Optimal Stochastic Algorithms for Convex-Concave Saddle Point Problems

Renbo Zhao

Operations Research Center, Massachusetts Institute of Technology

Department of ISEM, NUS
Singapore, May 2019

Problem Statement

Consider the following convex-concave saddle point problem (SPP)

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\left[S(x,y)\triangleq f(x)+g(x)+\Phi(x,y)-J(y)\right], \qquad \texttt{(SPP)}$$

Problem Statement

Consider the following convex-concave saddle point problem (SPP)

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \left[S(x,y) \triangleq f(x) + g(x) + \Phi(x,y) - J(y)\right], \qquad \texttt{(SPP)}$$

- $\mathcal{X} \subseteq \mathbb{X}$ and $\mathcal{Y} \subseteq \mathbb{Y}$ are nonempty, closed and convex sets, where $\mathbb{X}$ and $\mathbb{Y}$ be two finite-dimensional real normed spaces.

Problem Statement

Consider the following convex-concave saddle point problem (SPP)

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \left[S(x,y) \triangleq f(x)+g(x)+\Phi(x,y)-J(y)\right], \qquad \texttt{(SPP)}$$

- $\mathcal{X}\subseteq\mathbb{X}$ and $\mathcal{Y}\subseteq\mathbb{Y}$ are nonempty, closed and convex sets, where $\mathbb{X}$ and $\mathbb{Y}$ be two finite-dimensional real normed spaces.
- $\mathbb{X}^*$ and $\mathbb{Y}^*$ are the dual spaces of $\mathbb{X}$ and $\mathbb{Y}$, respectively.

Problem Statement

Consider the following convex-concave saddle point problem (SPP)

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \big[S(x,y) \triangleq f(x) + g(x) + \Phi(x,y) - J(y)\big], \qquad \texttt{(SPP)}$$

- ▷ $\mathcal{X} \subseteq \mathbb{X}$ and $\mathcal{Y} \subseteq \mathbb{Y}$ are nonempty, closed and convex sets, where $\mathbb{X}$ and $\mathbb{Y}$ be two finite-dimensional real normed spaces.
- ▷ $\mathbb{X}^*$ and $\mathbb{Y}^*$ are the dual spaces of $\mathbb{X}$ and $\mathbb{Y}$, respectively.
- ▷ $f : \mathbb{X} \to \overline{\mathbb{R}}$, $g : \mathbb{X} \to \overline{\mathbb{R}}$ and $J : \mathbb{Y} \to \overline{\mathbb{R}}$ are convex, closed and proper (CCP) functions, where $\overline{\mathbb{R}} \triangleq (-\infty, +\infty]$.

Problem Statement

Consider the following convex-concave saddle point problem (SPP)

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \left[S(x,y) \triangleq f(x) + g(x) + \Phi(x,y) - J(y)\right], \qquad \texttt{(SPP)}$$

- ▷ $\mathcal{X} \subseteq \mathbb{X}$ and $\mathcal{Y} \subseteq \mathbb{Y}$ are nonempty, closed and convex sets, where $\mathbb{X}$ and $\mathbb{Y}$ be two finite-dimensional real normed spaces.
- ▷ $\mathbb{X}^*$ and $\mathbb{Y}^*$ are the dual spaces of $\mathbb{X}$ and $\mathbb{Y}$, respectively.
- ▷ $f : \mathbb{X} \to \overline{\mathbb{R}}$, $g : \mathbb{X} \to \overline{\mathbb{R}}$ and $J : \mathbb{Y} \to \overline{\mathbb{R}}$ are convex, closed and proper (CCP) functions, where $\overline{\mathbb{R}} \triangleq (-\infty, +\infty]$.
- ▷ $\Phi : \mathbb{X} \times \mathbb{Y} \to [-\infty, +\infty]$ is convex-concave, i.e., $\Phi(\cdot, y)$ is convex and $\Phi(x, \cdot)$ is concave, for any $(x, y) \in \mathbb{X} \times \mathbb{Y}$.

Regularity Assumptions

Regularity Assumptions

- f is μ-strong convex (s.c.) and L-smooth on $\mathcal{X}$ ($L \geq \mu \geq 0$), i.e.,

$$\frac{\mu}{2}\left\|x - x'\right\|_{\mathbb{X}}^2 \leq f(x) - f(x') - \langle \nabla f(x'), x - x' \rangle \leq \frac{L}{2}\left\|x - x'\right\|_{\mathbb{X}}^2, \forall\, x, x' \in \mathcal{X}.$$

Regularity Assumptions

- f is μ-strong convex (s.c.) and L-smooth on $\mathcal{X}$ ($L \ge \mu \ge 0$), i.e.,

$$\frac{\mu}{2}\left\|x - x'\right\|_{\mathbb{X}}^2 \le f(x) - f(x') - \langle \nabla f(x'), x - x' \rangle \le \frac{L}{2}\left\|x - x'\right\|_{\mathbb{X}}^2, \forall\, x, x' \in \mathcal{X}.$$

- Both cases $\mu = 0$ and $\mu > 0$ will be considered.

Regularity Assumptions

- f is μ-strong convex (s.c.) and L-smooth on $\mathcal{X}$ ($L \geq \mu \geq 0$), i.e.,

$$\frac{\mu}{2}\left\|x-x'\right\|_{\mathbb{X}}^2 \leq f(x)-f(x')-\langle \nabla f(x'), x-x'\rangle \leq \frac{L}{2}\left\|x-x'\right\|_{\mathbb{X}}^2, \forall\, x, x' \in \mathcal{X}.$$

- Both cases $\mu = 0$ and $\mu > 0$ will be considered.
- g and J admit tractable *Bregman proximal projections* on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Also, $\mathsf{dom}\, g \cap \mathcal{X} \neq \emptyset$ and $\mathsf{dom}\, J \cap \mathcal{Y} \neq \emptyset$.

Regularity Assumptions

- ▷ f is μ-strong convex (s.c.) and L-smooth on $\mathcal{X}$ ($L \geq \mu \geq 0$), i.e.,

$$\frac{\mu}{2}\left\|x - x'\right\|_{\mathbb{X}}^2 \leq f(x) - f(x') - \langle \nabla f(x'), x - x' \rangle \leq \frac{L}{2}\left\|x - x'\right\|_{\mathbb{X}}^2, \forall\, x, x' \in \mathcal{X}.$$

- ▷ Both cases $\mu = 0$ and $\mu > 0$ will be considered.
- ▷ g and J admit tractable *Bregman proximal projections* on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Also, $\mathsf{dom}\, g \cap \mathcal{X} \neq \emptyset$ and $\mathsf{dom}\, J \cap \mathcal{Y} \neq \emptyset$.
- ▷ Φ is (L_{xx}, L_{yx}, L_{yy})-smooth on $\mathcal{X} \times \mathcal{Y}$, i.e.,

$$\left\|\nabla_x \Phi(x, y) - \nabla_x \Phi(x', y)\right\|_{\mathbb{X}^*} \leq L_{xx}\left\|x - x'\right\|_{\mathbb{X}}, \tag{1a}$$
$$\left\|\nabla_x \Phi(x, y) - \nabla_x \Phi(x, y')\right\|_{\mathbb{X}^*} \leq L_{yx}\left\|y - y'\right\|_{\mathbb{Y}}, \tag{1b}$$
$$\left\|\nabla_y \Phi(x, y) - \nabla_y \Phi(x', y)\right\|_{\mathbb{Y}^*} \leq L_{yx}\left\|x - x'\right\|_{\mathbb{X}}, \tag{1c}$$
$$\left\|\nabla_y \Phi(x, y) - \nabla_y \Phi(x, y')\right\|_{\mathbb{Y}^*} \leq L_{yy}\left\|y - y'\right\|_{\mathbb{Y}}. \tag{1d}$$

Regularity Assumptions

- ▷ f is μ-strong convex (s.c.) and L-smooth on $\mathcal{X}$ ($L \geq \mu \geq 0$), i.e.,

$$\frac{\mu}{2}\left\|x - x'\right\|_{\mathbb{X}}^2 \leq f(x) - f(x') - \langle \nabla f(x'), x - x' \rangle \leq \frac{L}{2}\left\|x - x'\right\|_{\mathbb{X}}^2, \forall\, x, x' \in \mathcal{X}.$$

- ▷ Both cases $\mu = 0$ and $\mu > 0$ will be considered.
- ▷ g and J admit tractable *Bregman proximal projections* on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Also, $\mathsf{dom}\, g \cap \mathcal{X} \neq \emptyset$ and $\mathsf{dom}\, J \cap \mathcal{Y} \neq \emptyset$.
- ▷ Φ is (L_{xx}, L_{yx}, L_{yy})-smooth on $\mathcal{X} \times \mathcal{Y}$, i.e.,

$$\left\|\nabla_x \Phi(x, y) - \nabla_x \Phi(x', y)\right\|_{\mathbb{X}^*} \leq L_{xx}\left\|x - x'\right\|_{\mathbb{X}}, \tag{1a}$$
$$\left\|\nabla_x \Phi(x, y) - \nabla_x \Phi(x, y')\right\|_{\mathbb{X}^*} \leq L_{yx}\left\|y - y'\right\|_{\mathbb{Y}}, \tag{1b}$$
$$\left\|\nabla_y \Phi(x, y) - \nabla_y \Phi(x', y)\right\|_{\mathbb{Y}^*} \leq L_{yx}\left\|x - x'\right\|_{\mathbb{X}}, \tag{1c}$$
$$\left\|\nabla_y \Phi(x, y) - \nabla_y \Phi(x, y')\right\|_{\mathbb{Y}^*} \leq L_{yy}\left\|y - y'\right\|_{\mathbb{Y}}. \tag{1d}$$

- ▷ A saddle point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ exists for (SPP), i.e.,

$$S(x^*, y) \leq S(x^*, y^*) \leq S(x, y^*), \quad \forall\, (x, y) \in \mathcal{X} \times \mathcal{Y}.$$

Applications

Applications

- Any bilinear SPP, i.e., $\Phi(x, y) = \langle \mathsf{A}x, y\rangle$, $\mathsf{A} \in \mathcal{L}(\mathbb{X}, \mathbb{Y}^*)$

Applications

- ▷ Any bilinear SPP, i.e., $\Phi(x, y) = \langle \mathsf{A}x, y \rangle$, $\mathsf{A} \in \mathcal{L}(\mathbb{X}, \mathbb{Y}^*)$
- ▷ Non-bilinear SPP

Applications

- Any bilinear SPP, i.e., $\Phi(x, y) = \langle \mathsf{A}x, y \rangle$, $\mathsf{A} \in \mathcal{L}(\mathbb{X}, \mathbb{Y}^*)$
- Non-bilinear SPP
  - Convex-concave game

Applications

- Any bilinear SPP, i.e., $\Phi(x, y) = \langle \mathsf{A}x, y \rangle$, $\mathsf{A} \in \mathcal{L}(\mathbb{X}, \mathbb{Y}^*)$
- Non-bilinear SPP
    - Convex-concave game
    - Convex Optimization with Functional Constraints

Applications

- Any bilinear SPP, i.e., $\Phi(x, y) = \langle \mathsf{A}x, y \rangle$, $\mathsf{A} \in \mathcal{L}(\mathbb{X}, \mathbb{Y}^*)$
- Non-bilinear SPP
  - Convex-concave game
  - Convex Optimization with Functional Constraints
  - Kernel Matrix Learning

Stochastic First-Order Oracles

$$f(x) \triangleq \mathbb{E}_\xi[\tilde{f}(x,\xi)] \qquad \Phi(x,y) \triangleq \mathbb{E}_\zeta[\widetilde{\Phi}(x,y,\zeta)]$$

Stochastic First-Order Oracles

$$f(x) \triangleq \mathbb{E}_\xi[\tilde{f}(x,\xi)] \qquad \Phi(x,y) \triangleq \mathbb{E}_\zeta[\widetilde{\Phi}(x,y,\zeta)]$$

Oracle model (Stochastic approximation):
Return estimators of ∇f, $\nabla\Phi(\cdot,y)$ and $\nabla\Phi(x,\cdot)$, i.e., $\hat{\nabla} f$, $\hat{\nabla}\Phi(\cdot,y)$ and $\hat{\nabla}\Phi(x,\cdot)$, that

- are unbiased
- have bounded variances
- (may also) obey "light-tailed" distributions

Gradient Noise	Mean	Variance
$\delta_{x,f} \triangleq \hat{\nabla} f - \nabla f$	0	$\sigma_{x,f}^2$
$\delta_{x,\Phi} \triangleq \hat{\nabla}_x \Phi(\cdot,y) - \nabla_x \Phi(\cdot,y)$	0	$\sigma_{x,\Phi}^2$
$\delta_{y,\Phi} \triangleq \hat{\nabla}_y \Phi(x,\cdot) - \nabla_y \Phi(x,\cdot)$	0	$\sigma_{y,\Phi}^2$

Stochastic First-Order Oracles

$$f(x) \triangleq \mathbb{E}_\xi[\tilde{f}(x,\xi)] \qquad \Phi(x,y) \triangleq \mathbb{E}_\zeta[\widetilde{\Phi}(x,y,\zeta)]$$

Oracle model (Stochastic approximation):
Return estimators of ∇f, $\nabla\Phi(\cdot,y)$ and $\nabla\Phi(x,\cdot)$, i.e., $\hat{\nabla} f$, $\hat{\nabla}\Phi(\cdot,y)$ and $\hat{\nabla}\Phi(x,\cdot)$, that

- are unbiased
- have bounded variances
- (may also) obey "light-tailed" distributions

Gradient Noise	Mean	Variance
$\delta_{x,f} \triangleq \hat{\nabla} f - \nabla f$	0	$\sigma_{x,f}^2$
$\delta_{x,\Phi} \triangleq \hat{\nabla}_x\Phi(\cdot,y) - \nabla_x\Phi(\cdot,y)$	0	$\sigma_{x,\Phi}^2$
$\delta_{y,\Phi} \triangleq \hat{\nabla}_y\Phi(x,\cdot) - \nabla_y\Phi(x,\cdot)$	0	$\sigma_{y,\Phi}^2$

- (SPP) $\to$ SPP(L, L_{xx}, L_{yx}, L_{yy}, σ, μ), where $\sigma \triangleq \sigma_{x,f} + \sigma_{x,\Phi} + \sigma_{y,\Phi}$.

Main Contribution ($\mu > 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, \mu)}$$

Main Contribution ($\mu > 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, \mu)}$$

- Develop the *first stochastic* restart scheme for SPP.

Main Contribution ($\mu > 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, \mu)}$$

- Develop the *first stochastic* restart scheme for SPP.
- Consider the *sub-Gaussian* gradient noises.

Main Contribution ($\mu > 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, \mu)}$$

- Develop the *first stochastic* restart scheme for SPP.
- Consider the *sub-Gaussian* gradient noises.
- To obtain an ϵ-duality gap w.p. $\geq 1 - \nu$, the oracle complexity is

$$O\left(\left(\sqrt{\frac{L}{\mu}} + \frac{L_{xx}}{\mu}\right)\log\left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}} + \frac{L_{yy}}{\epsilon} + \left(\frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2}{\mu\epsilon} + \frac{\sigma_{y,\Phi}^2}{\epsilon^2}\right)\log\left(\frac{\log(1/\epsilon)}{\nu}\right)\right).$$

Main Contribution ($\mu > 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, \mu)}$$

Main Contribution ($\mu > 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, \mu)}$$

- Furthermore, assume that $\mathsf{dom}\, g$ and $\mathsf{dom}\, J$ are closed.

Main Contribution ($\mu > 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, \mu)}$$

- Furthermore, assume that $\mathsf{dom}\, g$ and $\mathsf{dom}\, J$ are closed.
- Then to obtain an ϵ-expected duality gap, the oracle complexity is

$$O\left(\left(\sqrt{\frac{L}{\mu}} + \frac{L_{xx}}{\mu}\right)\log\left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}} + \frac{L_{yy}}{\epsilon} + \left(\frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2}{\mu\epsilon} + \frac{\sigma_{y,\Phi}^2}{\epsilon^2}\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

Main Contribution ($\mu > 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, \mu)}$$

- Furthermore, assume that $\mathsf{dom}\, g$ and $\mathsf{dom}\, J$ are closed.
- Then to obtain an ϵ-expected duality gap, the oracle complexity is

$$O\left(\left(\sqrt{\frac{L}{\mu}} + \frac{L_{xx}}{\mu}\right)\log\left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}} + \frac{L_{yy}}{\epsilon} + \left(\frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2}{\mu\epsilon} + \frac{\sigma_{y,\Phi}^2}{\epsilon^2}\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

  - The complexities of L and L_{yx} are optimal.

Main Contribution ($\mu > 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, \mu)}$$

- Furthermore, assume that $\mathsf{dom}\, g$ and $\mathsf{dom}\, J$ are closed.
- Then to obtain an ϵ-expected duality gap, the oracle complexity is

$$O\left(\left(\sqrt{\frac{L}{\mu}} + \frac{L_{xx}}{\mu}\right)\log\left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}} + \frac{L_{yy}}{\epsilon} + \left(\frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2}{\mu\epsilon} + \frac{\sigma_{y,\Phi}^2}{\epsilon^2}\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

  - The complexities of L and L_{yx} are optimal.
  - The complexities of $\sigma_{x,f}$, $\sigma_{x,\Phi}$ and $\sigma_{y,\Phi}$ are optimal up to a log factor, but still the best-known.

Main Contribution ($\mu > 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, \mu)}$$

- Furthermore, assume that $\mathsf{dom}\, g$ and $\mathsf{dom}\, J$ are closed.
- Then to obtain an ϵ-expected duality gap, the oracle complexity is

$$O\left(\left(\sqrt{\frac{L}{\mu}} + \frac{L_{xx}}{\mu}\right)\log\left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}} + \frac{L_{yy}}{\epsilon} + \left(\frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2}{\mu\epsilon} + \frac{\sigma_{y,\Phi}^2}{\epsilon^2}\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

  - The complexities of L and L_{yx} are optimal.
  - The complexities of $\sigma_{x,f}$, $\sigma_{x,\Phi}$ and $\sigma_{y,\Phi}$ are optimal up to a log factor, but still the best-known.
  - The complexities of L_{xx} and L_{yy} are the best-known. (Lower bound? Acceleration?)

Comparison with Other Methods

Algorithm	Problem Class	Oracle Complexity
PDHG-type [Hamedani & Aybat'18]	$\sigma = 0, L_{yy} = 0$	$O\left(\frac{L+L_{xx}+L_{yx}}{\sqrt{\mu\epsilon}}\right)$
Mirror-Prox-B [Juditsky & Nemirovski'12]	$\sigma = 0, L_{yy} = 0$	$O\left(\frac{L+L_{xx}}{\mu}\log\left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}}\right)$

Comparison with Other Methods

Algorithm	Problem Class	Oracle Complexity
PDHG-type [Hamedani & Aybat'18]	$\sigma = 0, L_{yy} = 0$	$O\left(\frac{L+L_{xx}+L_{yx}}{\sqrt{\mu\epsilon}}\right)$
Mirror-Prox-B [Juditsky & Nemirovski'12]	$\sigma = 0, L_{yy} = 0$	$O\left(\frac{L+L_{xx}}{\mu}\log\left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}}\right)$

▷ The oracle complexity of Algorithm 2 is

$$O\left(\left(\sqrt{\frac{L}{\mu}} + \frac{L_{xx}}{\mu}\right)\log\left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}} + \frac{L_{yy}}{\epsilon} + \left(\frac{(\sigma_{x,f}+\sigma_{x,\Phi})^2}{\mu\epsilon} + \frac{\sigma_{y,\Phi}^2}{\epsilon^2}\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

Comparison with Other Methods

Algorithm	Problem Class	Oracle Complexity
PDHG-type [Hamedani & Aybat'18]	$\sigma = 0, L_{yy} = 0$	$O\left(\frac{L+L_{xx}+L_{yx}}{\sqrt{\mu\epsilon}}\right)$
Mirror-Prox-B [Juditsky & Nemirovski'12]	$\sigma = 0, L_{yy} = 0$	$O\left(\frac{L+L_{xx}}{\mu}\log\left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}}\right)$

- The oracle complexity of Algorithm 2 is

$$O\left(\left(\sqrt{\frac{L}{\mu}} + \frac{L_{xx}}{\mu}\right)\log\left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}} + \frac{L_{yy}}{\epsilon} + \left(\frac{(\sigma_{x,f}+\sigma_{x,\Phi})^2}{\mu\epsilon} + \frac{\sigma_{y,\Phi}^2}{\epsilon^2}\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

  - For $\sigma = 0$ and $L_{yy} = 0$, strictly better than the previous methods.

Comparison with Other Methods

Algorithm	Problem Class	Oracle Complexity
PDHG-type [Hamedani & Aybat'18]	$\sigma = 0, L_{yy} = 0$	$O\left(\frac{L+L_{xx}+L_{yx}}{\sqrt{\mu\epsilon}}\right)$
Mirror-Prox-B [Juditsky & Nemirovski'12]	$\sigma = 0, L_{yy} = 0$	$O\left(\frac{L+L_{xx}}{\mu}\log\left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}}\right)$

▷ The oracle complexity of Algorithm 2 is

$$O\left(\left(\sqrt{\frac{L}{\mu}} + \frac{L_{xx}}{\mu}\right)\log\left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}} + \frac{L_{yy}}{\epsilon} + \left(\frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2}{\mu\epsilon} + \frac{\sigma_{y,\Phi}^2}{\epsilon^2}\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

- For $\sigma = 0$ and $L_{yy} = 0$, strictly better than the previous methods.
- For $\sigma > 0$ and $L_{yy} > 0$, the first complexity result.

Subroutine ($\mu = 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, 0)}$$

Subroutine ($\mu = 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, 0)}$$

- Extend the primal-dual hybrid gradient (PDHG) framework to the *non-bilinear stochastic* SPP.

Subroutine ($\mu = 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, 0)}$$

- Extend the primal-dual hybrid gradient (PDHG) framework to the *non-bilinear stochastic* SPP.
- To obtain an ϵ-expected duality gap, the oracle complexity is

$$O\left(\sqrt{\frac{L}{\epsilon}} + \frac{L_{xx} + L_{yx} + L_{yy}}{\epsilon} + \frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2 + \sigma_{y,\Phi}^2}{\epsilon^2}\right).$$

Subroutine ($\mu = 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, 0)}$$

- ▷ Extend the primal-dual hybrid gradient (PDHG) framework to the *non-bilinear stochastic* SPP.
- ▷ To obtain an ϵ-expected duality gap, the oracle complexity is

$$O\left(\sqrt{\frac{L}{\epsilon}} + \frac{L_{xx} + L_{yx} + L_{yy}}{\epsilon} + \frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2 + \sigma_{y,\Phi}^2}{\epsilon^2}\right).$$

  - The complexities of L, L_{yx}, $\sigma_{x,f}$, $\sigma_{x,\Phi}$ and $\sigma_{y,\Phi}$ are optimal.

Subroutine ($\mu = 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, 0)}$$

- Extend the primal-dual hybrid gradient (PDHG) framework to the *non-bilinear stochastic* SPP.
- To obtain an ϵ-expected duality gap, the oracle complexity is

$$O\left(\sqrt{\frac{L}{\epsilon}} + \frac{L_{xx} + L_{yx} + L_{yy}}{\epsilon} + \frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2 + \sigma_{y,\Phi}^2}{\epsilon^2}\right).$$

  - The complexities of L, L_{yx}, $\sigma_{x,f}$, $\sigma_{x,\Phi}$ and $\sigma_{y,\Phi}$ are optimal.
  - The complexities of L_{xx} and L_{yy} are the best-known. (Lower bound? Acceleration?)

Subroutine ($\mu = 0$)

$$\boxed{\mathrm{SPP}(L,\, L_{xx},\, L_{yx},\, L_{yy},\, \sigma,\, 0)}$$

- Extend the primal-dual hybrid gradient (PDHG) framework to the *non-bilinear stochastic* SPP.
- To obtain an ϵ-expected duality gap, the oracle complexity is

$$O\left(\sqrt{\frac{L}{\epsilon}} + \frac{L_{xx} + L_{yx} + L_{yy}}{\epsilon} + \frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2 + \sigma_{y,\Phi}^2}{\epsilon^2}\right).$$

  - The complexities of L, L_{yx}, $\sigma_{x,f}$, $\sigma_{x,\Phi}$ and $\sigma_{y,\Phi}$ are optimal.
  - The complexities of L_{xx} and L_{yy} are the best-known. (Lower bound? Acceleration?)

- If the gradient noises are sub-Gaussian, to obtain an ϵ-duality gap w.p. at least $1 - \nu$, the oracle complexity is

$$O\left(\sqrt{\frac{L}{\epsilon}} + \frac{L_{xx} + L_{yx} + L_{yy}}{\epsilon} + \frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2 + \sigma_{y,\Phi}^2}{\epsilon^2} \log\left(\frac{1}{\nu}\right)\right).$$

Comparison with Other Methods

Algorithm	Prob. Class	Oracle Complexity
PDHG-type [Hamedani & Aybat'18]	$\sigma = 0$	$O\left(\frac{L}{\epsilon} + \frac{L_{xx}+L_{yx}+L_{yy}}{\epsilon}\right)$
Mirror-Prox [Nemirovski'05]	$\sigma = 0$	$O\left(\frac{L}{\epsilon} + \frac{L_{xx}+L_{yx}+L_{yy}}{\epsilon}\right)$
Stoc. MP [Juditsky et al.'11]	$\sigma > 0$	$O\left(\frac{L}{\epsilon} + \frac{L_{xx}+L_{yx}+L_{yy}}{\epsilon} + \frac{(\sigma_{x,f}+\sigma_{x,\Phi})^2+\sigma_{y,\Phi}^2}{\epsilon^2}\right)$
Stoc. Acc. MP [Chen et al.'17]	$\sigma > 0$	$O\left(\sqrt{\frac{L}{\epsilon}} + \frac{L_{xx}+L_{yx}+L_{yy}}{\epsilon} + \frac{(\sigma_{x,f}+\sigma_{x,\Phi})^2+\sigma_{y,\Phi}^2}{\epsilon^2}\right)$
Algorithm 1 [Zhao'19]	$\sigma > 0$	$O\left(\sqrt{\frac{L}{\epsilon}} + \frac{L_{xx}+L_{yx}+L_{yy}}{\epsilon} + \frac{(\sigma_{x,f}+\sigma_{x,\Phi})^2+\sigma_{y,\Phi}^2}{\epsilon^2}\right)$

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 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 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 $D_{h_{\mathcal{U}}}(u, u') \geq (1/2) \|u - u'\|^2$, for any $u \in \mathcal{U}$ and $u' \in \mathcal{U}^o \triangleq \mathcal{U} \cap \mathsf{int\,dom}\, h_{\mathcal{U}}$.

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 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 $D_{h_{\mathcal{U}}}(u, u') \geq (1/2) \|u - u'\|^2$, for any $u \in \mathcal{U}$ and $u' \in \mathcal{U}^o \triangleq \mathcal{U} \cap \mathsf{int\,dom}\, h_{\mathcal{U}}$.

- Example: $\mathbb{U} = (\mathbb{R}^n, \|\cdot\|_1), \mathcal{U} = \Delta_n \triangleq \{u \in \mathbb{R}^n_+ : \sum_{i=1}^n u_i = 1\}$, $h_{\mathcal{U}} = \sum_{i=1}^n u_i \log u_i$, $\mathsf{dom}\, h_{\mathcal{U}} = \mathbb{R}^n_+$, $\mathcal{U}^o = \mathsf{ri}\, \Delta_n$.

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.

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

Bregman Proximal Projection (BPP)

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

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

- If $\inf_{u\in\mathcal{U}} \varphi(u) > -\infty$ and $\mathcal{U} \cap \mathsf{dom}\, \phi \cap \mathsf{dom}\, h_{\mathcal{U}} \neq \emptyset$, then u^+ is unique and lies in $\mathcal{U}^o \cap \mathsf{dom}\, \varphi$.

Bregman Proximal Projection (BPP)

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

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

- If $\inf_{u \in \mathcal{U}} \varphi(u) > -\infty$ and $\mathcal{U} \cap \mathsf{dom}\, \phi \cap \mathsf{dom}\, h_{\mathcal{U}} \neq \emptyset$, then u^+ is unique and lies in $\mathcal{U}^o \cap \mathsf{dom}\, \varphi$.
- We say φ has a *tractable* BPP on $\mathcal{U}$ 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.

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

- If $\inf_{u \in \mathcal{U}} \varphi(u) > -\infty$ and $\mathcal{U} \cap \mathsf{dom}\, \phi \cap \mathsf{dom}\, h_{\mathcal{U}} \neq \emptyset$, then u^+ is unique and lies in $\mathcal{U}^o \cap \mathsf{dom}\, \varphi$.
- We say φ has a *tractable* BPP on $\mathcal{U}$ 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

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

Primal and Dual Functions

$$(\mathbb{P}): \ \min_{x\in\mathcal{X}} \left[\bar{S}(x) \triangleq \sup\nolimits_{y\in\mathcal{Y}} S(x,y)\right], \quad (\mathbb{D}): \ \max_{y\in\mathcal{Y}} \left[\underline{S}(x) \triangleq \inf\nolimits_{x\in\mathcal{X}} S(x,y)\right].$$

Primal and Dual Functions

$$(\mathbb{P}): \min_{x\in\mathcal{X}} \left[\bar{S}(x) \triangleq \sup\nolimits_{y\in\mathcal{Y}} S(x,y)\right], \quad (\mathbb{D}): \max_{y\in\mathcal{Y}} \left[\underline{S}(x) \triangleq \inf\nolimits_{x\in\mathcal{X}} S(x,y)\right].$$

▷ $\bar{S} \to$ primal function, $\underline{S} \to$ dual function

Primal and Dual Functions

$$(\mathbb{P}):\ \min_{x\in\mathcal{X}}\left[\bar{S}(x)\triangleq \sup\nolimits_{y\in\mathcal{Y}} S(x,y)\right],\quad (\mathbb{D}):\ \max_{y\in\mathcal{Y}}\left[\underline{S}(x)\triangleq \inf\nolimits_{x\in\mathcal{X}} S(x,y)\right].$$

- $\bar{S} \to$ primal function, $\underline{S} \to$ dual function
- Since (SPP) has a saddle point (x^*, y^*), x^* and y^* are solutions of $(\mathbb{P})$ and $(\mathbb{D})$ respectively and $\bar{S}(x^*) = S(x^*, y^*) = \underline{S}(y^*)$.

Primal and Dual Functions

$$(\mathbb{P}): \min_{x\in\mathcal{X}} \left[\bar{S}(x) \triangleq \sup_{y\in\mathcal{Y}} S(x,y)\right], \quad (\mathbb{D}): \max_{y\in\mathcal{Y}} \left[\underline{S}(x) \triangleq \inf_{x\in\mathcal{X}} S(x,y)\right].$$

- $\bar{S} \to$ primal function, $\underline{S} \to$ dual function
- Since (SPP) has a saddle point (x^*, y^*), x^* and y^* are solutions of $(\mathbb{P})$ and $(\mathbb{D})$ respectively and $\bar{S}(x^*) = S(x^*, y^*) = \underline{S}(y^*)$.
- Define the *duality gap*

$$G(x,y) \triangleq \bar{S}(x) - \underline{S}(y) = \sup_{x'\in\mathcal{X}, y'\in\mathcal{Y}} S(x,y') - S(x',y).$$

Algorithm 1: An Optimal Algorithm for $\mu = 0$

Algorithm 1: An Optimal Algorithm for $\mu = 0$

- **Input**: Interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$

Algorithm 1: An Optimal Algorithm for $\mu = 0$

- **Input**: Interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$, $t = 1$

Algorithm 1: An Optimal Algorithm for $\mu = 0$

- **Input**: Interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$, $t = 1$
- **Repeat** (until some convergence criterion is met)

$$y^{t+1} := \arg\min_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{h_{\mathcal{Y}}}(y, y^t) \qquad \text{(Dual Ascent)}$$

Algorithm 1: An Optimal Algorithm for $\mu = 0$

- **Input**: Interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$, $t = 1$
- **Repeat** (until some convergence criterion is met)

$$y^{t+1} := \arg\min_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{h_{\mathcal{Y}}}(y, y^t) \qquad \text{(Dual Ascent)}$$

$$\tilde{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^t \qquad \text{(Interpolation)}$$

Algorithm 1: An Optimal Algorithm for $\mu = 0$

- **Input**: Interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$, $t = 1$
- **Repeat** (until some convergence criterion is met)

$$y^{t+1} := \arg\min_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{h_{\mathcal{Y}}}(y, y^t) \qquad \text{(Dual Ascent)}$$

$$\tilde{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^t \qquad \text{(Interpolation)}$$

$$x^{t+1} := \arg\min_{x\in\mathcal{X}} g(x) + \langle \hat{\nabla}_x \Phi(x^t, y^{t+1}, \zeta_x^t) + \hat{\nabla} f(\tilde{x}^{t+1}, \xi^t), x - x^t\rangle + \tau_t^{-1} D_{h_{\mathcal{X}}}(x, x^t)) \qquad \text{(Primal Descent)}$$

Algorithm 1: An Optimal Algorithm for $\mu = 0$

- **Input**: Interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$, $t = 1$
- **Repeat** (until some convergence criterion is met)

$$
\begin{aligned}
y^{t+1} &:= \arg\min\nolimits_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{h_{\mathcal{Y}}}(y, y^t) && \text{(Dual Ascent)}\\
\tilde{x}^{t+1} &:= (1-\beta_t)\bar{x}^t + \beta_t x^t && \text{(Interpolation)}\\
x^{t+1} &:= \arg\min_{x\in\mathcal{X}} g(x) + \langle \hat{\nabla}_x \Phi(x^t, y^{t+1}, \zeta_x^t) + \hat{\nabla} f(\tilde{x}^{t+1}, \xi^t), x - x^t\rangle \\
&\qquad\qquad + \tau_t^{-1} D_{h_{\mathcal{X}}}(x, x^t)) && \text{(Primal Descent)}\\
s^{t+1} &:= (1+\theta_{t+1})\hat{\nabla}_y \Phi(x^{t+1}, y^{t+1}, \zeta_y^{t+1}) - \theta_{t+1}\hat{\nabla}_y \Phi(x^t, y^t, \zeta_y^t) && \text{(Extrap.)}
\end{aligned}
$$

Algorithm 1: An Optimal Algorithm for $\mu = 0$

- **Input**: Interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$, $t = 1$
- **Repeat** (until some convergence criterion is met)

$$y^{t+1} := \arg\min_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{h_{\mathcal{Y}}}(y, y^t) \qquad \text{(Dual Ascent)}$$

$$\tilde{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^t \qquad \text{(Interpolation)}$$

$$x^{t+1} := \arg\min_{x\in\mathcal{X}} g(x) + \langle \hat{\nabla}_x \Phi(x^t, y^{t+1}, \zeta_x^t) + \hat{\nabla} f(\tilde{x}^{t+1}, \xi^t), x - x^t\rangle + \tau_t^{-1} D_{h_{\mathcal{X}}}(x, x^t)) \qquad \text{(Primal Descent)}$$

$$s^{t+1} := (1+\theta_{t+1})\hat{\nabla}_y \Phi(x^{t+1}, y^{t+1}, \zeta_y^{t+1}) - \theta_{t+1}\hat{\nabla}_y \Phi(x^t, y^t, \zeta_y^t) \qquad \text{(Extrap.)}$$

$$\bar{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^{t+1} \qquad \text{(Primal Averaging)}$$

Algorithm 1: An Optimal Algorithm for $\mu = 0$

- **Input**: Interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$, $t = 1$
- **Repeat** (until some convergence criterion is met)

$$y^{t+1} := \arg\min_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{h_{\mathcal{Y}}}(y, y^t) \qquad \text{(Dual Ascent)}$$

$$\tilde{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^t \qquad \text{(Interpolation)}$$

$$x^{t+1} := \arg\min_{x\in\mathcal{X}} g(x) + \langle \hat{\nabla}_x \Phi(x^t, y^{t+1}, \zeta_x^t) + \hat{\nabla} f(\tilde{x}^{t+1}, \xi^t), x - x^t\rangle + \tau_t^{-1} D_{h_{\mathcal{X}}}(x, x^t)) \qquad \text{(Primal Descent)}$$

$$s^{t+1} := (1+\theta_{t+1})\hat{\nabla}_y \Phi(x^{t+1}, y^{t+1}, \zeta_y^{t+1}) - \theta_{t+1}\hat{\nabla}_y \Phi(x^t, y^t, \zeta_y^t) \qquad \text{(Extrap.)}$$

$$\bar{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^{t+1} \qquad \text{(Primal Averaging)}$$

$$\bar{y}^{t+1} := (1-\beta_t)\bar{y}^t + \beta_t y^{t+1} \qquad \text{(Dual Averaging)}$$

Algorithm 1: An Optimal Algorithm for $\mu = 0$

- **Input**: Interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$, $t = 1$
- **Repeat** (until some convergence criterion is met)

$$y^{t+1} := \arg\min_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{h_{\mathcal{Y}}}(y, y^t) \qquad \text{(Dual Ascent)}$$

$$\tilde{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^t \qquad \text{(Interpolation)}$$

$$x^{t+1} := \arg\min_{x\in\mathcal{X}} g(x) + \langle \hat{\nabla}_x \Phi(x^t, y^{t+1}, \zeta_x^t) + \hat{\nabla} f(\tilde{x}^{t+1}, \xi^t), x - x^t\rangle + \tau_t^{-1} D_{h_{\mathcal{X}}}(x, x^t)) \qquad \text{(Primal Descent)}$$

$$s^{t+1} := (1+\theta_{t+1})\hat{\nabla}_y \Phi(x^{t+1}, y^{t+1}, \zeta_y^{t+1}) - \theta_{t+1}\hat{\nabla}_y \Phi(x^t, y^t, \zeta_y^t) \qquad \text{(Extrap.)}$$

$$\bar{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^{t+1} \qquad \text{(Primal Averaging)}$$

$$\bar{y}^{t+1} := (1-\beta_t)\bar{y}^t + \beta_t y^{t+1} \qquad \text{(Dual Averaging)}$$

$$t := t+1$$

Algorithm 1: An Optimal Algorithm for $\mu = 0$

- **Input**: Interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$, $t = 1$
- **Repeat** (until some convergence criterion is met)

$$y^{t+1} := \arg\min_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{h_{\mathcal{Y}}}(y, y^t) \qquad \text{(Dual Ascent)}$$

$$\tilde{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^t \qquad \text{(Interpolation)}$$

$$x^{t+1} := \arg\min_{x\in\mathcal{X}} g(x) + \langle \hat{\nabla}_x \Phi(x^t, y^{t+1}, \zeta_x^t) + \hat{\nabla} f(\tilde{x}^{t+1}, \xi^t), x - x^t\rangle + \tau_t^{-1} D_{h_{\mathcal{X}}}(x, x^t)) \qquad \text{(Primal Descent)}$$

$$s^{t+1} := (1+\theta_{t+1})\hat{\nabla}_y \Phi(x^{t+1}, y^{t+1}, \zeta_y^{t+1}) - \theta_{t+1}\hat{\nabla}_y \Phi(x^t, y^t, \zeta_y^t) \qquad \text{(Extrap.)}$$

$$\bar{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^{t+1} \qquad \text{(Primal Averaging)}$$

$$\bar{y}^{t+1} := (1-\beta_t)\bar{y}^t + \beta_t y^{t+1} \qquad \text{(Dual Averaging)}$$

$$t := t+1$$

- **Output**: $(\bar{x}^t, \bar{y}^t)$

Definitions and Assumptions

Definitions and Assumptions

- Bregman diameters:

$$\Omega_{h_{\mathcal{X}}} \triangleq \sup_{x\in\mathcal{X},x'\in\mathcal{X}^o} D_{h_{\mathcal{X}}}(x,x'), \quad \Omega_{h_{\mathcal{Y}}} \triangleq \sup_{y\in\mathcal{Y},y'\in\mathcal{Y}^o} D_{h_{\mathcal{Y}}}(y,y').$$

Definitions and Assumptions

- Bregman diameters:

$$\Omega_{h_{\mathcal{X}}} \triangleq \sup_{x\in\mathcal{X},x'\in\mathcal{X}^o} D_{h_{\mathcal{X}}}(x,x'), \quad \Omega_{h_{\mathcal{Y}}} \triangleq \sup_{y\in\mathcal{Y},y'\in\mathcal{Y}^o} D_{h_{\mathcal{Y}}}(y,y').$$

- Gradient noises at iteration t:

$$\begin{aligned}
\delta^t_{y,\Phi} &\triangleq \hat{\nabla}_y \Phi(x^t, y^t, \zeta^t_y) - \nabla_y \Phi(x^t, y^t), \\
\delta^t_{x,\Phi} &\triangleq \hat{\nabla}_x \Phi(x^t, y^{t+1}, \zeta^t_x) - \nabla_x \Phi(x^t, y^{t+1}), \\
\delta^t_{x,f} &\triangleq \hat{\nabla} f(\tilde{x}^{t+1}, \xi^t) - \nabla f(\tilde{x}^{t+1}).
\end{aligned}$$

Assumptions 1 (On Constraint Sets)

(A) *The Bregman diameters $\Omega_{h_{\mathcal{X}}}$ and $\Omega_{h_{\mathcal{Y}}}$ are bounded.*

(B) *The set $\mathcal{X}$ is bounded and the Bregman diameter $\Omega_{h_{\mathcal{Y}}}$ is bounded.*

Definitions and Assumptions

Assumptions 2 (On Gradient Noises)

Define $\mathbb{E}_t[\cdot] \triangleq \mathbb{E}[\cdot \,|\, \mathcal{F}_t]$*. For any* $x \in \mathcal{X}$ *and* $y \in \mathcal{Y}$ *and any* $t \in \mathbb{N}$*, there exist positive constants* $\sigma_{y,\Phi}$*,* $\sigma_{x,\Phi}$ *and* $\sigma_{x,f}$ *such that*

- A (Unbiasedness) $\mathbb{E}_{t-1}[\delta^t_{y,\Phi}] = 0$, $\mathbb{E}_{t-1}[\delta^t_{x,\Phi}] = 0$, $\mathbb{E}_{t-1}[\delta^t_{x,f}] = 0$ *a.s.*,
- B (Bounded variance) $\mathbb{E}_{t-1}[\|\delta^t_{y,\Phi}\|_*^2] \le \sigma^2_{y,\Phi}$, $\mathbb{E}_{t-1}[\|\delta^t_{x,\Phi}\|_*^2] \le \sigma^2_{x,\Phi}$, $\mathbb{E}_{t-1}[\|\delta^t_{x,f}\|_*^2] \le \sigma^2_{x,f}$ *a.s.*,
- C (Sub-Gaussian distributions) $\mathbb{E}_{t-1}\left[\exp\left(\|\delta^t_{y,\Phi}\|_*^2/\sigma^2_{y,\Phi}\right)\right] \le \exp(1)$, $\mathbb{E}_{t-1}\left[\exp\left(\|\delta^t_{x,\Phi}\|_*^2/\sigma^2_{x,\Phi}\right)\right] \le \exp(1)$, $\mathbb{E}_{t-1}\left[\exp\left(\|\delta^t_{x,f}\|_*^2/\sigma^2_{x,f}\right)\right] \le \exp(1)$ *a.s.*.

Convergence Results

Theorem 1

Let Assumptions 1(A) and 2(A) hold. In Algorithm 1, for any $t \in \mathbb{N}$, choose

$$\theta_t = \frac{t-1}{t}, \quad \beta_t = \frac{2}{t+1}, \quad \alpha_t = \frac{1}{16\left(L_{yx} + L_{yy} + \rho\sigma_{y,\Phi}\sqrt{t}\right)},$$
$$\tau_t = \frac{t}{2\left(2L + (L_{xx} + L_{yx})t + \rho'(\sigma_{x,\Phi} + \sigma_{x,f})t^{3/2}\right)},$$

where $\rho, \rho' > 0$ are constants independent of the parameters of interest, i.e., $(L, L_{xx}, L_{yx}, L_{yy}, \sigma_{x,f}, \sigma_{x,\Phi}, \sigma_{y,\Phi}, t)$.

1. If Assumption 2(B) also holds, then for any $T \geq 3$, we have

$$\mathbb{E}[G(\bar{x}^T, \bar{y}^T)] \leq B_{\mathrm{e}}(T) \triangleq \frac{16L}{T(T-1)}\Omega_{h_{\mathcal{X}}} + \frac{8(L_{xx} + L_{yx})}{T}\Omega_{h_{\mathcal{X}}}$$
$$+ \frac{128(L_{yx} + L_{yy})}{T}\Omega_{h_{\mathcal{Y}}} + \frac{8\sigma_{y,\Phi}}{\sqrt{T}}\left(\frac{1}{\rho} + 16\rho\Omega_{h_{\mathcal{Y}}}\right) + \frac{8(\sigma_{x,f} + \sigma_{x,\Phi})}{\sqrt{T}}\left(\frac{1}{\rho'} + \rho'\Omega_{h_{\mathcal{X}}}\right).$$

Convergence Results

Thus, the oracle complexity of obtaining an ϵ-*expected duality gap* is

$$O\left(\sqrt{L/\epsilon} + (L_{xx} + L_{yx} + L_{yy})/\epsilon + \left((\sigma_{x,f} + \sigma_{x,\Phi})^2 + \sigma_{y,\Phi}^2\right)/\epsilon^2\right).$$

2. Let $\nu \in (0, 1/6]$. If Assumption 2(C) also holds, then w.p. at least $1 - 6\nu$,

$$G(\bar{x}^T, \bar{y}^T) \le B_{\rm e}(T) + \frac{8\sigma_{y,\Phi}}{\sqrt{T}}\left(\frac{\log(1/\nu)}{\rho} + \sqrt{\log(1/\nu)\Omega_{h_{\mathcal{Y}}}}\right) + \frac{8(\sigma_{x,\Phi} + \sigma_{x,f})}{\sqrt{T}}\left(\frac{\log(1/\nu)}{\rho'} + \sqrt{\log(1/\nu)\Omega_{h_{\mathcal{X}}}}\right).$$

Thus, the oracle complexity of obtaining an ϵ-*duality gap w.p.* $\geq 1 - \nu$ is

$$O\left(\sqrt{\frac{L}{\epsilon}} + \frac{L_{xx} + L_{yx} + L_{yy}}{\epsilon} + \frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2 + \sigma_{y,\Phi}^2}{\epsilon^2}\log\left(\frac{1}{\nu}\right)\right).$$

Restart Scheme for Strongly Convex Minimization

Restart Scheme for Strongly Convex Minimization

- Most of the subroutines need to satisfy:
  For any starting point $\bar{x}^1$ and any $\epsilon, \delta > 0$, there exists $T \in \mathbb{N}$ such that

$$\mathbb{E}[\|\bar{x}^1 - x^*\|^2] \leq \delta \quad \Longrightarrow \quad \mathbb{E}[f(\bar{x}^T) - f(x^*)] \leq \epsilon.$$

  where $\bar{x}^T$ denotes the T-th iterate produced by the subroutine.

Restart Scheme for Strongly Convex Minimization

- Most of the subroutines need to satisfy:
  For any starting point $\bar{x}^1$ and any $\epsilon, \delta > 0$, there exists $T \in \mathbb{N}$ such that
  $$\mathbb{E}[\|\bar{x}^1 - x^*\|^2] \le \delta \quad \Longrightarrow \quad \mathbb{E}[f(\bar{x}^T) - f(x^*)] \le \epsilon.$$
  where $\bar{x}^T$ denotes the T-th iterate produced by the subroutine.
- By the strong convexity of f, we can bound
  $$\mathbb{E}[\|\bar{x}^1 - x^*\|^2] \le (2/\mu)\mathbb{E}[f(\bar{x}^1) - f(x^*)]$$
  and thus establish a recursion.

Restart Scheme for Strongly Convex Minimization

- Most of the subroutines need to satisfy:
  For any starting point $\bar{x}^1$ and any $\epsilon, \delta > 0$, there exists $T \in \mathbb{N}$ such that
  $$\mathbb{E}[\|\bar{x}^1 - x^*\|^2] \leq \delta \quad \Longrightarrow \quad \mathbb{E}[f(\bar{x}^T) - f(x^*)] \leq \epsilon.$$
  where $\bar{x}^T$ denotes the T-th iterate produced by the subroutine.
- By the strong convexity of f, we can bound
  $$\mathbb{E}[\|\bar{x}^1 - x^*\|^2] \leq (2/\mu)\mathbb{E}[f(\bar{x}^1) - f(x^*)]$$
  and thus establish a recursion.
- However, this *does not* work for SPP (convergence measured by *duality gap*, and only diameters $\Omega_{h_\mathcal{X}}$ and $\Omega_{h_\mathcal{Y}}$ appear in the bound)

Restart Scheme for Strongly Convex Minimization

- Most of the subroutines need to satisfy:
  For any starting point $\bar{x}^1$ and any $\epsilon, \delta > 0$, there exists $T \in \mathbb{N}$ such that
  $$\mathbb{E}[\|\bar{x}^1 - x^*\|^2] \le \delta \quad \Longrightarrow \quad \mathbb{E}[f(\bar{x}^T) - f(x^*)] \le \epsilon.$$
  where $\bar{x}^T$ denotes the T-th iterate produced by the subroutine.
- By the strong convexity of f, we can bound
  $$\mathbb{E}[\|\bar{x}^1 - x^*\|^2] \le (2/\mu)\mathbb{E}[f(\bar{x}^1) - f(x^*)]$$
  and thus establish a recursion.
- However, this *does not* work for SPP (convergence measured by *duality gap*, and only diameters $\Omega_{h_{\mathcal{X}}}$ and $\Omega_{h_{\mathcal{Y}}}$ appear in the bound)

  $\Longrightarrow$ New schemes need to be developed.

Rescaled Distance Generating Function (DGF)

Rescaled Distance Generating Function (DGF)

- Fix any $x_{\rm c} \in \mathcal{X}^o$ and define $\bar{\mathcal{X}}(x_{\rm c}, R) \triangleq R\mathcal{X} + x_{\rm c}$, where $R > 0$.

Rescaled Distance Generating Function (DGF)

- Fix any $x_{\mathrm{c}} \in \mathcal{X}^o$ and define $\bar{\mathcal{X}}(x_{\mathrm{c}}, R) \triangleq R\mathcal{X} + x_{\mathrm{c}}$, where $R > 0$.
- Define a rescaled DGFs on $\bar{\mathcal{X}}(x_{\mathrm{c}}, R)$:

$$\tilde{h}_{\bar{\mathcal{X}}(x_{\mathrm{c}},R)}(x) \triangleq R^2 h_{\mathcal{X}}\left(\frac{x - x_{\mathrm{c}}}{R}\right). \tag{2}$$

Rescaled Distance Generating Function (DGF)

- Fix any $x_{\rm c} \in \mathcal{X}^o$ and define $\bar{\mathcal{X}}(x_{\rm c}, R) \triangleq R\mathcal{X} + x_{\rm c}$, where $R > 0$.
- Define a rescaled DGFs on $\bar{\mathcal{X}}(x_{\rm c}, R)$:

$$\tilde{h}_{\bar{\mathcal{X}}(x_{\rm c},R)}(x) \triangleq R^2 h_{\mathcal{X}}\left(\frac{x - x_{\rm c}}{R}\right). \tag{2}$$

- The corresponding Bregman distances are

$$D_{\tilde{h}_{\bar{\mathcal{X}}(x_{\rm c},R)}}(x, x') = R^2 \left\{ h_{\mathcal{X}}\left(\frac{x - x_{\rm c}}{R}\right) - h_{\mathcal{X}}\left(\frac{x' - x_{\rm c}}{R}\right) - \left\langle \nabla h_{\mathcal{X}}\left(\frac{x' - x_{\rm c}}{R}\right), \frac{x - x'}{R} \right\rangle \right\}.$$

Rescaled Distance Generating Function (DGF)

- Fix any $x_{\rm c} \in \mathcal{X}^o$ and define $\bar{\mathcal{X}}(x_{\rm c}, R) \triangleq R\mathcal{X} + x_{\rm c}$, where $R > 0$.
- Define a rescaled DGFs on $\bar{\mathcal{X}}(x_{\rm c}, R)$:

$$\tilde{h}_{\bar{\mathcal{X}}(x_{\rm c},R)}(x) \triangleq R^2 h_{\mathcal{X}} \left(\frac{x - x_{\rm c}}{R} \right). \tag{2}$$

- The corresponding Bregman distances are

$$D_{\tilde{h}_{\bar{\mathcal{X}}(x_{\rm c},R)}}(x, x') = R^2 \left\{ h_{\mathcal{X}} \left(\frac{x - x_{\rm c}}{R} \right) - h_{\mathcal{X}} \left(\frac{x' - x_{\rm c}}{R} \right) - \left\langle \nabla h_{\mathcal{X}} \left(\frac{x' - x_{\rm c}}{R} \right), \frac{x - x'}{R} \right\rangle \right\}.$$

- Define $\mathcal{B}(x_{\rm c}, R) \triangleq \{x \in \mathbb{X} : \|x - x_{\rm c}\| \le R\}$. If $\mathcal{B}(0, 1) \subseteq \mathsf{dom}\, h_{\mathcal{X}}$, then

$$\sup_{x \in \mathcal{X} \cap \mathcal{B}(x_{\rm c}, R)} D_{\tilde{h}_{\bar{\mathcal{X}}(x_{\rm c},R)}}(x, x_{\rm c}) \le R^2 \Omega'_{h_{\mathcal{X}}},$$

where $\Omega'_{h_{\mathcal{X}}} \triangleq \sup_{z \in \mathcal{B}(0,1)} D_{h_{\mathcal{X}}}(z, 0) < +\infty$.

Rescaled Distance Generating Function (DGF)

- ▷ Fix any $x_{\rm c} \in \mathcal{X}^o$ and define $\bar{\mathcal{X}}(x_{\rm c}, R) \triangleq R\mathcal{X} + x_{\rm c}$, where $R > 0$.
- ▷ Define a rescaled DGFs on $\bar{\mathcal{X}}(x_{\rm c}, R)$:

$$\tilde{h}_{\bar{\mathcal{X}}(x_{\rm c},R)}(x) \triangleq R^2 h_{\mathcal{X}} \left(\frac{x - x_{\rm c}}{R} \right). \tag{2}$$

- ▷ The corresponding Bregman distances are

$$D_{\tilde{h}_{\bar{\mathcal{X}}(x_{\rm c},R)}}(x, x') = R^2 \left\{ h_{\mathcal{X}} \left(\frac{x - x_{\rm c}}{R} \right) - h_{\mathcal{X}} \left(\frac{x' - x_{\rm c}}{R} \right) - \left\langle \nabla h_{\mathcal{X}} \left(\frac{x' - x_{\rm c}}{R} \right), \frac{x - x'}{R} \right\rangle \right\}.$$

- ▷ Define $\mathcal{B}(x_{\rm c}, R) \triangleq \{x \in \mathbb{X} : \|x - x_{\rm c}\| \le R\}$. If $\mathcal{B}(0, 1) \subseteq \mathsf{dom}\, h_{\mathcal{X}}$, then

$$\sup_{x \in \mathcal{X} \cap \mathcal{B}(x_{\rm c},R)} D_{\tilde{h}_{\bar{\mathcal{X}}(x_{\rm c},R)}}(x, x_{\rm c}) \le R^2 \Omega'_{h_{\mathcal{X}}},$$

$$\text{where } \Omega'_{h_{\mathcal{X}}} \triangleq \sup_{z \in \mathcal{B}(0,1)} D_{h_{\mathcal{X}}}(z, 0) < +\infty.$$

- ▷ If $\mathbb{X}$ is a Hilbert space and $h_{\mathcal{X}} = (1/2) \|\cdot\|^2$, then

$$\tilde{h}_{\bar{\mathcal{X}}(x_{\rm c},R)}(x) = (1/2) \|x - x_{\rm c}\|^2, \quad D_{\tilde{h}_{\bar{\mathcal{X}}(x_{\rm c},R)}}(x, x') = (1/2) \|x - x'\|^2.$$

Algorithm 1R: Algorithm 1 with Rescaled Geometry

Algorithm 1R: Algorithm 1 with Rescaled Geometry

- **Input**: Starting primal variable $x^0 \in \mathcal{X}^o$, radius R, primal constraint set $\mathcal{X}'$ ($\mathcal{X}' \subseteq \mathcal{X}$), number of iterations T, interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$

Algorithm 1R: Algorithm 1 with Rescaled Geometry

- **Input**: Starting primal variable $x^0 \in \mathcal{X}^o$, radius R, primal constraint set $\mathcal{X}'$ ($\mathcal{X}' \subseteq \mathcal{X}$), number of iterations T, interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$

Algorithm 1R: Algorithm 1 with Rescaled Geometry

- **Input**: Starting primal variable $x^0 \in \mathcal{X}^o$, radius R, primal constraint set $\mathcal{X}'$ ($\mathcal{X}' \subseteq \mathcal{X}$), number of iterations T, interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$
- **Define**: $\bar{\mathcal{X}}(x^1, R)$ and $\tilde{h}_{\bar{\mathcal{X}}(x^1,R)}$ using $h_{\mathcal{X}}$, x^1 and R

Algorithm 1R: Algorithm 1 with Rescaled Geometry

- **Input**: Starting primal variable $x^0 \in \mathcal{X}^o$, radius R, primal constraint set $\mathcal{X}'$ ($\mathcal{X}' \subseteq \mathcal{X}$), number of iterations T, interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$
- **Define**: $\bar{\mathcal{X}}(x^1, R)$ and $\tilde{h}_{\bar{\mathcal{X}}(x^1,R)}$ using $h_{\mathcal{X}}$, x^1 and R
- **For** $t = 1, \ldots, T-1$

$$y^{t+1} := \arg\min_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{\tilde{h}_{\mathcal{Y}}}(y, y^t) \qquad \text{(Dual Ascent)}$$

Algorithm 1R: Algorithm 1 with Rescaled Geometry

- **Input**: Starting primal variable $x^0 \in \mathcal{X}^o$, radius R, primal constraint set $\mathcal{X}'$ ($\mathcal{X}' \subseteq \mathcal{X}$), number of iterations T, interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$
- **Define**: $\bar{\mathcal{X}}(x^1, R)$ and $\tilde{h}_{\bar{\mathcal{X}}(x^1,R)}$ using $h_{\mathcal{X}}$, x^1 and R
- **For** $t = 1, \ldots, T-1$

$$y^{t+1} := \arg\min_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{\tilde{h}_{\mathcal{Y}}}(y, y^t) \qquad \text{(Dual Ascent)}$$

$$\tilde{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^t \qquad \text{(Interpolation)}$$

Algorithm 1R: Algorithm 1 with Rescaled Geometry

- **Input**: Starting primal variable $x^0 \in \mathcal{X}^o$, radius R, primal constraint set $\mathcal{X}'$ ($\mathcal{X}' \subseteq \mathcal{X}$), number of iterations T, interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$
- **Define**: $\bar{\mathcal{X}}(x^1, R)$ and $\tilde{h}_{\bar{\mathcal{X}}(x^1,R)}$ using $h_{\mathcal{X}}$, x^1 and R
- **For** $t = 1, \ldots, T-1$

$$y^{t+1} := \arg\min_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{\tilde{h}_{\mathcal{Y}}}(y, y^t) \qquad \text{(Dual Ascent)}$$

$$\tilde{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^t \qquad \text{(Interpolation)}$$

$$x^{t+1} := \arg\min_{x\in\mathcal{X}'} g(x) + \langle \hat{\nabla}_x \Phi(x^t, y^{t+1}, \zeta_x^t) + \hat{\nabla} f(\tilde{x}^{t+1}, \xi^t), x - x^t\rangle + \tau_t^{-1} D_{\tilde{h}_{\bar{\mathcal{X}}(x^1,R)}}(x, x^t) \qquad \text{(Primal Descent)}$$

Algorithm 1R: Algorithm 1 with Rescaled Geometry

- **Input**: Starting primal variable $x^0 \in \mathcal{X}^o$, radius R, primal constraint set $\mathcal{X}'$ ($\mathcal{X}' \subseteq \mathcal{X}$), number of iterations T, interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$
- **Define**: $\bar{\mathcal{X}}(x^1, R)$ and $\tilde{h}_{\bar{\mathcal{X}}(x^1,R)}$ using $h_{\mathcal{X}}$, x^1 and R
- **For** $t = 1, \ldots, T-1$

$$y^{t+1} := \arg\min_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{\tilde{h}_{\mathcal{Y}}}(y, y^t) \qquad \text{(Dual Ascent)}$$

$$\tilde{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^t \qquad \text{(Interpolation)}$$

$$x^{t+1} := \arg\min_{x\in\mathcal{X}'} g(x) + \langle \hat{\nabla}_x \Phi(x^t, y^{t+1}, \zeta_x^t) + \hat{\nabla} f(\tilde{x}^{t+1}, \xi^t), x - x^t\rangle + \tau_t^{-1} D_{\tilde{h}_{\bar{\mathcal{X}}(x^1,R)}}(x, x^t) \qquad \text{(Primal Descent)}$$

$$s^{t+1} := (1+\theta_{t+1})\hat{\nabla}_y \Phi(x^{t+1}, y^{t+1}, \zeta_y^{t+1}) - \theta_{t+1}\hat{\nabla}_y \Phi(x^t, y^t, \zeta_y^t) \quad \text{(Extrap.)}$$

Algorithm 1R: Algorithm 1 with Rescaled Geometry

- **Input**: Starting primal variable $x^0 \in \mathcal{X}^o$, radius R, primal constraint set $\mathcal{X}'$ ($\mathcal{X}' \subseteq \mathcal{X}$), number of iterations T, interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$
- **Define**: $\bar{\mathcal{X}}(x^1, R)$ and $\tilde{h}_{\bar{\mathcal{X}}(x^1,R)}$ using $h_{\mathcal{X}}$, x^1 and R
- **For** $t = 1, \ldots, T-1$

$$y^{t+1} := \arg\min_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{\tilde{h}_{\mathcal{Y}}}(y, y^t) \qquad \text{(Dual Ascent)}$$

$$\tilde{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^t \qquad \text{(Interpolation)}$$

$$x^{t+1} := \arg\min_{x\in\mathcal{X}'} g(x) + \langle \hat{\nabla}_x \Phi(x^t, y^{t+1}, \zeta_x^t) + \hat{\nabla} f(\tilde{x}^{t+1}, \xi^t), x - x^t\rangle + \tau_t^{-1} D_{\tilde{h}_{\bar{\mathcal{X}}(x^1,R)}}(x, x^t) \qquad \text{(Primal Descent)}$$

$$s^{t+1} := (1+\theta_{t+1})\hat{\nabla}_y \Phi(x^{t+1}, y^{t+1}, \zeta_y^{t+1}) - \theta_{t+1}\hat{\nabla}_y \Phi(x^t, y^t, \zeta_y^t) \qquad \text{(Extrap.)}$$

$$\bar{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^{t+1}, \quad \bar{y}^{t+1} := (1-\beta_t)\bar{y}^t + \beta_t y^{t+1} \qquad \text{(Averaging)}$$

Algorithm 1R: Algorithm 1 with Rescaled Geometry

- **Input**: Starting primal variable $x^0 \in \mathcal{X}^o$, radius R, primal constraint set $\mathcal{X}'$ ($\mathcal{X}' \subseteq \mathcal{X}$), number of iterations T, interp. seq. $\{\beta_t\}_{t\in\mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t\in\mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t\in\mathbb{N}}$, relaxation seq. $\{\theta_t\}_{t\in\mathbb{N}}$, DGFs $h_{\mathcal{Y}} : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h_{\mathcal{X}} : \mathbb{X} \to \overline{\mathbb{R}}$
- **Init**: $(x^1, y^1) \in \mathcal{X}^o \times \mathcal{Y}^o$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \hat{\nabla}_y \Phi(x^1, y^1, \zeta_y^1)$
- **Define**: $\bar{\mathcal{X}}(x^1, R)$ and $\tilde{h}_{\bar{\mathcal{X}}(x^1,R)}$ using $h_{\mathcal{X}}$, x^1 and R
- **For** $t = 1, \ldots, T-1$

$$y^{t+1} := \arg\min_{y\in\mathcal{Y}} J(y) - \langle s^t, y - y^t\rangle + \alpha_t^{-1} D_{\tilde{h}_{\mathcal{Y}}}(y, y^t) \quad \text{(Dual Ascent)}$$

$$\tilde{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^t \quad \text{(Interpolation)}$$

$$x^{t+1} := \arg\min_{x\in\mathcal{X}'} g(x) + \langle \hat{\nabla}_x \Phi(x^t, y^{t+1}, \zeta_x^t) + \hat{\nabla} f(\tilde{x}^{t+1}, \xi^t), x - x^t\rangle + \tau_t^{-1} D_{\tilde{h}_{\bar{\mathcal{X}}(x^1,R)}}(x, x^t) \quad \text{(Primal Descent)}$$

$$s^{t+1} := (1+\theta_{t+1})\hat{\nabla}_y \Phi(x^{t+1}, y^{t+1}, \zeta_y^{t+1}) - \theta_{t+1}\hat{\nabla}_y \Phi(x^t, y^t, \zeta_y^t) \quad \text{(Extrap.)}$$

$$\bar{x}^{t+1} := (1-\beta_t)\bar{x}^t + \beta_t x^{t+1}, \quad \bar{y}^{t+1} := (1-\beta_t)\bar{y}^t + \beta_t y^{t+1} \quad \text{(Averaging)}$$

- **Output**: $(\bar{x}^T, \bar{y}^T)$

Easily Computable Solutions

$$\arg\min_{x \in \mathcal{X}'} g(x) + \langle x^*, x \rangle + \tau_t^{-1} R^2 h_{\mathcal{X}} \left(\frac{x - x_{\mathrm{c}}}{R} \right)$$

Has an easily computable solution if

Easily Computable Solutions

$$\boxed{\operatorname*{arg\,min}_{x\in\mathcal{X}'} g(x) + \langle x^*, x\rangle + \tau_t^{-1} R^2 h_{\mathcal{X}}\left(\frac{x - x_{\mathrm{c}}}{R}\right)}$$

Has an easily computable solution if

- $g \equiv 0$ and $\mathcal{X}' = \mathcal{X} = \mathbb{X}$,

Easily Computable Solutions

$$\boxed{\arg\min_{x \in \mathcal{X}'} g(x) + \langle x^*, x \rangle + \tau_t^{-1} R^2 h_{\mathcal{X}} \left(\frac{x - x_{\mathrm{c}}}{R} \right)}$$

Has an easily computable solution if

- $g \equiv 0$ and $\mathcal{X}' = \mathcal{X} = \mathbb{X}$,
- $\mathbb{X}$ is a Hilbert space

Easily Computable Solutions

$$\boxed{\operatorname*{arg\,min}_{x \in \mathcal{X}'} g(x) + \langle x^*, x \rangle + \tau_t^{-1} R^2 h_{\mathcal{X}} \left(\frac{x - x_{\mathrm{c}}}{R} \right)}$$

Has an easily computable solution if

- ▷ $g \equiv 0$ and $\mathcal{X}' = \mathcal{X} = \mathbb{X}$,
- ▷ $\mathbb{X}$ is a Hilbert space
  - $\mathcal{X}' = \mathcal{X}$ and $h_{\mathcal{X}} = (1/2) \|\cdot\|_{\mathbb{X}}^2$,

Easily Computable Solutions

$$\boxed{\underset{x \in \mathcal{X}'}{\arg\min}\, g(x) + \langle x^*, x \rangle + \tau_t^{-1} R^2 h_{\mathcal{X}} \left(\frac{x - x_{\mathrm{c}}}{R} \right)}$$

Has an easily computable solution if

- $g \equiv 0$ and $\mathcal{X}' = \mathcal{X} = \mathbb{X}$,
- $\mathbb{X}$ is a Hilbert space
  - $\mathcal{X}' = \mathcal{X}$ and $h_{\mathcal{X}} = (1/2) \|\cdot\|_{\mathbb{X}}^2$,
  - $g \equiv 0$, $\mathcal{X}'$ = any set with easily computable projection, $h_{\mathcal{X}} = (1/2) \|\cdot\|_{\mathbb{X}}^2$.

Convergence Results for Algorithm 1R

Theorem 2

Assume that $\mathcal{B}(0,1) \subseteq \mathsf{dom}\, h_{\mathcal{X}}$*, and let Assumptions 1(B), 2(A) and 2(C) hold. Fix any* $\varsigma \in (0, 1/6]$*. In Algorithm 1R, choose* $\mathcal{X}'$ *such that* $x^* \in \mathcal{X}'$ *and* $D_{\mathcal{X}'} \leq R$*, and choose*

Convergence Results for Algorithm 1R

Theorem 2

Assume that $\mathcal{B}(0,1) \subseteq \mathsf{dom}\, h_{\mathcal{X}}$*, and let Assumptions 1(B), 2(A) and 2(C) hold. Fix any* $\varsigma \in (0, 1/6]$*. In Algorithm 1R, choose* $\mathcal{X}'$ *such that* $x^* \in \mathcal{X}'$ *and* $D_{\mathcal{X}'} \leq R$*, and choose*

$$T \geq \Big\lceil \max\Big\{3,\ 64\sqrt{(L/\mu)\Omega'_{h_{\mathcal{X}}}},\ 2048(L_{xx}/\mu)\Omega'_{h_{\mathcal{X}}},\ 4096 L_{yx}(\mu R)^{-1}\sqrt{\Omega'_{h_{\mathcal{X}}}\Omega_{h_{\mathcal{Y}}}},$$

$$128^2 L_{yy}(\mu R^2)^{-1}\Omega_{h_{\mathcal{Y}}},\ 512^2(\sigma_{x,f}+\sigma_{x,\Phi})^2(\mu R)^{-2}\Big(4\sqrt{(1+\log(1/\nu))\Omega'_{h_{\mathcal{X}}}}+2\sqrt{\log(1/\nu)}\Big)^2,$$

$$512^2\sigma_{y,\Phi}^2(\mu R^2)^{-2}\Big(8\sqrt{2(1+\log(1/\nu))\Omega_{h_{\mathcal{Y}}}}+2\sqrt{\log(1/\nu)\Omega_{h_{\mathcal{Y}}}}\Big)^2\Big\}\Big\rceil.$$

Convergence Results for Algorithm 1R

Theorem 2

Assume that $\mathcal{B}(0,1) \subseteq \operatorname{dom} h_{\mathcal{X}}$, *and let Assumptions 1(B), 2(A) and 2(C) hold. Fix any* $\varsigma \in (0, 1/6]$. *In Algorithm 1R, choose* $\mathcal{X}'$ *such that* $x^* \in \mathcal{X}'$ *and* $D_{\mathcal{X}'} \le R$, *and choose*

$$T \ge \Big\lceil \max\Big\{3,\ 64\sqrt{(L/\mu)\Omega'_{h_{\mathcal{X}}}},\ 2048(L_{xx}/\mu)\Omega'_{h_{\mathcal{X}}},\ 4096 L_{yx}(\mu R)^{-1}\sqrt{\Omega'_{h_{\mathcal{X}}}\Omega_{h_{\mathcal{Y}}}},$$

$$128^2 L_{yy}(\mu R^2)^{-1}\Omega_{h_{\mathcal{Y}}},\ 512^2(\sigma_{x,f}+\sigma_{x,\Phi})^2(\mu R)^{-2}\Big(4\sqrt{(1+\log(1/\nu))\Omega'_{h_{\mathcal{X}}}}+2\sqrt{\log(1/\nu)}\Big)^2,$$

$$512^2\sigma_{y,\Phi}^2(\mu R^2)^{-2}\Big(8\sqrt{2(1+\log(1/\nu))\Omega_{h_{\mathcal{Y}}}}+2\sqrt{\log(1/\nu)\Omega_{h_{\mathcal{Y}}}}\Big)^2\Big\}\Big\rceil.$$

If we choose $R \ge 2\|x^0 - x^*\|$, $\{\beta_t\}_{t\in[T]}$ *and* $\{\theta_t\}_{t\in[T]}$ *as in Theorem 1, and* $\alpha_t = \alpha$ *and* $\tau_t = t\tau$ *for any* $t \in [T]$, *where*

$$\alpha = 1/\Big(16\big(\eta^{-1}L_{yx} + L_{yy} + \rho\sigma_{y,\Phi}\sqrt{T}\big)\Big), \quad \rho = (4R)^{-1}\sqrt{(1+\log(1/\varsigma))/(2\Omega'_{h_{\mathcal{X}}}\Omega_{h_{\mathcal{Y}}})},$$

$$\tau = 1/\Big(4L + 2(L_{xx}+\eta L_{yx})T + \rho'(\sigma_{x,\Phi}+\sigma_{x,f})T^{3/2}\Big), \quad \eta = (4/R)\sqrt{\Omega_{h_{\mathcal{Y}}}/\Omega'_{h_{\mathcal{X}}}},$$

$$\rho' = (8R)^{-1}\sqrt{(1+\log(1/\varsigma))/(\Omega'_{h_{\mathcal{X}}}\Omega_{h_{\mathcal{Y}}})},$$

Convergence Results for Algorithm 1R

then w.p. at least $1 - 6\nu$,

$$G(\bar{x}^T, \bar{y}^T) \leq B_R^{\mathrm{det}}(T) + B_R^{\mathrm{var}}(T) \leq \mu R^2/16,$$

Convergence Results for Algorithm 1R

then w.p. at least $1-6\nu$,

$$G(\bar{x}^T,\bar{y}^T) \le B_R^{\mathrm{det}}(T) + B_R^{\mathrm{var}}(T) \le \mu R^2/16,$$

where

$$\begin{aligned} B_R^{\mathrm{det}}(T) \triangleq & \frac{16LR^2}{T(T-1)}\Omega'_{h\mathcal{X}} + \frac{8L_{xx}R^2}{T-1}\Omega'_{h\mathcal{X}} \\ & + \frac{8L_{yx}R}{T-1}\left(\sqrt{\eta_x/\eta_y}\Omega'_{h\mathcal{X}} + 16\sqrt{\eta_y/\eta_x}\Omega_{h\mathcal{Y}}\right) + \frac{128L_{yy}}{T}\Omega_{h\mathcal{Y}}, \\ B_R^{\mathrm{var}}(T) \triangleq & \frac{4(\sigma_{x,\Phi}+\sigma_{x,f})R}{\sqrt{T}}\left\{4\sqrt{(1+\log(1/\nu))\Omega'_{h\mathcal{X}}} + 2\sqrt{\log(1/\nu)}\right\} \\ & + \frac{4\sigma_{y,\Phi}}{\sqrt{T}}\left\{8\sqrt{2(1+\log(1/\nu))\Omega_{h\mathcal{Y}}} + 2\sqrt{\log(1/\nu)\Omega_{h\mathcal{Y}}}\right\}. \end{aligned}$$

Convergence Results for Algorithm 1R

then w.p. at least $1 - 6\nu$,

$$G(\bar{x}^T, \bar{y}^T) \le B_R^{\rm det}(T) + B_R^{\rm var}(T) \le \mu R^2/16,$$

where

$$\begin{aligned} B_R^{\rm det}(T) \triangleq &\frac{16LR^2}{T(T-1)}\Omega'_{h\mathcal{X}} + \frac{8L_{xx}R^2}{T-1}\Omega'_{h\mathcal{X}} \\ &+ \frac{8L_{yx}R}{T-1}\left(\sqrt{\eta_x/\eta_y}\Omega'_{h\mathcal{X}} + 16\sqrt{\eta_y/\eta_x}\Omega_{h\mathcal{Y}}\right) + \frac{128L_{yy}}{T}\Omega_{h\mathcal{Y}}, \\ B_R^{\rm var}(T) \triangleq &\frac{4(\sigma_{x,\Phi} + \sigma_{x,f})R}{\sqrt{T}}\left\{4\sqrt{(1+\log(1/\nu))\Omega'_{h\mathcal{X}}} + 2\sqrt{\log(1/\nu)}\right\} \\ &+ \frac{4\sigma_{y,\Phi}}{\sqrt{T}}\left\{8\sqrt{2(1+\log(1/\nu))\Omega_{h\mathcal{Y}}} + 2\sqrt{\log(1/\nu)\Omega_{h\mathcal{Y}}}\right\}. \end{aligned}$$

Furthermore, $\|\bar{x}^T - x^*\| \le \sqrt{(2/\mu)(B_R^{\rm det}(T) + B_R^{\rm var}(T))} \le R/\big(2\sqrt{2}\big)$ *w.p. at least* $1 - 6\nu$.

Algorithm 2: Stochastic Restart Scheme

Algorithm 2: Stochastic Restart Scheme

- **Input**: Diameter estimate $U \geq D_{\mathcal{X}}$, starting primal variable $x_0 \in \mathcal{X}^o$, desired accuracy $\epsilon > 0$, error probability $\nu \in (0,1]$, $K = \left\lceil \max\left\{0, \log_2\left(\mu U^2/(4\epsilon)\right)\right\}\right\rceil + 1$, $\varsigma = \nu/(6K)$

Algorithm 2: Stochastic Restart Scheme

- **Input**: Diameter estimate $U \geq D_{\mathcal{X}}$, starting primal variable $x_0 \in \mathcal{X}^o$, desired accuracy $\epsilon > 0$, error probability $\nu \in (0,1]$, $K = \left\lceil \max\left\{0, \log_2\left(\mu U^2/(4\epsilon)\right)\right\}\right\rceil + 1$, $\varsigma = \nu/(6K)$
- **Init**: $R_1 = 2U$, $x_1 = x_0$, $y_0 \in \mathcal{Y}^o$

Algorithm 2: Stochastic Restart Scheme

- **Input**: Diameter estimate $U \geq D_{\mathcal{X}}$, starting primal variable $x_0 \in \mathcal{X}^o$, desired accuracy $\epsilon > 0$, error probability $\nu \in (0, 1]$, $K = \left\lceil \max\left\{0, \log_2\left(\mu U^2/(4\epsilon)\right)\right\}\right\rceil + 1$, $\varsigma = \nu/(6K)$
- **Init**: $R_1 = 2U$, $x_1 = x_0$, $y_0 \in \mathcal{Y}^o$
- **For** $k = 1, \ldots, K$

Algorithm 2: Stochastic Restart Scheme

- **Input**: Diameter estimate $U \geq D_{\mathcal{X}}$, starting primal variable $x_0 \in \mathcal{X}^o$, desired accuracy $\epsilon > 0$, error probability $\nu \in (0,1]$, $K = \left\lceil \max\left\{0, \log_2\left(\mu U^2/(4\epsilon)\right)\right\}\right\rceil + 1$, $\varsigma = \nu/(6K)$
- **Init**: $R_1 = 2U$, $x_1 = x_0$, $y_0 \in \mathcal{Y}^o$
- **For** $k = 1, \ldots, K$
  - $T_k := \Big\lceil \max\Big\{3,\ 64\sqrt{(L/\mu)\Omega'_{h_\mathcal{X}}},\ 2048(L_{xx}/\mu)\Omega'_{h_\mathcal{X}},$
    $512^2(\sigma_{x,f} + \sigma_{x,\Phi})^2(\mu R_k)^{-2}\Big(4\sqrt{(1+\log(1/\varsigma))\Omega'_{h_\mathcal{X}}} + 2\sqrt{\log(1/\varsigma)}\Big)^2,$
    $128^2 L_{yy}(\mu R_k^2)^{-1}\Omega_{h_\mathcal{Y}},\ 4096 L_{yx}(\mu R_k)^{-1}\sqrt{\Omega'_{h_\mathcal{X}}\Omega_{h_\mathcal{Y}}},$
    $512^2\sigma_{y,\Phi}^2(\mu R_k^2)^{-2}\Big(8\sqrt{2(1+\log(1/\varsigma))\Omega_{h_\mathcal{Y}}} + 2\sqrt{\log(1/\varsigma)\Omega_{h_\mathcal{Y}}}\Big)^2\ \Big\}\Big\rceil.$

Algorithm 2: Stochastic Restart Scheme

- **Input**: Diameter estimate $U \geq D_{\mathcal{X}}$, starting primal variable $x_0 \in \mathcal{X}^o$, desired accuracy $\epsilon > 0$, error probability $\nu \in (0,1]$, $K = \left\lceil \max\left\{0, \log_2\left(\mu U^2/(4\epsilon)\right)\right\}\right\rceil + 1$, $\varsigma = \nu/(6K)$
- **Init**: $R_1 = 2U$, $x_1 = x_0$, $y_0 \in \mathcal{Y}^o$
- **For** $k = 1, \ldots, K$
  - $$T_k := \Big\lceil \max\Big\{3,\ 64\sqrt{(L/\mu)\Omega'_{h_\mathcal{X}}},\ 2048(L_{xx}/\mu)\Omega'_{h_\mathcal{X}},\ 512^2(\sigma_{x,f} + \sigma_{x,\Phi})^2(\mu R_k)^{-2}\Big(4\sqrt{(1+\log(1/\varsigma))\Omega'_{h_\mathcal{X}}} + 2\sqrt{\log(1/\varsigma)}\Big)^2,\ 128^2 L_{yy}(\mu R_k^2)^{-1}\Omega_{h_\mathcal{Y}},\ 4096 L_{yx}(\mu R_k)^{-1}\sqrt{\Omega'_{h_\mathcal{X}}\Omega_{h_\mathcal{Y}}},\ 512^2\sigma_{y,\Phi}^2(\mu R_k^2)^{-2}\Big(8\sqrt{2(1+\log(1/\varsigma))\Omega_{h_\mathcal{Y}}} + 2\sqrt{\log(1/\varsigma)\Omega_{h_\mathcal{Y}}}\Big)^2\Big\}\Big\rceil.$$
  - Run Algorithm 1S for T_k iterations with starting primal variable x_k, radius R_k, constraint set $\mathcal{X}_k = \{x \in \mathcal{X} : \|x - x_k\| \leq R_k/2\}$ and other input parameters set as in Theorem 2, with output $(\bar{x}_k^{T_k}, \bar{y}_k^{T_k})$.

Algorithm 2: Stochastic Restart Scheme

- **Input**: Diameter estimate $U \geq D_{\mathcal{X}}$, starting primal variable $x_0 \in \mathcal{X}^o$, desired accuracy $\epsilon > 0$, error probability $\nu \in (0,1]$, $K = \left\lceil \max\left\{0, \log_2\left(\mu U^2/(4\epsilon)\right)\right\}\right\rceil + 1$, $\varsigma = \nu/(6K)$
- **Init**: $R_1 = 2U$, $x_1 = x_0$, $y_0 \in \mathcal{Y}^o$
- **For** $k = 1, \ldots, K$
  - $$T_k := \Big\lceil \max\Big\{3,\ 64\sqrt{(L/\mu)\Omega'_{h_{\mathcal{X}}}},\ 2048(L_{xx}/\mu)\Omega'_{h_{\mathcal{X}}},\\ 512^2(\sigma_{x,f}+\sigma_{x,\Phi})^2(\mu R_k)^{-2}\Big(4\sqrt{(1+\log(1/\varsigma))\Omega'_{h_{\mathcal{X}}}}+2\sqrt{\log(1/\varsigma)}\Big)^2,\\ 128^2 L_{yy}(\mu R_k^2)^{-1}\Omega_{h_{\mathcal{Y}}},\ 4096 L_{yx}(\mu R_k)^{-1}\sqrt{\Omega'_{h_{\mathcal{X}}}\Omega_{h_{\mathcal{Y}}}},\\ 512^2\sigma_{y,\Phi}^2(\mu R_k^2)^{-2}\Big(8\sqrt{2(1+\log(1/\varsigma))\Omega_{h_{\mathcal{Y}}}}+2\sqrt{\log(1/\varsigma)\Omega_{h_{\mathcal{Y}}}}\Big)^2\Big\}\Big\rceil.$$
  - Run Algorithm 1S for T_k iterations with starting primal variable x_k, radius R_k, constraint set $\mathcal{X}_k = \{x \in \mathcal{X} : \|x - x_k\| \leq R_k/2\}$ and other input parameters set as in Theorem 2, with output $(\bar{x}_k^{T_k}, \bar{y}_k^{T_k})$.
  - $R_{k+1} := R_k/\sqrt{2}$, $x_{k+1} := \bar{x}_k^{T_k}$.

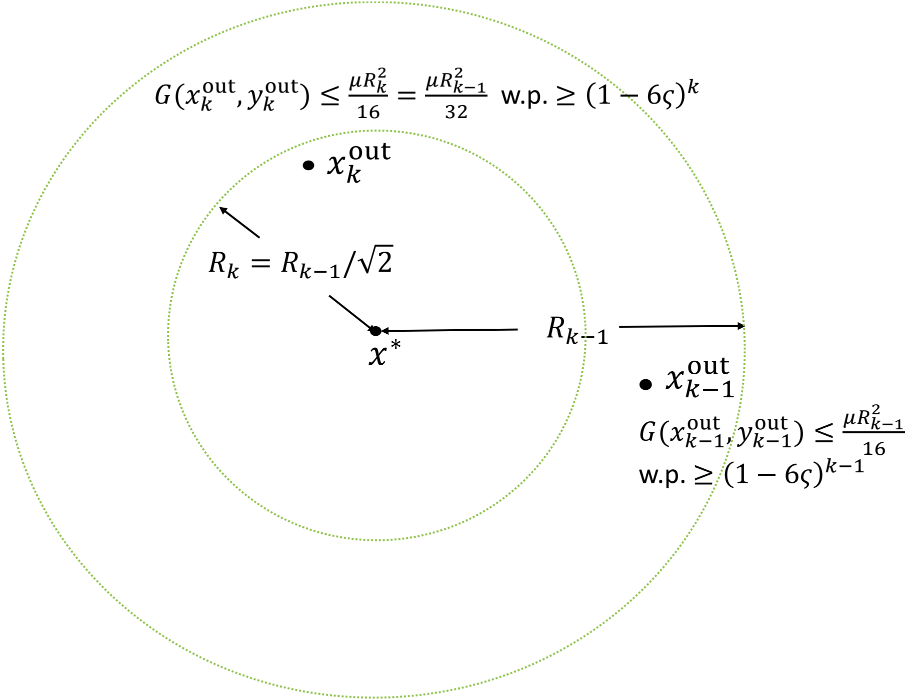
Algorithm 2: Stochastic Restart Scheme

- **Input**: Diameter estimate $U \geq D_{\mathcal{X}}$, starting primal variable $x_0 \in \mathcal{X}^o$, desired accuracy $\epsilon > 0$, error probability $\nu \in (0,1]$, $K = \lceil \max\{0, \log_2(\mu U^2/(4\epsilon))\}\rceil + 1$, $\varsigma = \nu/(6K)$
- **Init**: $R_1 = 2U$, $x_1 = x_0$, $y_0 \in \mathcal{Y}^o$
- **For** $k = 1, \ldots, K$
  - $T_k := \Big\lceil \max\Big\{3,\ 64\sqrt{(L/\mu)\Omega'_{h_{\mathcal{X}}}},\ 2048(L_{xx}/\mu)\Omega'_{h_{\mathcal{X}}},$

    $512^2(\sigma_{x,f} + \sigma_{x,\Phi})^2(\mu R_k)^{-2}\big(4\sqrt{(1+\log(1/\varsigma))\Omega'_{h_{\mathcal{X}}}} + 2\sqrt{\log(1/\varsigma)}\big)^2,$

    $128^2 L_{yy}(\mu R_k^2)^{-1}\Omega_{h_{\mathcal{Y}}},\ 4096 L_{yx}(\mu R_k)^{-1}\sqrt{\Omega'_{h_{\mathcal{X}}}\Omega_{h_{\mathcal{Y}}}},$

    $512^2\sigma_{y,\Phi}^2(\mu R_k^2)^{-2}\big(8\sqrt{2(1+\log(1/\varsigma))\Omega_{h_{\mathcal{Y}}}} + 2\sqrt{\log(1/\varsigma)\Omega_{h_{\mathcal{Y}}}}\big)^2\Big\}\Big\rceil.$
  - Run Algorithm 1S for T_k iterations with starting primal variable x_k, radius R_k, constraint set $\mathcal{X}_k = \{x \in \mathcal{X} : \|x - x_k\| \leq R_k/2\}$ and other input parameters set as in Theorem 2, with output $(\bar{x}_k^{T_k}, \bar{y}_k^{T_k})$.
  - $R_{k+1} := R_k/\sqrt{2}$, $x_{k+1} := \bar{x}_k^{T_k}$.
- **Output**: (x_{K+1}, y_{K+1})



$G(x_k^{\text{out}}, y_k^{\text{out}}) \leq \frac{\mu R_k^2}{16} = \frac{\mu R_{k-1}^2}{32}$ w.p. $\geq (1-6\varsigma)^k$
x_k^{out}
$R_k = R_{k-1}/\sqrt{2}$
R_{k-1}
x^*
x_{k-1}^{out}
$G(x_{k-1}^{\text{out}}, y_{k-1}^{\text{out}}) \leq \frac{\mu R_{k-1}^2}{16}$
w.p. $\geq (1-6\varsigma)^{k-1}$

Oracle Complexity

Theorem 3

Assume $\mathcal{B}(0,1) \subseteq \mathsf{dom}\, h_{\mathcal{X}}$ *and let Assumptions 1(B), 2(A) and 2(C) hold. In Algorithm 2, for any* $x_0 \in \mathcal{X}^o$*, desired accuracy* $\epsilon \in (0, \mu U^2/4]$ *and error probability* $\nu \in (0,1]$*, it holds that* $G(x_{K+1}, y_{K+1}) \le \epsilon$ *w.p. at least* $1-\nu$*.*

Oracle Complexity

Theorem 3

Assume $\mathcal{B}(0,1) \subseteq \mathsf{dom}\, h_{\mathcal{X}}$ and let Assumptions 1(B), 2(A) and 2(C) hold. In Algorithm 2, for any $x_0 \in \mathcal{X}^o$, desired accuracy $\epsilon \in (0, \mu U^2/4]$ and error probability $\nu \in (0,1]$, it holds that $G(x_{K+1}, y_{K+1}) \le \epsilon$ w.p. at least $1-\nu$.

Furthermore, the number of oracle calls

$$
\begin{aligned}
C_\epsilon^{\mathrm{st}} \le & \left(3 + 64\sqrt{(L/\mu)\Omega'_{h_{\mathcal{X}}}} + 2048(L_{xx}/\mu)\Omega'_{h_{\mathcal{X}}}\right)\left(\left\lceil \log_2\left(\mu U^2/(4\epsilon)\right)\right\rceil + 1\right) \\
& + 256^2\left(L_{yx}/\sqrt{\mu\epsilon}\right)\sqrt{\Omega'_{h_{\mathcal{X}}}\Omega_{h_{\mathcal{Y}}}} + 64^2\left(L_{yy}/\epsilon\right)\Omega_{h_{\mathcal{Y}}} \\
& + 1024^2\left\{(\sigma_{x,f} + \sigma_{x,\Phi})^2/(\epsilon\mu)\right\}\left\{(4\Omega'_{h_{\mathcal{X}}} + 1)\log\left(6\left[\log_2\left(\mu U^2(4\epsilon)^{-1}\right) + 2\right]/\nu\right) + 4\Omega'_{h_{\mathcal{X}}}\right\} \\
& + 1024^2\left(\sigma_{y,\Phi}^2/\epsilon^2\right)\left\{1 + \log\left(6\left[\log_2\left(\mu U^2(4\epsilon)^{-1}\right) + 2\right]/\nu\right)\right\}\Omega_{h_{\mathcal{Y}}}
\end{aligned}
$$

Oracle Complexity

Theorem 3

Assume $\mathcal{B}(0,1) \subseteq \mathsf{dom}\, h_{\mathcal{X}}$ *and let Assumptions 1(B), 2(A) and 2(C) hold. In Algorithm 2, for any* $x_0 \in \mathcal{X}^o$*, desired accuracy* $\epsilon \in (0, \mu U^2/4]$ *and error probability* $\nu \in (0,1]$*, it holds that* $G(x_{K+1}, y_{K+1}) \le \epsilon$ *w.p. at least* $1-\nu$*.*

Furthermore, the number of oracle calls

$$
\begin{aligned}
C_\epsilon^{\rm st} \le & \left(3 + 64\sqrt{(L/\mu)\Omega'_{h_{\mathcal{X}}}} + 2048(L_{xx}/\mu)\Omega'_{h_{\mathcal{X}}}\right)\left(\left\lceil \log_2\left(\mu U^2/(4\epsilon)\right)\right\rceil + 1\right) \\
& + 256^2\left(L_{yx}/\sqrt{\mu\epsilon}\right)\sqrt{\Omega'_{h_{\mathcal{X}}}\Omega_{h_{\mathcal{Y}}}} + 64^2\left(L_{yy}/\epsilon\right)\Omega_{h_{\mathcal{Y}}} \\
& + 1024^2\left\{(\sigma_{x,f}+\sigma_{x,\Phi})^2/(\epsilon\mu)\right\}\left\{(4\Omega'_{h_{\mathcal{X}}}+1)\log\left(6\left[\log_2\left(\mu U^2(4\epsilon)^{-1}\right)+2\right]/\nu\right) + 4\Omega'_{h_{\mathcal{X}}}\right\} \\
& + 1024^2\left(\sigma_{y,\Phi}^2/\epsilon^2\right)\left\{1 + \log\left(6\left[\log_2\left(\mu U^2(4\epsilon)^{-1}\right)+2\right]/\nu\right)\right\}\Omega_{h_{\mathcal{Y}}} \\
= & \, O\left(\left(\sqrt{\frac{L}{\mu}} + \frac{L_{xx}}{\mu}\right)\log\left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}} + \frac{L_{yy}}{\epsilon} + \left(\frac{(\sigma_{x,f}+\sigma_{x,\Phi})^2}{\mu\epsilon} + \frac{\sigma_{y,\Phi}^2}{\epsilon^2}\right)\log\left(\frac{\log(1/\epsilon)}{\nu}\right)\right).
\end{aligned}
$$

Complexity of convergence in expectation

Complexity of convergence in expectation

- Assume that $\mathsf{dom}\, g$ and $\mathsf{dom}\, J$ are closed.

Complexity of convergence in expectation

- ▷ Assume that $\mathsf{dom}\, g$ and $\mathsf{dom}\, J$ are closed.
- ▷ By compactness of $\mathcal{X}$ and $\mathcal{Y}$, invoke *Berge's maximum theorem* to conclude that $\bar{S}$ and $\underline{S}$ are continuous on $\mathcal{X} \cap \mathsf{dom}\, g$ and $\mathcal{Y} \cap \mathsf{dom}\, J$, respectively, so there exists $\Gamma < +\infty$ such that

Complexity of convergence in expectation

- Assume that $\mathsf{dom}\, g$ and $\mathsf{dom}\, J$ are closed.
- By compactness of $\mathcal{X}$ and $\mathcal{Y}$, invoke *Berge's maximum theorem* to conclude that $\bar{S}$ and $\underline{S}$ are continuous on $\mathcal{X} \cap \mathsf{dom}\, g$ and $\mathcal{Y} \cap \mathsf{dom}\, J$, respectively, so there exists $\Gamma < +\infty$ such that

$$\begin{aligned} & \sup_{x \in \mathsf{dom}\, g \cap \mathcal{X}} \sup_{y \in \mathsf{dom}\, J \cap \mathcal{Y}} G(x,y) \\ = \; & \sup_{x \in \mathsf{dom}\, g \cap \mathcal{X}} \bar{S}(x) - \inf_{y \in \mathsf{dom}\, J \cap \mathcal{Y}} \underline{S}(y) \leq \Gamma. \end{aligned}$$

Complexity of convergence in expectation

- Assume that $\mathsf{dom}\, g$ and $\mathsf{dom}\, J$ are closed.
- By compactness of $\mathcal{X}$ and $\mathcal{Y}$, invoke *Berge's maximum theorem* to conclude that $\bar{S}$ and $\underline{S}$ are continuous on $\mathcal{X} \cap \mathsf{dom}\, g$ and $\mathcal{Y} \cap \mathsf{dom}\, J$, respectively, so there exists $\Gamma < +\infty$ such that

$$\begin{aligned} & \sup\nolimits_{x \in \mathsf{dom}\, g \cap \mathcal{X}} \sup\nolimits_{y \in \mathsf{dom}\, J \cap \mathcal{Y}} G(x,y) \\ = \; & \sup\nolimits_{x \in \mathsf{dom}\, g \cap \mathcal{X}} \bar{S}(x) - \inf\nolimits_{y \in \mathsf{dom}\, J \cap \mathcal{Y}} \underline{S}(y) \le \Gamma. \end{aligned}$$

Theorem 4

Assume $\mathcal{B}(0,1) \subseteq \mathsf{dom}\, h_{\mathcal{X}}$ and let Assumptions 1(B), 2(A) and 2(C) hold. In Algorithm 2, for any $x_0 \in \mathcal{X}^o$ and $\varepsilon \in (0, \mu U^2/2]$, choose $\nu = \min\{\varepsilon/(2\Gamma), 1\}$ and $K = \lceil \log_2 \left(\mu U^2/(2\varepsilon)\right) \rceil + 1$. Then it holds that $\mathbb{E}[G(x_{K+1}, y_{K+1})] \le \varepsilon$.

Complexity of convergence in expectation

- Assume that $\mathsf{dom}\, g$ and $\mathsf{dom}\, J$ are closed.
- By compactness of $\mathcal{X}$ and $\mathcal{Y}$, invoke *Berge's maximum theorem* to conclude that $\bar{S}$ and $\underline{S}$ are continuous on $\mathcal{X} \cap \mathsf{dom}\, g$ and $\mathcal{Y} \cap \mathsf{dom}\, J$, respectively, so there exists $\Gamma < +\infty$ such that

$$
\begin{aligned}
& \sup_{x \in \mathsf{dom}\, g \cap \mathcal{X}} \sup_{y \in \mathsf{dom}\, J \cap \mathcal{Y}} G(x,y) \\
= \; & \sup_{x \in \mathsf{dom}\, g \cap \mathcal{X}} \bar{S}(x) - \inf_{y \in \mathsf{dom}\, J \cap \mathcal{Y}} \underline{S}(y) \le \Gamma.
\end{aligned}
$$

Theorem 4

Assume $\mathcal{B}(0,1) \subseteq \mathsf{dom}\, h_{\mathcal{X}}$ and let Assumptions 1(B), 2(A) and 2(C) hold. In Algorithm 2, for any $x_0 \in \mathcal{X}^o$ and $\varepsilon \in (0, \mu U^2/2]$, choose $\nu = \min\{\varepsilon/(2\Gamma), 1\}$ and $K = \lceil \log_2 (\mu U^2/(2\varepsilon)) \rceil + 1$. Then it holds that $\mathbb{E}[G(x_{K+1}, y_{K+1})] \le \varepsilon$.

Furthermore, the oracle complexity is

$$
O\left(\left(\sqrt{\frac{L}{\mu}} + \frac{L_{xx}}{\mu}\right) \log\left(\frac{1}{\varepsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\varepsilon}} + \frac{L_{yy}}{\varepsilon} + \left(\frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2}{\mu\varepsilon} + \frac{\sigma_{y,\Phi}^2}{\varepsilon^2}\right) \log\left(\frac{1}{\varepsilon}\right)\right).
$$

Future Directions

Future Directions

- Lower bounds on the complexities of L_{xx} and L_{yy}.

Future Directions

- Lower bounds on the complexities of L_{xx} and L_{yy}.
- In the strongly convex case ($\mu > 0$):

Future Directions

- Lower bounds on the complexities of L_{xx} and L_{yy}.
- In the strongly convex case ($\mu > 0$):
  - Relax the sub-Gaussian assumption on the gradient noises.

Future Directions

- Lower bounds on the complexities of L_{xx} and L_{yy}.
- In the strongly convex case ($\mu > 0$):
  - Relax the sub-Gaussian assumption on the gradient noises.
  - Remove the additional $\log(1/\epsilon)$ factors in the oracle complexities of $\sigma_{x,f}$, $\sigma_{x,\Phi}$ and $\sigma_{y,\Phi}$, in obtaining the ϵ-expected duality gap.

Thank you!