that

$$S(\tilde{x}_{\gamma_k}(\hat{\lambda}^k), \hat{\lambda}^k) - \psi^{\mathrm{D}}(\hat{\lambda}^k) \leq \gamma_k. \tag{PS}$$

- $x^{k+1} := \tau_k x^k + (1-\tau_k)\tilde{x}_{\gamma_k}(\hat{\lambda}^k)$, $\rho_{k+1} := \tau_k \rho_k$.

Deterministic Smoothing Framework (DSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences; $\{\tau_k\}_{k\geq 0}$: interpolation sequence; N_1, N_2: deterministic first-order solvers.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

- Use N_1 to find $\tilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that

$$\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \tilde{\lambda}_{\rho_k,\eta_k}(x^k)) \leq \eta_k. \qquad \text{(DS1)}$$

- $\hat{\lambda}^k := \tau_k \lambda^k + (1-\tau_k)\tilde{\lambda}_{\rho_k,\eta_k}(x^k)$.
- Use N_2 to find $\tilde{x}_{\gamma_k}(\hat{\lambda}^k) \in \mathcal{X}$ such that

$$S(\tilde{x}_{\gamma_k}(\hat{\lambda}^k), \hat{\lambda}^k) - \psi^{\mathrm{D}}(\hat{\lambda}^k) \leq \gamma_k. \qquad \text{(PS)}$$

- $x^{k+1} := \tau_k x^k + (1-\tau_k)\tilde{x}_{\gamma_k}(\hat{\lambda}^k)$, $\rho_{k+1} := \tau_k \rho_k$.
- Use N_1 to find $\tilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1}) \in \Lambda$ such that

$$\psi^{\mathrm{P}}_{\rho_{k+1}}(x^{k+1}) - S_{\rho_{k+1}}(x^{k+1}, \tilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})) \leq \eta_k. \qquad \text{(DS2)}$$

Deterministic Smoothing Framework (DSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences; $\{\tau_k\}_{k\geq 0}$: interpolation sequence; N_1, N_2: deterministic first-order solvers.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

- Use N_1 to find $\tilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that

$$\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \tilde{\lambda}_{\rho_k,\eta_k}(x^k)) \leq \eta_k. \tag{DS1}$$

- $\hat{\lambda}^k := \tau_k \lambda^k + (1-\tau_k)\tilde{\lambda}_{\rho_k,\eta_k}(x^k)$.
- Use N_2 to find $\tilde{x}_{\gamma_k}(\hat{\lambda}^k) \in \mathcal{X}$ such that

$$S(\tilde{x}_{\gamma_k}(\hat{\lambda}^k), \hat{\lambda}^k) - \psi^{\mathrm{D}}(\hat{\lambda}^k) \leq \gamma_k. \tag{PS}$$

- $x^{k+1} := \tau_k x^k + (1-\tau_k)\tilde{x}_{\gamma_k}(\hat{\lambda}^k)$, $\rho_{k+1} := \tau_k \rho_k$.
- Use N_1 to find $\tilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1}) \in \Lambda$ such that

$$\psi^{\mathrm{P}}_{\rho_{k+1}}(x^{k+1}) - S_{\rho_{k+1}}(x^{k+1}, \tilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})) \leq \eta_k. \tag{DS2}$$

- $\lambda^{k+1} := \tau_k \lambda^k + (1-\tau_k)\tilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})$, $k := k+1$.

Solving Sub-problems Inexactly

$$\min_{x\in\mathcal{X}} \left\{P(x,\lambda):=f(x)+g(x)+\Phi(x,\lambda)\right\} \tag{PS}$$

Solving Sub-problems Inexactly

$$\min_{x\in\mathcal{X}} \left\{P(x,\lambda){:=}f(x)+g(x)+\Phi(x,\lambda)\right\} \tag{PS}$$

▷ (PS) is μ-s.c., $(L+L_{xx})$-smooth, composite.

Solving Sub-problems Inexactly

$$\min_{x\in\mathcal{X}} \left\{ P(x,\lambda) := f(x) + g(x) + \Phi(x,\lambda) \right\} \tag{PS}$$

- (PS) is μ-s.c., $(L + L_{xx})$-smooth, composite.
- The constraint set $\mathcal{X}$ is compact.

Solving Sub-problems Inexactly

$$\min_{x\in\mathcal{X}} \{P(x,\lambda) := f(x) + g(x) + \Phi(x,\lambda)\} \tag{PS}$$

- (PS) is μ-s.c., $(L + L_{xx})$-smooth, composite.
- The constraint set $\mathcal{X}$ is compact.
- Use optimal first-order solver, e.g., APG in [Nesterov'13].

Solving Sub-problems Inexactly

$$\min_{x\in\mathcal{X}} \left\{P(x,\lambda):=f(x)+g(x)+\Phi(x,\lambda)\right\} \qquad \text{(PS)}$$

- ▷ (PS) is μ-s.c., $(L+L_{xx})$-smooth, composite.
- ▷ The constraint set $\mathcal{X}$ is compact.
- ▷ Use optimal first-order solver, e.g., APG in [Nesterov'13].
- ▷ $\kappa_{\mathcal{X}}:=(L+L_{xx})/\mu$ and $D_{\mathcal{X}}:=\max_{x,x'\in\mathcal{X}}\|x-x'\|<+\infty$, then

$$P(\tilde{x}^N,\lambda)-P^*(\lambda)\le L_P\left(1+\sqrt{\kappa_{\mathcal{X}}/2}\right)^{-2(N-1)} D_{\mathcal{X}}^2.$$

Solving Sub-problems Inexactly

$$\min_{x\in\mathcal{X}} \{P(x,\lambda) := f(x) + g(x) + \Phi(x,\lambda)\} \tag{PS}$$

- (PS) is μ-s.c., $(L + L_{xx})$-smooth, composite.
- The constraint set $\mathcal{X}$ is compact.
- Use optimal first-order solver, e.g., APG in [Nesterov'13].
- $\kappa_{\mathcal{X}} := (L + L_{xx})/\mu$ and $D_{\mathcal{X}} := \max_{x,x'\in\mathcal{X}} \|x - x'\| < +\infty$, then

$$P(\tilde{x}^N, \lambda) - P^*(\lambda) \le L_P \left(1 + \sqrt{\kappa_{\mathcal{X}}/2}\right)^{-2(N-1)} D_{\mathcal{X}}^2.$$

$$N \ge \left\lceil \sqrt{\kappa_{\mathcal{X}}} \log\left(L_P D_X^2/\epsilon\right)\right\rceil \implies P(\tilde{x}^N, \lambda) - P^*(\lambda) \le \epsilon.$$

No need to know $P^*(\lambda)$ or $x^*(\lambda)$!

Outer Iteration Complexity

Theorem 2 (Outer Iteration Complexity of DSF)

If we choose $\rho_0 = 8L_{\mathrm{D}}$ $(L_{\mathrm{D}} = L_{\lambda\lambda} + 2L_{\lambda x}^2/\mu)$ *and for any* $k \in \mathbb{Z}_+$,

$$\tau_k = \frac{k+1}{k+3}, \quad \gamma_k = \frac{\varepsilon}{4(k+3)} \quad \textit{and} \quad \eta_k = \frac{\varepsilon}{4(k+3)}, \tag{1}$$

then for any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \Lambda$ *and* $K \in \mathbb{N}$,

$$\Delta(x^K, \lambda^K) \leq \frac{32L_{\mathrm{D}}D_\Lambda^2 + 2\Delta(x^0, \lambda^0)}{(K+1)(K+2)} + \frac{\varepsilon}{2}. \tag{2}$$

Outer Iteration Complexity

Theorem 2 (Outer Iteration Complexity of DSF)

If we choose $\rho_0 = 8L_{\mathrm{D}}$ $(L_{\mathrm{D}} = L_{\lambda\lambda} + 2L_{\lambda x}^2/\mu)$ *and for any* $k \in \mathbb{Z}_+$,

$$\tau_k = \frac{k+1}{k+3}, \quad \gamma_k = \frac{\varepsilon}{4(k+3)} \quad \textit{and} \quad \eta_k = \frac{\varepsilon}{4(k+3)}, \tag{1}$$

then for any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \Lambda$ *and* $K \in \mathbb{N}$,

$$\Delta(x^K, \lambda^K) \leq \frac{32L_{\mathrm{D}}D_\Lambda^2 + 2\Delta(x^0, \lambda^0)}{(K+1)(K+2)} + \frac{\varepsilon}{2}. \tag{2}$$

Thus, to achieve an ε*-duality gap, the outer iteration complexity is* $O(\sqrt{L_{\mathrm{D}}/\varepsilon}) = O(\sqrt{L_{\lambda\lambda}/\varepsilon} + L_{\lambda x}/\sqrt{\mu\varepsilon})$.

Inner Iteration Complexity (Oracle Complexity)

Theorem 3 (Oracle complexity of DSF)

For any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \Lambda$*, let* $C^{\mathrm{P}}_{\mathsf{det}}$ *and* $C^{\mathrm{D}}_{\mathsf{det}}$ *denote the primal and dual oracle complexities to achieve an* ε*-duality gap, respectively. Then we have*

$$C^{\mathrm{P}}_{\mathsf{det}} = O\left(n\sqrt{\kappa_{\mathcal{X}} L_{\mathrm{D}}/\varepsilon}\,\log\left((L + L_{xx})L_{\mathrm{D}}/\varepsilon\right)\right),$$

$$C^{\mathrm{D}}_{\mathsf{det}} = O\left(n(\sqrt{L_{\lambda\lambda}L_{\mathrm{D}}}/\varepsilon)\,\log\left(L_{\lambda\lambda}L_{\mathrm{D}}/\varepsilon\right)\right).$$

Randomized Smoothing Framework (RSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences, $\{\tau_k\}_{k\geq 0}$: interpolation sequence; M_1, M_2: randomized subroutines.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

Randomized Smoothing Framework (RSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences, $\{\tau_k\}_{k\geq 0}$: interpolation sequence; M_1, M_2: randomized subroutines.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

- Use M_1 to find $\tilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that

$$\mathbb{E}\big[\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \tilde{\lambda}_{\rho_k,\eta_k}(x^k)) \,\big|\, \mathcal{F}_{k,0}\big] \leq \eta_k \quad \text{a.s.} \tag{rDS1}$$

Randomized Smoothing Framework (RSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences, $\{\tau_k\}_{k\geq 0}$: interpolation sequence; M_1, M_2: randomized subroutines.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

- Use M_1 to find $\tilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that

$$\mathbb{E}\big[\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \tilde{\lambda}_{\rho_k,\eta_k}(x^k)) \,\big|\, \mathcal{F}_{k,0}\big] \leq \eta_k \quad \text{a.s.} \tag{rDS1}$$

- $\hat{\lambda}^k := \tau_k \lambda^k + (1-\tau_k)\tilde{\lambda}_{\rho_k,\eta_k}(x^k)$.

Randomized Smoothing Framework (RSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences, $\{\tau_k\}_{k\geq 0}$: interpolation sequence; M_1, M_2: randomized subroutines.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

- Use M_1 to find $\tilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that

$$\mathbb{E}\big[\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \tilde{\lambda}_{\rho_k,\eta_k}(x^k)) \,\big|\, \mathcal{F}_{k,0}\big] \leq \eta_k \quad \text{a.s.} \tag{rDS1}$$

- $\hat{\lambda}^k := \tau_k \lambda^k + (1-\tau_k)\tilde{\lambda}_{\rho_k,\eta_k}(x^k)$.
- Use M_2 to find $\tilde{x}_{\gamma_k}(\hat{\lambda}^k) \in \mathcal{X}$ such that

$$\mathbb{E}\big[S(\tilde{x}_{\gamma_k}(\hat{\lambda}^k), \hat{\lambda}^k) - \psi^{\mathrm{D}}(\hat{\lambda}^k) \,\big|\, \mathcal{F}_{k,1}\big] \leq \gamma_k \quad \text{a.s.} \tag{rPS}$$

Randomized Smoothing Framework (RSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences, $\{\tau_k\}_{k\geq 0}$: interpolation sequence; M_1, M_2: randomized subroutines.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

- Use M_1 to find $\tilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that

$$\mathbb{E}\left[\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \tilde{\lambda}_{\rho_k,\eta_k}(x^k)) \,\middle|\, \mathcal{F}_{k,0}\right] \leq \eta_k \quad \text{a.s.} \tag{rDS1}$$

- $\hat{\lambda}^k {:=} \tau_k \lambda^k + (1-\tau_k)\tilde{\lambda}_{\rho_k,\eta_k}(x^k)$.
- Use M_2 to find $\tilde{x}_{\gamma_k}(\hat{\lambda}^k) \in \mathcal{X}$ such that

$$\mathbb{E}\left[S(\tilde{x}_{\gamma_k}(\hat{\lambda}^k), \hat{\lambda}^k) - \psi^{\mathrm{D}}(\hat{\lambda}^k) \,\middle|\, \mathcal{F}_{k,1}\right] \leq \gamma_k \quad \text{a.s.} \tag{rPS}$$

- $x^{k+1} = \tau_k x^k + (1-\tau_k)\tilde{x}_{\gamma_k}(\hat{\lambda}^k)$, $\rho_{k+1} = \tau_k \rho_k$.

Randomized Smoothing Framework (RSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences, $\{\tau_k\}_{k\geq 0}$: interpolation sequence; M_1, M_2: randomized subroutines.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

- Use M_1 to find $\tilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that

$$\mathbb{E}\big[\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \tilde{\lambda}_{\rho_k,\eta_k}(x^k)) \,\big|\, \mathcal{F}_{k,0}\big] \leq \eta_k \quad \text{a.s.} \tag{rDS1}$$

- $\hat{\lambda}^k := \tau_k \lambda^k + (1-\tau_k)\tilde{\lambda}_{\rho_k,\eta_k}(x^k)$.
- Use M_2 to find $\tilde{x}_{\gamma_k}(\hat{\lambda}^k) \in \mathcal{X}$ such that

$$\mathbb{E}\big[S(\tilde{x}_{\gamma_k}(\hat{\lambda}^k), \hat{\lambda}^k) - \psi^{\mathrm{D}}(\hat{\lambda}^k) \,\big|\, \mathcal{F}_{k,1}\big] \leq \gamma_k \quad \text{a.s.} \tag{rPS}$$

- $x^{k+1} = \tau_k x^k + (1-\tau_k)\tilde{x}_{\gamma_k}(\hat{\lambda}^k)$, $\rho_{k+1} = \tau_k \rho_k$.
- Use M_1 to find $\tilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1}) \in \Lambda$ such that

$$\mathbb{E}\big[\psi^{\mathrm{P}}_{\rho_{k+1}}(x^{k+1}) - S_{\rho_{k+1}}(x^{k+1}, \tilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})) \,\big|\, \mathcal{F}_{k,2}\big] \leq \eta_k \quad \text{a.s.} \tag{rDS2}$$

Randomized Smoothing Framework (RSF)

Input: ρ_0: smoothing parameter; $\{\eta_k\}_{k\geq 0}$, $\{\gamma_k\}_{k\geq 0}$: error sequences, $\{\tau_k\}_{k\geq 0}$: interpolation sequence; M_1, M_2: randomized subroutines.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$

Repeat (until some convergence criterion is met)

- Use M_1 to find $\tilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that

$$\mathbb{E}\big[\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \tilde{\lambda}_{\rho_k,\eta_k}(x^k)) \,\big|\, \mathcal{F}_{k,0}\big] \leq \eta_k \quad \text{a.s.} \tag{rDS1}$$

- $\hat{\lambda}^k := \tau_k \lambda^k + (1-\tau_k)\tilde{\lambda}_{\rho_k,\eta_k}(x^k)$.
- Use M_2 to find $\tilde{x}_{\gamma_k}(\hat{\lambda}^k) \in \mathcal{X}$ such that

$$\mathbb{E}\big[S(\tilde{x}_{\gamma_k}(\hat{\lambda}^k), \hat{\lambda}^k) - \psi^{\mathrm{D}}(\hat{\lambda}^k) \,\big|\, \mathcal{F}_{k,1}\big] \leq \gamma_k \quad \text{a.s.} \tag{rPS}$$

- $x^{k+1} = \tau_k x^k + (1-\tau_k)\tilde{x}_{\gamma_k}(\hat{\lambda}^k)$, $\rho_{k+1} = \tau_k \rho_k$.
- Use M_1 to find $\tilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1}) \in \Lambda$ such that

$$\mathbb{E}\big[\psi^{\mathrm{P}}_{\rho_{k+1}}(x^{k+1}) - S_{\rho_{k+1}}(x^{k+1}, \tilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})) \,\big|\, \mathcal{F}_{k,2}\big] \leq \eta_k \quad \text{a.s.} \tag{rDS2}$$

- $\lambda^{k+1} := \tau_k \lambda^k + (1-\tau_k)\tilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})$, $k := k+1$.

Solving Subproblems Inexactly

$$\min_{x\in X}\left\{P(x,\lambda):=f(x)+g(x)+\Phi(x,\lambda)\right\},\quad \Phi(x,\lambda)=\frac{1}{n}\sum_{i=1}^{n}\Phi_i(x,\lambda)$$

Solving Subproblems Inexactly

$$\min_{x \in X} \{P(x,\lambda) := f(x) + g(x) + \Phi(x,\lambda)\}, \quad \Phi(x,\lambda) = \frac{1}{n}\sum_{i=1}^{n} \Phi_i(x,\lambda)$$

▷ Recall $\kappa_{\mathcal{X}} := (L + L_{xx})/\mu$. Use optimal randomized first-order solver, e.g., RPDG in [Lan & Zhou'18], we have

$$N = \Omega\left((n + \sqrt{n\kappa_{\mathcal{X}}}) \log(1/\epsilon)\right) \implies \mathbb{E}[P(\tilde{x}^N, \lambda) - P^*(\lambda)] \le \epsilon.$$

Outer Iteration Complexity

Theorem 4 (Outer Iteration Complexity of RSF)

If we choose $\rho_0 = 8L_{\mathrm{D}}$ $(L_{\mathrm{D}} = L_{\lambda\lambda} + 2L_{\lambda x}^2/\mu)$ *and for any* $k \in \mathbb{Z}_+$,

$$\tau_k = \frac{k+1}{k+3}, \quad \gamma_k = \frac{\varepsilon}{4(k+3)} \quad \textit{and} \quad \eta_k = \frac{\varepsilon}{4(k+3)}, \tag{3}$$

then for any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \Lambda$ *and* $K \in \mathbb{N}$,

$$\mathbb{E}[\Delta(x^K, \lambda^K)] \leq \frac{32L_{\mathrm{D}}D_{\Lambda}^2 + 2\Delta(x^0, \lambda^0)}{(K+1)(K+2)} + \frac{\varepsilon}{2}. \tag{4}$$

Outer Iteration Complexity

Theorem 4 (Outer Iteration Complexity of RSF)

If we choose $\rho_0 = 8L_{\mathrm{D}}$ $(L_{\mathrm{D}} = L_{\lambda\lambda} + 2L_{\lambda x}^2/\mu)$ *and for any* $k \in \mathbb{Z}_+$,

$$\tau_k = \frac{k+1}{k+3}, \quad \gamma_k = \frac{\varepsilon}{4(k+3)} \quad \text{and} \quad \eta_k = \frac{\varepsilon}{4(k+3)}, \tag{3}$$

then for any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \Lambda$ *and* $K \in \mathbb{N}$,

$$\mathbb{E}[\Delta(x^K, \lambda^K)] \leq \frac{32L_{\mathrm{D}}D_\Lambda^2 + 2\Delta(x^0, \lambda^0)}{(K+1)(K+2)} + \frac{\varepsilon}{2}. \tag{4}$$

Thus, to achieve an ε-*expected duality gap, the outer iteration complexity is* $O(\sqrt{L_{\mathrm{D}}/\varepsilon}) = O(\sqrt{L_{\lambda\lambda}/\varepsilon} + L_{\lambda x}/\sqrt{\mu\varepsilon})$.

Inner Iteration Complexity (Oracle Complexity)

Theorem 5 (Oracle complexity of RSF)

For any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \Lambda$*, let* $C^{\mathrm{P}}_{\mathsf{stoc}}$ *and* $C^{\mathrm{D}}_{\mathsf{stoc}}$ *denote the primal and dual oracle complexities to achieve an* ε*-expected duality gap, respectively. Then we have*

$$C^{\mathrm{P}}_{\mathsf{stoc}} = O\left((n + \sqrt{n\kappa_{\mathcal{X}}})\sqrt{\frac{L_{\mathrm{D}}}{\varepsilon}}\log\left(\frac{\kappa_{\mathcal{X}} L_{\mathrm{D}}(n + \sqrt{n\kappa_{\mathcal{X}}})}{\varepsilon}\right)\right),$$

$$C^{\mathrm{D}}_{\mathsf{stoc}} = O\left(\left(n\sqrt{\frac{L_{\mathrm{D}}}{\varepsilon}} + \frac{\sqrt{nL_{\lambda\lambda}L_{\mathrm{D}}}}{\varepsilon}\right)\log\left(\frac{L_{\lambda\lambda}(n + \sqrt{nL_{\lambda\lambda}/L_{\mathrm{D}}})}{\varepsilon}\right)\right).$$

Comparison of Oracle Complexities

Figure 1: Each $\Phi_i(x, \cdot)$ is concave (not necessarily linear).

Algorithms	Primal Oracle Comp.	Dual Oracle Comp.
PDHG-type [HA18]	$O(n/\varepsilon)$	$O(n/\varepsilon)$
Mirror-Prox [Nem05]	$O(n/\varepsilon)$	$O(n/\varepsilon)$
Det. IPDS	$\widetilde{O}(n\sqrt{\kappa_{\mathcal{X}}/\varepsilon})$	$\widetilde{O}(n/\varepsilon)$
Rand. IPDS	$\widetilde{O}((n+\sqrt{n\kappa_{\mathcal{X}}})/\sqrt{\varepsilon})$	$\widetilde{O}(n/\sqrt{\varepsilon}+\sqrt{n}/\varepsilon)$

Constrained Optimization Revisited

$$\min_{x \in \mathcal{X}} f(x) + r(x) \quad \text{s. t.} \quad g_i(x) \leq 0, \forall\, i \in [n]$$

- f is μ-strongly convex (s.c.) and L-smooth on $\mathcal{X}$.
- r is CCP with an easily computable proximal operator.
- For each $i \in [n]$, g_i is convex and α_i-smooth on $\mathcal{X}$.
- Slater condition holds $\Rightarrow$ no duality gap and an optimal primal-dual pair (x^*, λ^*) exists.

Constrained Optimization Revisited

$$\min_{x \in \mathcal{X}} f(x) + r(x) \quad \text{s. t.} \quad g_i(x) \leq 0, \, \forall \, i \in [n]$$

- f is μ-strongly convex (s.c.) and L-smooth on $\mathcal{X}$.
- r is CCP with an easily computable proximal operator.
- For each $i \in [n]$, g_i is convex and α_i-smooth on $\mathcal{X}$.
- Slater condition holds $\Rightarrow$ no duality gap and an optimal primal-dual pair (x^*, λ^*) exists.
- $\bar{x} \in \mathcal{X}$ is an ε-optimal and ε-feasible solution if

$$f(\bar{x}) - f(x^*) \leq \varepsilon, \quad \text{and} \quad [g_i(\bar{x})]_+ \leq \varepsilon, \, \forall \, i \in [n].$$

Lagrangian Form

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\mathbb{R}^n_+} \{S(x,\lambda) = f(x) + r(x) + (1/n)\textstyle\sum_{i=1}^n n\lambda_i g_i(x)\} \qquad \text{(Lag)}$$

Although $\Lambda = \mathbb{R}^n_+$ is unbounded, but

Lagrangian Form

$$\min_{x \in \mathcal{X}} \max_{\lambda \in \mathbb{R}^n_+} \left\{ S(x, \lambda) = f(x) + r(x) + (1/n) \textstyle\sum_{i=1}^n n \lambda_i g_i(x) \right\} \qquad \text{(Lag)}$$

Although $\Lambda = \mathbb{R}^n_+$ is unbounded, but

- The dual smoothing sub-problem has closed-form solution:

  $$\left([g_i(x)]_+/\rho\right)_{i=1}^n = \arg\max_{\lambda \in \mathbb{R}^n_+} \textstyle\sum_{i=1}^n \lambda_i g_i(x) - (\rho/2)\|\lambda\|_2^2$$

  $\Longrightarrow$ No need for first-order solver, and frameworks implementable.

Lagrangian Form

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\mathbb{R}^n_+} \{S(x,\lambda) = f(x) + r(x) + (1/n)\textstyle\sum_{i=1}^n n\lambda_i g_i(x)\} \qquad \text{(Lag)}$$

Although $\Lambda = \mathbb{R}^n_+$ is unbounded, but

- ▷ The dual smoothing sub-problem has closed-form solution:

  $$\big([g_i(x)]_+/\rho\big)_{i=1}^n = \arg\max_{\lambda\in\mathbb{R}^n_+} \textstyle\sum_{i=1}^n \lambda_i g_i(x) - (\rho/2)\|\lambda\|_2^2$$

  $\Longrightarrow$ No need for first-order solver, and frameworks implementable.

- ▷ Primal sub-optimality and constraint violation are used as convergence criteria, not duality gap.

Lagrangian Form

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\mathbb{R}^n_+} \{S(x,\lambda) = f(x) + r(x) + (1/n)\textstyle\sum_{i=1}^n n\lambda_i g_i(x)\} \qquad \text{(Lag)}$$

Although $\Lambda = \mathbb{R}^n_+$ is unbounded, but

- ▷ The dual smoothing sub-problem has closed-form solution:

  $$\left([g_i(x)]_+/\rho\right)_{i=1}^n = \arg\max_{\lambda\in\mathbb{R}^n_+} \textstyle\sum_{i=1}^n \lambda_i g_i(x) - (\rho/2)\|\lambda\|_2^2$$

  $\Longrightarrow$ No need for first-order solver, and frameworks implementable.

- ▷ Primal sub-optimality and constraint violation are used as convergence criteria, not duality gap.

- ▷ $L_{xx}(\lambda) = \sum_{i=1}^n \lambda_i \alpha_i$ is unbounded $\Longrightarrow$ Bound $\|\hat{\lambda}^k\|_\infty$ adaptively.

Convergence Rate of DSF for Constrained Opt.

Theorem 6 (Convergence Rate of DSF)

Let $(x^*, \lambda^*) \in \mathcal{X} \times \mathbb{R}^n_+$ *be a saddle point of* (Lag)*. If we apply DSF to solving* (Lag)*, then for any starting point* $(x^0, \lambda^0) \in \mathcal{X} \times \mathbb{R}^n_+$*,*

$$f(x^K) - f(x^*) \leq \frac{2[\Delta_{\rho_0}(x^0, \lambda^0)]_+}{(K+1)(K+2)} + \frac{\varepsilon}{2},$$

$$[g_i(x^K)]_+ \leq \frac{16\,(\lambda_i^* + \|\lambda^*\|_2)\,L_{\mathrm{D}} + 8\sqrt{L_{\mathrm{D}}[\Delta_{\rho_0}(x^0, \lambda^0)]_+}}{(K+1)(K+2)} + \frac{4\sqrt{L_{\mathrm{D}}\varepsilon}}{K+1},$$

for any $K \in \mathbb{N}$ *and* $i \in [m]$*.*

Oracle Complexity of DSF for Constrained Opt.

$$M := \sum_{i=1}^{n} \alpha_i D_{\mathcal{X}} + \inf_{x \in \mathcal{X}} \|\nabla g_i(x)\|_* \quad \text{and} \quad \alpha := \sum_{i=1}^{n} \alpha_i.$$

Oracle Complexity of DSF for Constrained Opt.

$$M := \sum_{i=1}^{n} \alpha_i D_{\mathcal{X}} + \inf_{x \in \mathcal{X}} \|\nabla g_i(x)\|_* \quad \text{and} \quad \alpha := \sum_{i=1}^{n} \alpha_i.$$

Lemma 7 (Bound on $\|\hat{\lambda}^k\|_\infty$)

If we apply DSF to (Lag)*, then for any* $k \in \mathbb{N}$,

$$\|\hat{\lambda}^k\|_\infty = O(1 + k\sqrt{\varepsilon\mu}/M).$$

Oracle Complexity of DSF for Constrained Opt.

$$M := \sum_{i=1}^n \alpha_i D_{\mathcal{X}} + \inf_{x\in\mathcal{X}} \|\nabla g_i(x)\|_* \quad \text{and} \quad \alpha := \sum_{i=1}^n \alpha_i.$$

Lemma 7 (Bound on $\|\hat{\lambda}^k\|_\infty$)

If we apply DSF to (Lag), then for any $k \in \mathbb{N}$,

$$\|\hat{\lambda}^k\|_\infty = O(1 + k\sqrt{\varepsilon\mu}/M).$$

Theorem 8 (Oracle Complexity of DSF)

For any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \mathbb{R}_+$, the oracle complexity of DSF to obtain an ε-optimal and ε-feasible solution is

$$O\left(\frac{nM}{\sqrt{\mu\varepsilon}}\sqrt{(L+\alpha)/\mu}\log\left(\frac{L+\alpha}{\varepsilon}\right)\right).$$

Convergence Rate of RSF for Constrained Opt.

Theorem 9 (Convergence Rate of RSF)

Let $(x^*, \lambda^*) \in \mathcal{X} \times \mathbb{R}^n_+$ *be a saddle point of* (Lag)*. If we apply RSF to solving* (Lag)*, then for any starting point* $(x^0, \lambda^0) \in \mathcal{X} \times \mathbb{R}^n_+$*,*

$$\mathbb{E}[f(x^K)] - f(x^*) \leq \frac{2[\Delta_{\rho_0}(x^0, \lambda^0)]_+}{(K+1)(K+2)} + \frac{\varepsilon}{2},$$

$$\mathbb{E}[[g_i(x^K)]_+] \leq \frac{16\,(\lambda_i^* + \|\lambda^*\|_2)\,L_{\mathrm{D}} + 8\sqrt{L_{\mathrm{D}}[\Delta_{\rho_0}(x^0, \lambda^0)]_+}}{(K+1)(K+2)} + \frac{4\sqrt{L_{\mathrm{D}}\varepsilon}}{K+1},$$

for any $K \in \mathbb{N}$ *and* $i \in [m]$*.*

Oracle Complexity of RSF for Constrained Opt.

Theorem 10 (Oracle Complexity of RSF)

For any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \mathbb{R}_+$*, the oracle complexity of RSF to obtain an* ε*-optimal and* ε*-feasible solution is*

$$O\left(\frac{\sqrt{n}M}{\sqrt{\mu\varepsilon}}\left(\sqrt{n} + \sqrt{(L+\alpha)/\mu}\right)\log\left(\frac{nM(L+\alpha)}{\mu\varepsilon}\right)\right).$$

Thank you!

References

[HA18] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. *A Primal-Dual Algorithm for General Convex-Concave Saddle Point Problems*. arXiv:1803.01401. 2018.

[Nem05] Arkadi Nemirovski. “Prox-Method with Rate of Convergence $O(1/t)$ for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-Concave Saddle Point Problems”. In: *SIAM J. Optim.* 15.1 (2005), pp. 229–251.