An Optimal Algorithm for Stochastic Three-Composite Optimization

Renbo Zhao[1], William B. Haskell[2], Vincent Y. F. Tan[3]

[1]ORC, Massachusetts Institute of Technology

[2]Dept. ISEM, National University of Singapore

[3]Dept. ECE & Math, National University of Singapore

INFORMS Annual Meeting
Phoenix, Arizona, Nov. 2018

Three-Composite Convex Minimization

Consider the following convex three-composite problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left[P(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})\right]. \qquad \texttt{(P)}$$

Three-Composite Convex Minimization

Consider the following convex three-composite problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left[P(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})\right]. \tag{P}$$

- ▷ $f : \mathbb{R}^d \to \mathbb{R}$, $g : \mathbb{R}^d \to \overline{\mathbb{R}}$ and $h : \mathbb{R}^m \to \overline{\mathbb{R}}$ are convex, closed, and proper (CCP) functions, where $\overline{\mathbb{R}} \triangleq \mathbb{R} \cup \{+\infty\}$.

Three-Composite Convex Minimization

Consider the following convex three-composite problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left[P(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})\right]. \quad \texttt{(P)}$$

- $f : \mathbb{R}^d \to \mathbb{R}$, $g : \mathbb{R}^d \to \overline{\mathbb{R}}$ and $h : \mathbb{R}^m \to \overline{\mathbb{R}}$ are convex, closed, and proper (CCP) functions, where $\overline{\mathbb{R}} \triangleq \mathbb{R} \cup \{+\infty\}$.
- $\mathbf{A} : \mathbb{R}^d \to \mathbb{R}^m$ is a linear operator ($\mathbf{A} \neq \mathbf{0}$, $\|\mathbf{A}\| = B$).

Three-Composite Convex Minimization

Consider the following convex three-composite problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \Big[P(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})\Big]. \qquad (\texttt{P})$$

- ▷ $f : \mathbb{R}^d \to \mathbb{R}$, $g : \mathbb{R}^d \to \overline{\mathbb{R}}$ and $h : \mathbb{R}^m \to \overline{\mathbb{R}}$ are convex, closed, and proper (CCP) functions, where $\overline{\mathbb{R}} \triangleq \mathbb{R}\cup\{+\infty\}$.
- ▷ $\mathbf{A} : \mathbb{R}^d \to \mathbb{R}^m$ is a linear operator ($\mathbf{A}\neq\mathbf{0}$, $\|\mathbf{A}\|=B$).
- ▷ f is differentiable with L-Lipschitz gradient on $\mathbb{R}^d$ ($L>0$).

Three-Composite Convex Minimization

Consider the following convex three-composite problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left[P(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})\right]. \tag{P}$$

- ▷ $f : \mathbb{R}^d \to \mathbb{R}$, $g : \mathbb{R}^d \to \overline{\mathbb{R}}$ and $h : \mathbb{R}^m \to \overline{\mathbb{R}}$ are convex, closed, and proper (CCP) functions, where $\overline{\mathbb{R}} \triangleq \mathbb{R}\cup\{+\infty\}$.
- ▷ $\mathbf{A} : \mathbb{R}^d \to \mathbb{R}^m$ is a linear operator ($\mathbf{A}\neq\mathbf{0}$, $\|\mathbf{A}\|=B$).
- ▷ f is differentiable with L-Lipschitz gradient on $\mathbb{R}^d$ ($L>0$).
- ▷ g and h have tractable proximal operators, but $h\circ\mathbf{A}$ **may not**.

Three-Composite Convex Minimization

Consider the following convex three-composite problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left[P(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x}) \right]. \qquad \texttt{(P)}$$

- ▷ $f:\mathbb{R}^d \to \mathbb{R}$, $g:\mathbb{R}^d \to \overline{\mathbb{R}}$ and $h:\mathbb{R}^m \to \overline{\mathbb{R}}$ are convex, closed, and proper (CCP) functions, where $\overline{\mathbb{R}} \triangleq \mathbb{R}\cup\{+\infty\}$.
- ▷ $\mathbf{A}:\mathbb{R}^d \to \mathbb{R}^m$ is a linear operator ($\mathbf{A}\neq\mathbf{0}$, $\|\mathbf{A}\|=B$).
- ▷ f is differentiable with L-Lipschitz gradient on $\mathbb{R}^d$ ($L>0$).
- ▷ g and h have tractable proximal operators, but $h\circ\mathbf{A}$ **may not**.
- ▷ Assume at least one minimizer $\mathbf{x}^*$ exists on $\mathbf{dom}P$.

Stochastic Setting

We focus on the stochastic setting, i.e.,

$$f(\mathbf{x}) \triangleq \mathbb{E}_{\boldsymbol{\xi}\sim\nu}[F(\mathbf{x},\boldsymbol{\xi})]. \tag{1}$$

Stochastic Setting

We focus on the stochastic setting, i.e.,

$$f(\mathbf{x}) \triangleq \mathbb{E}_{\boldsymbol{\xi}\sim\nu}[F(\mathbf{x},\boldsymbol{\xi})]. \tag{1}$$

- $\boldsymbol{\xi}$ is a random vector with distribution ν.

Stochastic Setting

We focus on the stochastic setting, i.e.,

$$f(\mathbf{x}) \triangleq \mathbb{E}_{\boldsymbol{\xi}\sim\nu}[F(\mathbf{x},\boldsymbol{\xi})]. \tag{1}$$

- $\boldsymbol{\xi}$ is a random vector with distribution ν.
- Corresponding to *large-scale* or *online* setting
  - If $\nu = n^{-1}\sum_{i=1}^{n}\delta_{\xi_i}$, then $f(\mathbf{x}) = n^{-1}\sum_{i=1}^{n}F(\mathbf{x},\boldsymbol{\xi}_i)$.

Stochastic Setting

We focus on the stochastic setting, i.e.,

$$f(\mathbf{x}) \triangleq \mathbb{E}_{\boldsymbol{\xi}\sim\nu}[F(\mathbf{x}, \boldsymbol{\xi})]. \tag{1}$$

- $\boldsymbol{\xi}$ is a random vector with distribution ν.
- Corresponding to *large-scale* or *online* setting
  - If $\nu = n^{-1} \sum_{i=1}^{n} \delta_{\xi_i}$, then $f(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} F(\mathbf{x}, \boldsymbol{\xi}_i)$.
- Assume a stochastic first-order oracle $\mathsf{SFO}(f, \sigma)$ that returns an unbiased estimate of $\nabla f(\mathbf{x})$ with variance σ^2, for any $\mathbf{x} \in \mathbf{dom} P$.

Applications

- Constrained Stochastic (Two-)Composite Optimization
  - Dual Soft-Margin Kernelized SVM
  - Two-Stage Stochastic Programming
  - Constrained TV-Denoising/Deblurring

Applications

- ▷ Constrained Stochastic (Two-)Composite Optimization
  - Dual Soft-Margin Kernelized SVM
  - Two-Stage Stochastic Programming
  - Constrained TV-Denoising/Deblurring
- ▷ Three-Composite Expected Risk Minimization
  - Graph-Guided Fused Lasso
  - Graph-Guided Sparse Logistic Regression
  - Robust Matrix Recovery

Saddle-Point Reformulation

$$\boxed{\min_{\mathbf{x}\in\mathbb{R}^d} \; f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})}$$

$\mathbf{prox}_{h\circ\mathbf{A}}$ is intractable in general $\Rightarrow$ Consider saddle-point form

$$\min_{\mathbf{x}\in\mathbb{R}^d} \max_{\mathbf{y}\in\mathbb{R}^m} \left[S(\mathbf{x},\mathbf{y}) \triangleq f(\mathbf{x}) + g(\mathbf{x}) + \langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - h^*(\mathbf{y}) \right]. \qquad (\texttt{SP})$$

Saddle-Point Reformulation

$$\min_{\mathbf{x}\in\mathbb{R}^d} \; f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})$$

$\mathbf{prox}_{h\circ\mathbf{A}}$ is intractable in general $\Rightarrow$ Consider saddle-point form

$$\min_{\mathbf{x}\in\mathbb{R}^d}\max_{\mathbf{y}\in\mathbb{R}^m}\left[S(\mathbf{x},\mathbf{y}) \triangleq f(\mathbf{x}) + g(\mathbf{x}) + \langle \mathbf{A}\mathbf{x},\mathbf{y}\rangle - h^*(\mathbf{y})\right]. \qquad (\texttt{SP})$$

▷ $h^* : \mathbb{R}^m \to \mathbb{R}$ denotes the Fenchel conjugate of h.

Saddle-Point Reformulation

$$\min_{\mathbf{x}\in\mathbb{R}^d} \; f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})$$

$\mathbf{prox}_{h\circ\mathbf{A}}$ is intractable in general $\Rightarrow$ Consider saddle-point form

$$\min_{\mathbf{x}\in\mathbb{R}^d} \max_{\mathbf{y}\in\mathbb{R}^m} \left[S(\mathbf{x},\mathbf{y}) \triangleq f(\mathbf{x}) + g(\mathbf{x}) + \langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - h^*(\mathbf{y}) \right]. \qquad (\texttt{SP})$$

- $h^* : \mathbb{R}^m \to \mathbb{R}$ denotes the Fenchel conjugate of h.
- Under Slater's condition, $\mathbf{x}^*$ is a minimizer of $(\texttt{P}) \Leftrightarrow \exists\, \mathbf{y}^* \in \mathbb{R}^m$ such that $(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point of $(\texttt{SP})$.

Saddle-Point Reformulation

$$\boxed{\min_{\mathbf{x}\in\mathbb{R}^d} \; f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})}$$

$\mathbf{prox}_{h\circ\mathbf{A}}$ is intractable in general $\Rightarrow$ Consider saddle-point form

$$\min_{\mathbf{x}\in\mathbb{R}^d}\max_{\mathbf{y}\in\mathbb{R}^m}\left[S(\mathbf{x},\mathbf{y}) \triangleq f(\mathbf{x}) + g(\mathbf{x}) + \langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - h^*(\mathbf{y})\right]. \qquad (\texttt{SP})$$

- $h^* : \mathbb{R}^m \to \mathbb{R}$ denotes the Fenchel conjugate of h.
- Under Slater's condition, $\mathbf{x}^*$ is a minimizer of ($\texttt{P}$) $\Leftrightarrow \exists\, \mathbf{y}^* \in \mathbb{R}^m$ such that $(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point of ($\texttt{SP}$).
- Develop a primal-dual algorithm for ($\texttt{SP}$).

Existing methods

$$\min_{\mathbf{x}\in\mathbb{R}^d} \; f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})$$

Algorithm	Reference	Type	Convergence Rate[1]	K Known?
Stoc. Subgradient	[Lan12]	Primal	$O\left(\frac{L}{K^2} + \frac{M_g + M_h B + \sigma}{\sqrt{K}}\right)$	Yes
Stoc. F-B Splitting	[DS09]	Primal	$O\left(\frac{\max\{L + M_h B, M_g\} + \sigma}{\sqrt{K}}\right)$	Yes
Stoc. ADMM	[Ouy13]	Primal-Dual	$O\left(\frac{L + M_g}{\sqrt{K}} + \frac{B}{K} + \frac{\sigma}{\sqrt{K}}\right)$	No
Stoc. E-ADMM	[Lin18]	Primal-Dual	$O\left(\frac{L}{K} + \frac{B}{K} + \frac{\sigma^2}{\sqrt{K}}\right)$	No
Stoc. NSPA	[ZK14]	Primal-Dual	$O\left(\frac{L}{K^2} + \frac{M_g^2}{K^{3/2}} + \frac{B^2}{K} + \frac{\sigma}{\sqrt{K}}\right)$	No
Stoc. PD3CM	[ZC18]	Primal-Dual	$O\left(\frac{L}{K} + \frac{B}{K} + \frac{\sigma}{\sqrt{K}}\right)$	Yes

L: Smoothness of f
B: Operator norm of $\mathbf{A}$
M_g: Lipschitz constant of g
M_h: Lipschitz constant of h
σ^2: Variance of stochastic (sub-)gradient
K: Total number of iterations

[1] In terms of expected primal sub-optimality gap or primal-dual gap.

Lower Bound of Convergence Rate

Under $\mathsf{SFO}(f, \sigma)$, when $g \equiv 0$, for any algorithm that solves (**SP**), the convergence rate is no better than[2]

$$\Omega\left(\frac{L}{K^2} + \frac{B}{K} + \frac{\sigma}{\sqrt{K}}\right). \tag{LB}$$

[2] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. "Optimal Primal-Dual Methods for a Class of Saddle Point Problems". In: *SIAM J. Optim.* 24.4 (2014), pp. 1779–1814.

Lower Bound of Convergence Rate

Under $\mathsf{SFO}(f, \sigma)$, when $g \equiv 0$, for any algorithm that solves (**SP**), the convergence rate is no better than[2]

$$\Omega\left(\frac{L}{K^2} + \frac{B}{K} + \frac{\sigma}{\sqrt{K}}\right). \tag{LB}$$

Questions:

- If this lower bound achievable for (**SP**)?
- Can we develop an (minimax) optimal algorithm that achieves (**LB**)?

[2] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. "Optimal Primal-Dual Methods for a Class of Saddle Point Problems". In: *SIAM J. Optim.* 24.4 (2014), pp. 1779–1814.

Lower Bound of Convergence Rate

Under $\mathsf{SFO}(f, \sigma)$, when $g \equiv 0$, for any algorithm that solves (**SP**), the convergence rate is no better than[2]

$$\Omega\left(\frac{L}{K^2} + \frac{B}{K} + \frac{\sigma}{\sqrt{K}}\right). \tag{LB}$$

Questions:

- If this lower bound achievable for (**SP**)?
- Can we develop an (minimax) optimal algorithm that achieves (**LB**)?

Yes, we can!

[2] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. "Optimal Primal-Dual Methods for a Class of Saddle Point Problems". In: *SIAM J. Optim.* 24.4 (2014), pp. 1779–1814.

Algorithm I: An Optimal Algorithm for (SP)

- **Input**: Interpolation sequence $\{\beta_k\}_{k\in\mathbb{Z}^+}$, dual stepsizes $\{\alpha_k\}_{k\in\mathbb{Z}^+}$, primal stepsizes $\{\tau_k\}_{k\in\mathbb{Z}^+}$ and extrapolation sequence $\{\theta_k\}_{k\in\mathbb{Z}^+}$
- **Initialize**: $\mathbf{x}^0 \in \mathbf{dom}\, g$, $\mathbf{y}^0 \in \mathbf{dom}\, h^*$, $\overline{\mathbf{x}}^0 = \mathbf{x}^0$, $\overline{\mathbf{y}}^0 = \mathbf{y}^0$, $\mathbf{z}^0 = \mathbf{x}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

  $\tilde{\mathbf{x}}^k := \beta_k^{-1}\mathbf{x}^k + (1-\beta_k^{-1})\overline{\mathbf{x}}^k$ (Interpolation)

Algorithm I: An Optimal Algorithm for (SP)

- **Input**: Interpolation sequence $\{\beta_k\}_{k\in\mathbb{Z}^+}$, dual stepsizes $\{\alpha_k\}_{k\in\mathbb{Z}^+}$, primal stepsizes $\{\tau_k\}_{k\in\mathbb{Z}^+}$ and extrapolation sequence $\{\theta_k\}_{k\in\mathbb{Z}^+}$
- **Initialize**: $\mathbf{x}^0 \in \mathbf{dom}\, g$, $\mathbf{y}^0 \in \mathbf{dom}\, h^*$, $\overline{\mathbf{x}}^0 = \mathbf{x}^0$, $\overline{\mathbf{y}}^0 = \mathbf{y}^0$, $\mathbf{z}^0 = \mathbf{x}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

  $\tilde{\mathbf{x}}^k := \beta_k^{-1}\mathbf{x}^k + (1-\beta_k^{-1})\overline{\mathbf{x}}^k$ (Interpolation)

  Sample $\boldsymbol{\xi}^k \sim \nu$ and define $\mathbf{v}^k \triangleq \nabla_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\xi}^k)|_{\mathbf{x}=\tilde{\mathbf{x}}^k}$ (Stoc. Gradient)

Algorithm I: An Optimal Algorithm for (SP)

- **Input**: Interpolation sequence $\{\beta_k\}_{k\in\mathbb{Z}^+}$, dual stepsizes $\{\alpha_k\}_{k\in\mathbb{Z}^+}$, primal stepsizes $\{\tau_k\}_{k\in\mathbb{Z}^+}$ and extrapolation sequence $\{\theta_k\}_{k\in\mathbb{Z}^+}$
- **Initialize**: $\mathbf{x}^0 \in \mathbf{dom}\, g$, $\mathbf{y}^0 \in \mathbf{dom}\, h^*$, $\overline{\mathbf{x}}^0 = \mathbf{x}^0$, $\overline{\mathbf{y}}^0 = \mathbf{y}^0$, $\mathbf{z}^0 = \mathbf{x}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

$\tilde{\mathbf{x}}^k := \beta_k^{-1}\mathbf{x}^k + (1-\beta_k^{-1})\overline{\mathbf{x}}^k$ (Interpolation)

Sample $\boldsymbol{\xi}^k \sim \nu$ and define $\mathbf{v}^k \triangleq \nabla_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\xi}^k)|_{\mathbf{x}=\tilde{\mathbf{x}}^k}$ (Stoc. Gradient)

$\mathbf{y}^{k+1} := \mathbf{prox}_{\alpha_k h^*}(\mathbf{y}^k + \alpha_k \mathbf{A}\mathbf{z}^k)$ (Dual Ascent)

Algorithm I: An Optimal Algorithm for (SP)

- **Input**: Interpolation sequence $\{\beta_k\}_{k\in\mathbb{Z}^+}$, dual stepsizes $\{\alpha_k\}_{k\in\mathbb{Z}^+}$, primal stepsizes $\{\tau_k\}_{k\in\mathbb{Z}^+}$ and extrapolation sequence $\{\theta_k\}_{k\in\mathbb{Z}^+}$
- **Initialize**: $\mathbf{x}^0 \in \mathbf{dom}\, g$, $\mathbf{y}^0 \in \mathbf{dom}\, h^*$, $\overline{\mathbf{x}}^0 = \mathbf{x}^0$, $\overline{\mathbf{y}}^0 = \mathbf{y}^0$, $\mathbf{z}^0 = \mathbf{x}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

$$\tilde{\mathbf{x}}^k := \beta_k^{-1}\mathbf{x}^k + (1-\beta_k^{-1})\overline{\mathbf{x}}^k \qquad \text{(Interpolation)}$$

Sample $\boldsymbol{\xi}^k \sim \nu$ and define $\mathbf{v}^k \triangleq \nabla_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\xi}^k)|_{\mathbf{x}=\tilde{\mathbf{x}}^k}$ (Stoc. Gradient)

$$\mathbf{y}^{k+1} := \mathbf{prox}_{\alpha_k h^*}(\mathbf{y}^k + \alpha_k \mathbf{A}\mathbf{z}^k) \qquad \text{(Dual Ascent)}$$

$$\mathbf{x}^{k+1} := \mathbf{prox}_{\tau_k g}(\mathbf{x}^k - \tau_k(\mathbf{A}^T\mathbf{y}^{k+1} + \mathbf{v}^k)) \qquad \text{(Primal Descent)}$$

Algorithm I: An Optimal Algorithm for (SP)

- **Input**: Interpolation sequence $\{\beta_k\}_{k\in\mathbb{Z}^+}$, dual stepsizes $\{\alpha_k\}_{k\in\mathbb{Z}^+}$, primal stepsizes $\{\tau_k\}_{k\in\mathbb{Z}^+}$ and extrapolation sequence $\{\theta_k\}_{k\in\mathbb{Z}^+}$
- **Initialize**: $\mathbf{x}^0 \in \mathbf{dom}\, g$, $\mathbf{y}^0 \in \mathbf{dom}\, h^*$, $\overline{\mathbf{x}}^0 = \mathbf{x}^0$, $\overline{\mathbf{y}}^0 = \mathbf{y}^0$, $\mathbf{z}^0 = \mathbf{x}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

$\tilde{\mathbf{x}}^k := \beta_k^{-1}\mathbf{x}^k + (1-\beta_k^{-1})\overline{\mathbf{x}}^k$ (Interpolation)

Sample $\boldsymbol{\xi}^k \sim \nu$ and define $\mathbf{v}^k \triangleq \nabla_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\xi}^k)|_{\mathbf{x}=\tilde{\mathbf{x}}^k}$ (Stoc. Gradient)

$\mathbf{y}^{k+1} := \mathbf{prox}_{\alpha_k h^*}(\mathbf{y}^k + \alpha_k \mathbf{A}\mathbf{z}^k)$ (Dual Ascent)

$\mathbf{x}^{k+1} := \mathbf{prox}_{\tau_k g}(\mathbf{x}^k - \tau_k(\mathbf{A}^T\mathbf{y}^{k+1} + \mathbf{v}^k))$ (Primal Descent)

$\mathbf{z}^{k+1} := \mathbf{x}^{k+1} + \theta_{k+1}(\mathbf{x}^{k+1} - \mathbf{x}^k)$ (Extrapolation)

Algorithm I: An Optimal Algorithm for ($\mathtt{SP}$)

- **Input**: Interpolation sequence $\{\beta_k\}_{k\in\mathbb{Z}^+}$, dual stepsizes $\{\alpha_k\}_{k\in\mathbb{Z}^+}$, primal stepsizes $\{\tau_k\}_{k\in\mathbb{Z}^+}$ and extrapolation sequence $\{\theta_k\}_{k\in\mathbb{Z}^+}$
- **Initialize**: $\mathbf{x}^0 \in \mathbf{dom}\, g$, $\mathbf{y}^0 \in \mathbf{dom}\, h^*$, $\overline{\mathbf{x}}^0 = \mathbf{x}^0$, $\overline{\mathbf{y}}^0 = \mathbf{y}^0$, $\mathbf{z}^0 = \mathbf{x}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

$$\tilde{\mathbf{x}}^k := \beta_k^{-1}\mathbf{x}^k + (1-\beta_k^{-1})\overline{\mathbf{x}}^k \quad \text{(Interpolation)}$$

Sample $\boldsymbol{\xi}^k \sim \nu$ and define $\mathbf{v}^k \triangleq \nabla_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\xi}^k)|_{\mathbf{x}=\tilde{\mathbf{x}}^k}$ (Stoc. Gradient)

$$\mathbf{y}^{k+1} := \mathbf{prox}_{\alpha_k h^*}(\mathbf{y}^k + \alpha_k \mathbf{A}\mathbf{z}^k) \quad \text{(Dual Ascent)}$$

$$\mathbf{x}^{k+1} := \mathbf{prox}_{\tau_k g}(\mathbf{x}^k - \tau_k(\mathbf{A}^T\mathbf{y}^{k+1} + \mathbf{v}^k)) \quad \text{(Primal Descent)}$$

$$\mathbf{z}^{k+1} := \mathbf{x}^{k+1} + \theta_{k+1}(\mathbf{x}^{k+1} - \mathbf{x}^k) \quad \text{(Extrapolation)}$$

$$\overline{\mathbf{x}}^{k+1} := \beta_k^{-1}\mathbf{x}^{k+1} + (1-\beta_k^{-1})\overline{\mathbf{x}}^k \quad \text{(Primal Averaging)}$$

Algorithm I: An Optimal Algorithm for (SP)

- **Input**: Interpolation sequence $\{\beta_k\}_{k\in\mathbb{Z}^+}$, dual stepsizes $\{\alpha_k\}_{k\in\mathbb{Z}^+}$, primal stepsizes $\{\tau_k\}_{k\in\mathbb{Z}^+}$ and extrapolation sequence $\{\theta_k\}_{k\in\mathbb{Z}^+}$
- **Initialize**: $\mathbf{x}^0 \in \mathbf{dom}\, g$, $\mathbf{y}^0 \in \mathbf{dom}\, h^*$, $\overline{\mathbf{x}}^0 = \mathbf{x}^0$, $\overline{\mathbf{y}}^0 = \mathbf{y}^0$, $\mathbf{z}^0 = \mathbf{x}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

$$\tilde{\mathbf{x}}^k := \beta_k^{-1}\mathbf{x}^k + (1-\beta_k^{-1})\overline{\mathbf{x}}^k \quad \text{(Interpolation)}$$

Sample $\boldsymbol{\xi}^k \sim \nu$ and define $\mathbf{v}^k \triangleq \nabla_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\xi}^k)|_{\mathbf{x}=\tilde{\mathbf{x}}^k}$ (Stoc. Gradient)

$$\mathbf{y}^{k+1} := \mathbf{prox}_{\alpha_k h^*}(\mathbf{y}^k + \alpha_k \mathbf{A}\mathbf{z}^k) \quad \text{(Dual Ascent)}$$

$$\mathbf{x}^{k+1} := \mathbf{prox}_{\tau_k g}(\mathbf{x}^k - \tau_k(\mathbf{A}^T\mathbf{y}^{k+1} + \mathbf{v}^k)) \quad \text{(Primal Descent)}$$

$$\mathbf{z}^{k+1} := \mathbf{x}^{k+1} + \theta_{k+1}(\mathbf{x}^{k+1} - \mathbf{x}^k) \quad \text{(Extrapolation)}$$

$$\overline{\mathbf{x}}^{k+1} := \beta_k^{-1}\mathbf{x}^{k+1} + (1-\beta_k^{-1})\overline{\mathbf{x}}^k \quad \text{(Primal Averaging)}$$

$$\overline{\mathbf{y}}^{k+1} := \beta_k^{-1}\mathbf{y}^{k+1} + (1-\beta_k^{-1})\overline{\mathbf{y}}^k \quad \text{(Dual Averaging)}$$

Algorithm I: An Optimal Algorithm for (SP)

- **Input**: Interpolation sequence $\{\beta_k\}_{k\in\mathbb{Z}^+}$, dual stepsizes $\{\alpha_k\}_{k\in\mathbb{Z}^+}$, primal stepsizes $\{\tau_k\}_{k\in\mathbb{Z}^+}$ and extrapolation sequence $\{\theta_k\}_{k\in\mathbb{Z}^+}$
- **Initialize**: $\mathbf{x}^0 \in \mathbf{dom}\, g$, $\mathbf{y}^0 \in \mathbf{dom}\, h^*$, $\overline{\mathbf{x}}^0 = \mathbf{x}^0$, $\overline{\mathbf{y}}^0 = \mathbf{y}^0$, $\mathbf{z}^0 = \mathbf{x}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

$\tilde{\mathbf{x}}^k := \beta_k^{-1}\mathbf{x}^k + (1-\beta_k^{-1})\overline{\mathbf{x}}^k$ (Interpolation)

Sample $\boldsymbol{\xi}^k \sim \nu$ and define $\mathbf{v}^k \triangleq \nabla_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\xi}^k)|_{\mathbf{x}=\tilde{\mathbf{x}}^k}$ (Stoc. Gradient)

$\mathbf{y}^{k+1} := \mathbf{prox}_{\alpha_k h^*}(\mathbf{y}^k + \alpha_k \mathbf{A}\mathbf{z}^k)$ (Dual Ascent)

$\mathbf{x}^{k+1} := \mathbf{prox}_{\tau_k g}(\mathbf{x}^k - \tau_k(\mathbf{A}^T\mathbf{y}^{k+1} + \mathbf{v}^k))$ (Primal Descent)

$\mathbf{z}^{k+1} := \mathbf{x}^{k+1} + \theta_{k+1}(\mathbf{x}^{k+1} - \mathbf{x}^k)$ (Extrapolation)

$\overline{\mathbf{x}}^{k+1} := \beta_k^{-1}\mathbf{x}^{k+1} + (1-\beta_k^{-1})\overline{\mathbf{x}}^k$ (Primal Averaging)

$\overline{\mathbf{y}}^{k+1} := \beta_k^{-1}\mathbf{y}^{k+1} + (1-\beta_k^{-1})\overline{\mathbf{y}}^k$ (Dual Averaging)

$k := k+1$

Algorithm I: An Optimal Algorithm for (SP)

- **Input**: Interpolation sequence $\{\beta_k\}_{k\in\mathbb{Z}^+}$, dual stepsizes $\{\alpha_k\}_{k\in\mathbb{Z}^+}$, primal stepsizes $\{\tau_k\}_{k\in\mathbb{Z}^+}$ and extrapolation sequence $\{\theta_k\}_{k\in\mathbb{Z}^+}$
- **Initialize**: $\mathbf{x}^0 \in \mathbf{dom}\, g$, $\mathbf{y}^0 \in \mathbf{dom}\, h^*$, $\overline{\mathbf{x}}^0 = \mathbf{x}^0$, $\overline{\mathbf{y}}^0 = \mathbf{y}^0$, $\mathbf{z}^0 = \mathbf{x}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

  $\tilde{\mathbf{x}}^k := \beta_k^{-1}\mathbf{x}^k + (1-\beta_k^{-1})\overline{\mathbf{x}}^k$ (Interpolation)

  Sample $\boldsymbol{\xi}^k \sim \nu$ and define $\mathbf{v}^k \triangleq \nabla_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\xi}^k)|_{\mathbf{x}=\tilde{\mathbf{x}}^k}$ (Stoc. Gradient)

  $\mathbf{y}^{k+1} := \mathbf{prox}_{\alpha_k h^*}(\mathbf{y}^k + \alpha_k \mathbf{A}\mathbf{z}^k)$ (Dual Ascent)

  $\mathbf{x}^{k+1} := \mathbf{prox}_{\tau_k g}(\mathbf{x}^k - \tau_k(\mathbf{A}^T\mathbf{y}^{k+1} + \mathbf{v}^k))$ (Primal Descent)

  $\mathbf{z}^{k+1} := \mathbf{x}^{k+1} + \theta_{k+1}(\mathbf{x}^{k+1} - \mathbf{x}^k)$ (Extrapolation)

  $\overline{\mathbf{x}}^{k+1} := \beta_k^{-1}\mathbf{x}^{k+1} + (1-\beta_k^{-1})\overline{\mathbf{x}}^k$ (Primal Averaging)

  $\overline{\mathbf{y}}^{k+1} := \beta_k^{-1}\mathbf{y}^{k+1} + (1-\beta_k^{-1})\overline{\mathbf{y}}^k$ (Dual Averaging)

  $k := k+1$
- **Output**: $(\overline{\mathbf{x}}^k, \overline{\mathbf{y}}^k)$

Choice of Parameters

For any constants $\rho, \rho' > 0$ and $k \in \mathbb{Z}^+$,

Choice of Parameters

For any constants $\rho, \rho' > 0$ and $k \in \mathbb{Z}^+$,

- Interpolation parameter

$$\beta_k = \frac{(k+1)(k+4)}{2(k+2)} = \Theta(k)$$

Choice of Parameters

For any constants $\rho, \rho' > 0$ and $k \in \mathbb{Z}^+$,

- Interpolation parameter

$$\beta_k = \frac{(k+1)(k+4)}{2(k+2)} = \Theta(k)$$

- Extrapolation parameter

$$\theta_k = \frac{k+1}{k+2} = \Theta(1)$$

Choice of Parameters

For any constants $\rho, \rho' > 0$ and $k \in \mathbb{Z}^+$,

- Interpolation parameter

$$\beta_k = \frac{(k+1)(k+4)}{2(k+2)} = \Theta(k)$$

- Extrapolation parameter

$$\theta_k = \frac{k+1}{k+2} = \Theta(1)$$

- Primal stepsize

$$\tau_k^{-1} = \frac{4L}{k+2} + 2\rho' B + \rho\sigma\sqrt{k+2}$$
$$= \Theta(L/k + B + \sigma\sqrt{k})$$

Choice of Parameters

For any constants $\rho, \rho' > 0$ and $k \in \mathbb{Z}^+$,

▷ Interpolation parameter

$$\beta_k = \frac{(k+1)(k+4)}{2(k+2)} = \Theta(k)$$

▷ Extrapolation parameter

$$\theta_k = \frac{k+1}{k+2} = \Theta(1)$$

▷ Primal stepsize

$$\tau_k^{-1} = \frac{4L}{k+2} + 2\rho' B + \rho\sigma\sqrt{k+2}$$
$$= \Theta(L/k + B + \sigma\sqrt{k})$$

▷ Dual stepsize

$$\alpha_k = \frac{\rho'}{B}$$

Choice of Parameters

For any constants $\rho, \rho' > 0$ and $k \in \mathbb{Z}^+$,

- Interpolation parameter

$$\beta_k = \frac{(k+1)(k+4)}{2(k+2)} = \Theta(k)$$

- Primal stepsize

$$\tau_k^{-1} = \frac{4L}{k+2} + 2\rho' B + \rho\sigma\sqrt{k+2} = \Theta(L/k + B + \sigma\sqrt{k})$$

- Extrapolation parameter

$$\theta_k = \frac{k+1}{k+2} = \Theta(1)$$

- Dual stepsize

$$\alpha_k = \frac{\rho'}{B}$$

The convergence rate of our algorithm matches the lower bound (**LB**) for any values of ρ and ρ'.

Extension to Multi-Composite Problems

$$\boxed{\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + g(\mathbf{x}) + \sum_{i=1}^{p} h_i(\mathbf{A}_i\mathbf{x})} \quad \Longrightarrow \text{Product-Space Technique}$$

Extension to Multi-Composite Problems

$$\boxed{\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + g(\mathbf{x}) + \sum_{i=1}^{p} h_i(\mathbf{A}_i\mathbf{x})} \quad \Longrightarrow \text{Product-Space Technique}$$

- ▷ Introduce multiple dual variables $\{\mathbf{y}_i\}_{i=1}^{p}$.

Extension to Multi-Composite Problems

$$\boxed{\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + g(\mathbf{x}) + \textstyle\sum_{i=1}^{p} h_i(\mathbf{A}_i\mathbf{x})} \quad \Longrightarrow \text{Product-Space Technique}$$

- Introduce multiple dual variables $\{\mathbf{y}_i\}_{i=1}^{p}$.
- Steps to change

For each $i \in [p]$, perform (in parallel)

$$\mathbf{y}_i^{k+1} := \mathbf{prox}_{\alpha_k h_i^*}(\mathbf{y}_i^k + \alpha_k \mathbf{A}_i \mathbf{z}^k) \qquad \text{(Dual Ascent)}$$

Extension to Multi-Composite Problems

$$\boxed{\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + g(\mathbf{x}) + \sum_{i=1}^{p} h_i(\mathbf{A}_i\mathbf{x})} \implies \text{Product-Space Technique}$$

- Introduce multiple dual variables $\{\mathbf{y}_i\}_{i=1}^{p}$.
- Steps to change

For each $i \in [p]$, perform (in parallel)

$$\mathbf{y}_i^{k+1} := \mathbf{prox}_{\alpha_k h_i^*}(\mathbf{y}_i^k + \alpha_k \mathbf{A}_i \mathbf{z}^k) \qquad \text{(Dual Ascent)}$$

$$\overline{\mathbf{y}}_i^{k+1} := \beta_k^{-1}\mathbf{y}_i^{k+1} + (1-\beta_k^{-1})\overline{\mathbf{y}}_i^k \qquad \text{(Dual Averaging)}$$

Extension to Multi-Composite Problems

$$\boxed{\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + g(\mathbf{x}) + \textstyle\sum_{i=1}^{p} h_i(\mathbf{A}_i\mathbf{x})} \quad \Longrightarrow \text{Product-Space Technique}$$

- Introduce multiple dual variables $\{\mathbf{y}_i\}_{i=1}^{p}$.
- Steps to change

For each $i \in [p]$, perform (in parallel)

$$\mathbf{y}_i^{k+1} := \mathbf{prox}_{\alpha_k h_i^*}(\mathbf{y}_i^k + \alpha_k \mathbf{A}_i \mathbf{z}^k) \qquad \text{(Dual Ascent)}$$

$$\overline{\mathbf{y}}_i^{k+1} := \beta_k^{-1}\mathbf{y}_i^{k+1} + (1-\beta_k^{-1})\overline{\mathbf{y}}_i^k \qquad \text{(Dual Averaging)}$$

$$\mathbf{x}^{k+1} := \mathbf{prox}_{\tau_k g}(\mathbf{x}^k - \tau_k(\textstyle\sum_{i=1}^{p}\mathbf{A}_i^T\mathbf{y}_i^{k+1} + \mathbf{v}^k)) \qquad \text{(Primal Descent)}$$

Extension to Multi-Composite Problems

$$\boxed{\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + g(\mathbf{x}) + \textstyle\sum_{i=1}^{p} h_i(\mathbf{A}_i\mathbf{x})} \quad \Longrightarrow \text{Product-Space Technique}$$

- ▷ Introduce multiple dual variables $\{\mathbf{y}_i\}_{i=1}^{p}$.
- ▷ Steps to change

For each $i \in [p]$, perform (in parallel)

$$\mathbf{y}_i^{k+1} := \mathbf{prox}_{\alpha_k h_i^*}(\mathbf{y}_i^k + \alpha_k \mathbf{A}_i \mathbf{z}^k) \qquad \text{(Dual Ascent)}$$

$$\overline{\mathbf{y}}_i^{k+1} := \beta_k^{-1}\mathbf{y}_i^{k+1} + (1-\beta_k^{-1})\overline{\mathbf{y}}_i^k \qquad \text{(Dual Averaging)}$$

$$\mathbf{x}^{k+1} := \mathbf{prox}_{\tau_k g}(\mathbf{x}^k - \tau_k(\textstyle\sum_{i=1}^{p}\mathbf{A}_i^T\mathbf{y}_i^{k+1} + \mathbf{v}^k)) \qquad \text{(Primal Descent)}$$

- ▷ For large p, can further introduce randomization on the dual update and averaging steps.

Convergence Analysis – Preliminaries

Definition 1

Convergence Analysis – Preliminaries

Definition 1

- Stochastic noise $\boldsymbol{\varepsilon}^k \triangleq \mathbf{v}^k - \nabla f(\tilde{\mathbf{x}}^k)$.

Convergence Analysis – Preliminaries

Definition 1

- Stochastic noise $\boldsymbol{\varepsilon}^k \triangleq \mathbf{v}^k - \nabla f(\tilde{\mathbf{x}}^k)$.
- Filtration $\{\mathcal{F}_k\}_{k\in\mathbb{Z}^+}$ s.t. $\mathcal{F}_0 \triangleq \emptyset$ and $\mathcal{F}_k = \sigma\left(\{\boldsymbol{\xi}^i\}_{i=0}^{k-1}\right)$.

Convergence Analysis – Preliminaries

Definition 1

- Stochastic noise $\boldsymbol{\varepsilon}^k \triangleq \mathbf{v}^k - \nabla f(\tilde{\mathbf{x}}^k)$.
- Filtration $\{\mathcal{F}_k\}_{k \in \mathbb{Z}^+}$ s.t. $\mathcal{F}_0 \triangleq \emptyset$ and $\mathcal{F}_k = \sigma\left(\{\boldsymbol{\xi}^i\}_{i=0}^{k-1}\right)$.
- $D_g \triangleq \sup_{\mathbf{x},\mathbf{x}' \in \mathbf{dom}\, g} \|\mathbf{x} - \mathbf{x}'\|$ and D_{h^*} similarly.

Convergence Analysis – Preliminaries

Definition 1

- Stochastic noise $\boldsymbol{\varepsilon}^k \triangleq \mathbf{v}^k - \nabla f(\tilde{\mathbf{x}}^k)$.
- Filtration $\{\mathcal{F}_k\}_{k\in\mathbb{Z}^+}$ s.t. $\mathcal{F}_0 \triangleq \emptyset$ and $\mathcal{F}_k = \sigma\left(\{\boldsymbol{\xi}^i\}_{i=0}^{k-1}\right)$.
- $D_g \triangleq \sup_{\mathbf{x},\mathbf{x}'\in\mathbf{dom}\, g} \|\mathbf{x}-\mathbf{x}'\|$ and D_{h^*} similarly.
- Primal-dual gap

$$G(\mathbf{x},\mathbf{y}) \triangleq \sup_{\mathbf{y}'\in\mathbf{dom}\, h^*} S(\mathbf{x},\mathbf{y}') - \inf_{\mathbf{x}'\in\mathbf{dom}\, g} S(\mathbf{x}',\mathbf{y}).$$

Convergence Analysis – Preliminaries

Definition 1

- Stochastic noise $\boldsymbol{\varepsilon}^k \triangleq \mathbf{v}^k - \nabla f(\tilde{\mathbf{x}}^k)$.
- Filtration $\{\mathcal{F}_k\}_{k\in\mathbb{Z}^+}$ s.t. $\mathcal{F}_0 \triangleq \emptyset$ and $\mathcal{F}_k = \sigma\left(\{\boldsymbol{\xi}^i\}_{i=0}^{k-1}\right)$.
- $D_g \triangleq \sup_{\mathbf{x},\mathbf{x}'\in\mathbf{dom}\,g} \|\mathbf{x}-\mathbf{x}'\|$ and D_{h^*} similarly.
- Primal-dual gap

$$G(\mathbf{x},\mathbf{y}) \triangleq \sup_{\mathbf{y}'\in\mathbf{dom}\,h^*} S(\mathbf{x},\mathbf{y}') - \inf_{\mathbf{x}'\in\mathbf{dom}\,g} S(\mathbf{x}',\mathbf{y}).$$

Assumptions

Convergence Analysis – Preliminaries

Definition 1

- Stochastic noise $\boldsymbol{\varepsilon}^k \triangleq \mathbf{v}^k - \nabla f(\tilde{\mathbf{x}}^k)$.
- Filtration $\{\mathcal{F}_k\}_{k\in\mathbb{Z}^+}$ s.t. $\mathcal{F}_0 \triangleq \emptyset$ and $\mathcal{F}_k = \sigma\left(\{\boldsymbol{\xi}^i\}_{i=0}^{k-1}\right)$.
- $D_g \triangleq \sup_{\mathbf{x},\mathbf{x}'\in\mathbf{dom}\,g} \|\mathbf{x}-\mathbf{x}'\|$ and D_{h^*} similarly.
- Primal-dual gap

$$G(\mathbf{x},\mathbf{y}) \triangleq \sup_{\mathbf{y}'\in\mathbf{dom}\,h^*} S(\mathbf{x},\mathbf{y}') - \inf_{\mathbf{x}'\in\mathbf{dom}\,g} S(\mathbf{x}',\mathbf{y}).$$

Assumptions

A1 $\mathbb{E}_{\boldsymbol{\xi}^k}\left[\boldsymbol{\varepsilon}^k \,|\, \mathcal{F}_k\right] = 0$ *almost surely (a.s.).*

Convergence Analysis – Preliminaries

Definition 1

- Stochastic noise $\boldsymbol{\varepsilon}^k \triangleq \mathbf{v}^k - \nabla f(\tilde{\mathbf{x}}^k)$.
- Filtration $\{\mathcal{F}_k\}_{k \in \mathbb{Z}^+}$ s.t. $\mathcal{F}_0 \triangleq \emptyset$ and $\mathcal{F}_k = \sigma\left(\{\boldsymbol{\xi}^i\}_{i=0}^{k-1}\right)$.
- $D_g \triangleq \sup_{\mathbf{x},\mathbf{x}' \in \mathbf{dom}\, g} \|\mathbf{x} - \mathbf{x}'\|$ and D_{h^*} similarly.
- Primal-dual gap

$$G(\mathbf{x}, \mathbf{y}) \triangleq \sup_{\mathbf{y}' \in \mathbf{dom}\, h^*} S(\mathbf{x}, \mathbf{y}') - \inf_{\mathbf{x}' \in \mathbf{dom}\, g} S(\mathbf{x}', \mathbf{y}).$$

Assumptions

A1 $\mathbb{E}_{\boldsymbol{\xi}^k}\left[\boldsymbol{\varepsilon}^k \,|\, \mathcal{F}_k\right] = 0$ *almost surely (a.s.).*

A2 $\mathbb{E}_{\boldsymbol{\xi}^k}\left[\|\boldsymbol{\varepsilon}^k\|^2 \,\middle|\, \mathcal{F}_k\right] \le \sigma^2$ *a.s.*

Convergence Analysis – Preliminaries

Definition 1

- Stochastic noise $\boldsymbol{\varepsilon}^k \triangleq \mathbf{v}^k - \nabla f(\tilde{\mathbf{x}}^k)$.
- Filtration $\{\mathcal{F}_k\}_{k\in\mathbb{Z}^+}$ s.t. $\mathcal{F}_0 \triangleq \emptyset$ and $\mathcal{F}_k = \sigma\left(\{\boldsymbol{\xi}^i\}_{i=0}^{k-1}\right)$.
- $D_g \triangleq \sup_{\mathbf{x},\mathbf{x}'\in\mathbf{dom}\, g} \|\mathbf{x}-\mathbf{x}'\|$ and D_{h^*} similarly.
- Primal-dual gap

$$G(\mathbf{x},\mathbf{y}) \triangleq \sup_{\mathbf{y}'\in\mathbf{dom}\, h^*} S(\mathbf{x},\mathbf{y}') - \inf_{\mathbf{x}'\in\mathbf{dom}\, g} S(\mathbf{x}',\mathbf{y}).$$

Assumptions

A1 $\mathbb{E}_{\boldsymbol{\xi}^k}\left[\boldsymbol{\varepsilon}^k \,\middle|\, \mathcal{F}_k\right] = 0$ *almost surely (a.s.).*

A2 $\mathbb{E}_{\boldsymbol{\xi}^k}\left[\|\boldsymbol{\varepsilon}^k\|^2 \,\middle|\, \mathcal{F}_k\right] \le \sigma^2$ *a.s.*

A3 $\mathbb{E}_{\boldsymbol{\xi}^k}\left[\exp\left\{\varsigma\|\boldsymbol{\varepsilon}^k\|^2/\sigma^2\right\} \,\middle|\, \mathcal{F}_k\right] \le \exp\{\varsigma^2+\varsigma\}$ *a.s.*

An Important Lemma

Lemma 2 (Z.-Haskell-Tan, 2018)

Let **dom** g *be compact and* **dom** h^* *be bounded. In Algorithm I, let* $\beta_0 = 1$,

$$\beta_{k-1}\theta_k + 1 = \beta_k, \forall\, k \in \mathbb{Z}^+,$$
$$0 < \theta_k \leq \min\{\tau_{k-1}/\tau_k, \alpha_{k-1}/\alpha_k\}, \forall\, k \in \mathbb{N},$$
$$B^2\alpha_{k-1} + L/\beta_{k-1} \leq (1-\zeta)/\tau_{k-1}, \forall\, k \in \mathbb{N},$$

for some $\zeta \in (0, 1)$.

An Important Lemma

Lemma 2 (Z.-Haskell-Tan, 2018)

Let $\mathbf{dom}\, g$ *be compact and* $\mathbf{dom}\, h^*$ *be bounded. In Algorithm I, let* $\beta_0 = 1$,

$$\beta_{k-1}\theta_k + 1 = \beta_k, \forall\, k \in \mathbb{Z}^+,$$
$$0 < \theta_k \le \min\{\tau_{k-1}/\tau_k, \alpha_{k-1}/\alpha_k\}, \forall\, k \in \mathbb{N},$$
$$B^2\alpha_{k-1} + L/\beta_{k-1} \le (1-\zeta)/\tau_{k-1}, \forall\, k \in \mathbb{N},$$

for some $\zeta \in (0,1)$.

Define $\Gamma_K \triangleq \sum_{k=0}^{K-1} \gamma_k \tau_k$ *and* $\Gamma'_K \triangleq (\sum_{k=0}^{K-1} \gamma_k^2)^{1/2}$. *If* ***A1*** *and* ***A2*** *hold, then*

$$\mathbb{E}_{\Xi_K}\left[G(\overline{\mathbf{x}}^K, \overline{\mathbf{y}}^K)\right] \le \frac{D_g^2}{\beta_{K-1}\tau_{K-1}} + \frac{D_{h^*}^2}{2\beta_{K-1}\alpha_{K-1}} + \frac{(1+\zeta)\Gamma_K}{2\zeta\beta_{K-1}\gamma_{K-1}}\sigma^2,\ \forall\, K \in \mathbb{N}.$$

An Important Lemma (Cont'd)

*Also, if **A1** and **A3** hold, then for any* $\delta \in (0,1)$*,*

$$G(\overline{\mathbf{x}}^K, \overline{\mathbf{y}}^K) \leq \frac{1}{\beta_{K-1}} \left\{ \frac{4\sqrt{\log(2/\delta)}D_g}{\gamma_{K-1}} \Gamma'_K \sigma + \frac{D_g^2}{\tau_{K-1}} + \frac{D_{h^*}^2}{2\alpha_{K-1}} + \frac{1 + 2\sqrt{\log(2/\delta)}}{2\zeta\gamma_{K-1}(1+\zeta)^{-1}} \Gamma_K \sigma^2 \right\}$$

with probability (w.p.) at least $1 - \delta$*.*

Main Results

Theorem 3 (Z.-Haskell-Tan, 2018)

Let $\mathbf{dom}\, g$ *be compact,* $\mathbf{dom}\, h^*$ *be bounded and* $(\overline{\mathbf{x}}^K, \overline{\mathbf{y}}^K)$ *be produced by Algorithm I. If* ***A1*** *and* ***A2*** *hold, then for any* $K \in \mathbb{N}$*,*

$$\mathbb{E}_{\Xi_K}\left[G(\overline{\mathbf{x}}^K, \overline{\mathbf{y}}^K)\right] \leq \frac{8L}{K(K+3)} D_g^2 + \frac{4B}{K}\left(\rho' D_g^2 + \frac{D_{h^*}^2}{4\rho'}\right) + \frac{4\sigma}{\sqrt{K+3}}\left(\rho D_g^2 + \frac{2}{\rho}\right).$$

$$= O\left(\frac{L}{K^2} + \frac{B}{K} + \frac{\sigma}{\sqrt{K}}\right)$$

Main Results (Cont'd)

*In addition, if **A1** and **A3** hold, then for any* $\delta \in (0,1)$*,*

$$\begin{aligned} G(\overline{\mathbf{x}}^K, \overline{\mathbf{y}}^K) &\leq \frac{8L}{K(K+3)} D_g^2 + \frac{4B}{K}\left(\rho' D_g^2 + \frac{D_{h^*}^2}{4\rho'}\right) \\ &\quad + \frac{16\sigma}{\sqrt{K+3}}\left(D_g + \frac{2}{\rho}\right)\sqrt{\log(2/\delta)} \\ &= O\left(\frac{L}{K^2} + \frac{B}{K} + \frac{\sigma\sqrt{\log(1/\delta)}}{\sqrt{K}}\right) \end{aligned} \tag{2}$$

w.p. at least $1-\delta$.

Choose ρ and ρ'

If D_g and D_{h^*} are known or can be estimated, then we can choose $\rho' = D_{h^*}/(2D_g)$ and $\rho = 2/D_g$. As a result,

$$\mathbb{E}_{\Xi_K}\left[G(\overline{\mathbf{x}}^K, \overline{\mathbf{y}}^K)\right] \leq \frac{8L}{K(K+3)} D_g^2 + \frac{4B}{K} D_g D_{h^*} + \frac{12\sigma}{\sqrt{K+3}} D_g$$

and for any $\delta \in (0,1)$, w.p. at least $1-\delta$,

$$G(\overline{\mathbf{x}}^K, \overline{\mathbf{y}}^K) \leq \frac{8L}{K(K+3)} D_g^2 + \frac{4B}{K} D_g D_{h^*} + \frac{32\sigma}{\sqrt{K+3}} \sqrt{\log(2/\delta)} D_g.$$

Constrained Minimization Reformulation

$$\boxed{\min_{\mathbf{u}\in\mathbb{R}^d,\boldsymbol{\omega}\in\mathbb{R}^m} \; f(\mathbf{u}) + g(\mathbf{u}) + h(\boldsymbol{\omega}) \quad \text{s.t.} \quad \mathbf{A}\mathbf{u} = \boldsymbol{\omega}} \tag{\texttt{CSP}}$$

Constrained Minimization Reformulation

$$\boxed{\min_{\mathbf{u}\in\mathbb{R}^d,\boldsymbol{\omega}\in\mathbb{R}^m} \quad f(\mathbf{u}) + g(\mathbf{u}) + h(\boldsymbol{\omega}) \quad \text{s.t.} \quad \mathbf{A}\mathbf{u} = \boldsymbol{\omega}} \qquad (\texttt{CSP})$$

When $g \equiv 0$:

- ▷ Many stochastic ADMM algorithms proposed [Ouy13; Suz13; AS14].
- ▷ The algorithm in [AS14] obtains the optimal convergence rate
  $\longrightarrow$ The convergence rate of the smooth term f is $O(L/K^2)$.

Constrained Minimization Reformulation

$$\boxed{\min_{\mathbf{u}\in\mathbb{R}^d,\boldsymbol{\omega}\in\mathbb{R}^m} \quad f(\mathbf{u}) + g(\mathbf{u}) + h(\boldsymbol{\omega}) \quad \text{s.t.} \quad \mathbf{A}\mathbf{u} = \boldsymbol{\omega}} \qquad (\texttt{CSP})$$

When $g \equiv 0$:

- ▷ Many stochastic ADMM algorithms proposed [Ouy13; Suz13; AS14].
- ▷ The algorithm in [AS14] obtains the optimal convergence rate
  $\longrightarrow$ The convergence rate of the smooth term f is $O(L/K^2)$.

When g is CCP:

- ▷ How to design an optimal ADMM algorithm for ($\texttt{CSP}$)?
- ▷ Moreover, how is it related to Algorithm I?

Algorithm II: A Stochastic ADMM Algorithm

- **Define**: $L_k^{\varrho}(\mathbf{u},\boldsymbol{\omega},\boldsymbol{\lambda}) \triangleq f(\mathbf{u}^k) + \langle \mathbf{v}^k, \mathbf{u}-\mathbf{u}^k\rangle + g(\mathbf{u}) + h(\boldsymbol{\omega}) - \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{u}-\boldsymbol{\omega}\rangle$
$+r_k(2\eta_k)^{-1}\langle \mathbf{u}-\mathbf{u}^k, \mathbf{W}^k(\mathbf{u}-\mathbf{u}^k)\rangle + (\varrho/2)\|\mathbf{A}\mathbf{u}-\boldsymbol{\omega}\|^2$

Algorithm II: A Stochastic ADMM Algorithm

- **Define**: $L_k^{\varrho}(\mathbf{u}, \boldsymbol{\omega}, \boldsymbol{\lambda}) \triangleq f(\mathbf{u}^k) + \langle \mathbf{v}^k, \mathbf{u} - \mathbf{u}^k \rangle + g(\mathbf{u}) + h(\boldsymbol{\omega}) - \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{u} - \boldsymbol{\omega} \rangle$
$+ r_k (2\eta_k)^{-1} \langle \mathbf{u} - \mathbf{u}^k, \mathbf{W}^k (\mathbf{u} - \mathbf{u}^k) \rangle + (\varrho/2) \|\mathbf{A}\mathbf{u} - \boldsymbol{\omega}\|^2$
- **Input**: Interpolation sequence $\{r_k\}_{k \in \mathbb{Z}^+}$, stepsizes $\{\eta_k\}_{k \in \mathbb{Z}^+}$, penalty parameter $\varrho > 0$

Algorithm II: A Stochastic ADMM Algorithm

- **Define**: $L_k^{\varrho}(\mathbf{u},\boldsymbol{\omega},\boldsymbol{\lambda}) \triangleq f(\mathbf{u}^k) + \langle \mathbf{v}^k, \mathbf{u}-\mathbf{u}^k\rangle + g(\mathbf{u}) + h(\boldsymbol{\omega}) - \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{u}-\boldsymbol{\omega}\rangle$
  $+ r_k(2\eta_k)^{-1}\langle \mathbf{u}-\mathbf{u}^k, \mathbf{W}^k(\mathbf{u}-\mathbf{u}^k)\rangle + (\varrho/2)\|\mathbf{A}\mathbf{u}-\boldsymbol{\omega}\|^2$
- **Input**: Interpolation sequence $\{r_k\}_{k\in\mathbb{Z}^+}$, stepsizes $\{\eta_k\}_{k\in\mathbb{Z}^+}$, penalty parameter $\varrho > 0$
- **Initialize**: $\mathbf{u}^0 \in \mathbf{dom}\, g$, $\boldsymbol{\omega}^0 \in \mathbf{dom}\, h$, $\boldsymbol{\lambda}^0 \in \mathbb{R}^m$, $\overline{\mathbf{u}}^0 = \mathbf{u}^0$, $\overline{\boldsymbol{\omega}}^0 = \boldsymbol{\omega}^0$, $\overline{\boldsymbol{\lambda}}^0 = \boldsymbol{\lambda}^0$, $k = 0$

Algorithm II: A Stochastic ADMM Algorithm

- **Define**: $L_k^{\varrho}(\mathbf{u},\boldsymbol{\omega},\boldsymbol{\lambda}) \triangleq f(\mathbf{u}^k) + \langle \mathbf{v}^k, \mathbf{u}-\mathbf{u}^k\rangle + g(\mathbf{u}) + h(\boldsymbol{\omega}) - \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{u}-\boldsymbol{\omega}\rangle$
  $+ r_k(2\eta_k)^{-1}\langle \mathbf{u}-\mathbf{u}^k, \mathbf{W}^k(\mathbf{u}-\mathbf{u}^k)\rangle + (\varrho/2)\|\mathbf{A}\mathbf{u}-\boldsymbol{\omega}\|^2$
- **Input**: Interpolation sequence $\{r_k\}_{k\in\mathbb{Z}^+}$, stepsizes $\{\eta_k\}_{k\in\mathbb{Z}^+}$, penalty parameter $\varrho > 0$
- **Initialize**: $\mathbf{u}^0 \in \mathbf{dom}\, g$, $\boldsymbol{\omega}^0 \in \mathbf{dom}\, h$, $\boldsymbol{\lambda}^0 \in \mathbb{R}^m$, $\overline{\mathbf{u}}^0 = \mathbf{u}^0$, $\overline{\boldsymbol{\omega}}^0 = \boldsymbol{\omega}^0$, $\overline{\boldsymbol{\lambda}}^0 = \boldsymbol{\lambda}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

Algorithm II: A Stochastic ADMM Algorithm

- **Define**: $L_k^{\varrho}(\mathbf{u},\boldsymbol{\omega},\boldsymbol{\lambda}) \triangleq f(\mathbf{u}^k) + \langle \mathbf{v}^k, \mathbf{u}-\mathbf{u}^k\rangle + g(\mathbf{u}) + h(\boldsymbol{\omega}) - \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{u}-\boldsymbol{\omega}\rangle$ $+ r_k(2\eta_k)^{-1}\langle \mathbf{u}-\mathbf{u}^k, \mathbf{W}^k(\mathbf{u}-\mathbf{u}^k)\rangle + (\varrho/2)\|\mathbf{A}\mathbf{u}-\boldsymbol{\omega}\|^2$
- **Input**: Interpolation sequence $\{r_k\}_{k\in\mathbb{Z}^+}$, stepsizes $\{\eta_k\}_{k\in\mathbb{Z}^+}$, penalty parameter $\varrho > 0$
- **Initialize**: $\mathbf{u}^0 \in \mathbf{dom}\, g$, $\boldsymbol{\omega}^0 \in \mathbf{dom}\, h$, $\boldsymbol{\lambda}^0 \in \mathbb{R}^m$, $\overline{\mathbf{u}}^0 = \mathbf{u}^0$, $\overline{\boldsymbol{\omega}}^0 = \boldsymbol{\omega}^0$, $\overline{\boldsymbol{\lambda}}^0 = \boldsymbol{\lambda}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

$$\tilde{\mathbf{u}}^k := r_k\mathbf{u}^k + (1-r_k)\overline{\mathbf{u}}^k \qquad \text{(Interpolation)}$$

Algorithm II: A Stochastic ADMM Algorithm

- **Define**: $L_k^{\varrho}(\mathbf{u}, \boldsymbol{\omega}, \boldsymbol{\lambda}) \triangleq f(\mathbf{u}^k) + \langle \mathbf{v}^k, \mathbf{u} - \mathbf{u}^k \rangle + g(\mathbf{u}) + h(\boldsymbol{\omega}) - \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{u} - \boldsymbol{\omega} \rangle$
$+ r_k (2\eta_k)^{-1} \langle \mathbf{u} - \mathbf{u}^k, \mathbf{W}^k (\mathbf{u} - \mathbf{u}^k) \rangle + (\varrho/2) \| \mathbf{A}\mathbf{u} - \boldsymbol{\omega} \|^2$
- **Input**: Interpolation sequence $\{r_k\}_{k \in \mathbb{Z}^+}$, stepsizes $\{\eta_k\}_{k \in \mathbb{Z}^+}$, penalty parameter $\varrho > 0$
- **Initialize**: $\mathbf{u}^0 \in \mathbf{dom}\, g$, $\boldsymbol{\omega}^0 \in \mathbf{dom}\, h$, $\boldsymbol{\lambda}^0 \in \mathbb{R}^m$, $\overline{\mathbf{u}}^0 = \mathbf{u}^0$, $\overline{\boldsymbol{\omega}}^0 = \boldsymbol{\omega}^0$, $\overline{\boldsymbol{\lambda}}^0 = \boldsymbol{\lambda}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

$$\tilde{\mathbf{u}}^k := r_k \mathbf{u}^k + (1 - r_k) \overline{\mathbf{u}}^k \qquad \text{(Interpolation)}$$

Sample $\tilde{\boldsymbol{\xi}}^k \sim \nu$ and define $\tilde{\mathbf{v}}^k \triangleq \nabla_{\mathbf{u}} F(\mathbf{u}, \tilde{\boldsymbol{\xi}}^k)|_{\mathbf{u} = \tilde{\mathbf{u}}^k}$

Algorithm II: A Stochastic ADMM Algorithm

- **Define**: $L_k^{\varrho}(\mathbf{u},\boldsymbol{\omega},\boldsymbol{\lambda}) \triangleq f(\mathbf{u}^k) + \langle \mathbf{v}^k, \mathbf{u}-\mathbf{u}^k\rangle + g(\mathbf{u}) + h(\boldsymbol{\omega}) - \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{u}-\boldsymbol{\omega}\rangle$
  $+r_k(2\eta_k)^{-1}\langle \mathbf{u}-\mathbf{u}^k, \mathbf{W}^k(\mathbf{u}-\mathbf{u}^k)\rangle + (\varrho/2)\|\mathbf{A}\mathbf{u}-\boldsymbol{\omega}\|^2$
- **Input**: Interpolation sequence $\{r_k\}_{k\in\mathbb{Z}^+}$, stepsizes $\{\eta_k\}_{k\in\mathbb{Z}^+}$, penalty parameter $\varrho > 0$
- **Initialize**: $\mathbf{u}^0 \in \mathbf{dom}\, g$, $\boldsymbol{\omega}^0 \in \mathbf{dom}\, h$, $\boldsymbol{\lambda}^0 \in \mathbb{R}^m$, $\overline{\mathbf{u}}^0 = \mathbf{u}^0$, $\overline{\boldsymbol{\omega}}^0 = \boldsymbol{\omega}^0$, $\overline{\boldsymbol{\lambda}}^0 = \boldsymbol{\lambda}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

$$\tilde{\mathbf{u}}^k := r_k\mathbf{u}^k + (1-r_k)\overline{\mathbf{u}}^k \qquad \text{(Interpolation)}$$

Sample $\tilde{\boldsymbol{\xi}}^k \sim \nu$ and define $\tilde{\mathbf{v}}^k \triangleq \nabla_{\mathbf{u}} F(\mathbf{u}, \tilde{\boldsymbol{\xi}}^k)|_{\mathbf{u}=\tilde{\mathbf{u}}^k}$

$$\boldsymbol{\omega}^{k+1} := \operatorname*{arg\,min}_{\boldsymbol{\omega}\in\mathbf{dom}\, h} L_k^{\varrho}(\mathbf{u}^k, \boldsymbol{\omega}, \boldsymbol{\lambda}^k)$$

$$\mathbf{u}^{k+1} := \operatorname*{arg\,min}_{\mathbf{u}\in\mathbf{dom}\, g} L_k^{\varrho}(\mathbf{u}, \boldsymbol{\omega}^{k+1}, \boldsymbol{\lambda}^k) \qquad \text{(Alternating Update)}$$

Algorithm II: A Stochastic ADMM Algorithm

- **Define**: $L_k^\varrho(\mathbf{u},\boldsymbol{\omega},\boldsymbol{\lambda}) \triangleq f(\mathbf{u}^k) + \langle \mathbf{v}^k, \mathbf{u}-\mathbf{u}^k\rangle + g(\mathbf{u}) + h(\boldsymbol{\omega}) - \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{u}-\boldsymbol{\omega}\rangle$
$+r_k(2\eta_k)^{-1}\langle \mathbf{u}-\mathbf{u}^k, \mathbf{W}^k(\mathbf{u}-\mathbf{u}^k)\rangle + (\varrho/2)\|\mathbf{A}\mathbf{u}-\boldsymbol{\omega}\|^2$
- **Input**: Interpolation sequence $\{r_k\}_{k\in\mathbb{Z}^+}$, stepsizes $\{\eta_k\}_{k\in\mathbb{Z}^+}$, penalty parameter $\varrho > 0$
- **Initialize**: $\mathbf{u}^0 \in \mathbf{dom}\, g$, $\boldsymbol{\omega}^0 \in \mathbf{dom}\, h$, $\boldsymbol{\lambda}^0 \in \mathbb{R}^m$, $\overline{\mathbf{u}}^0 = \mathbf{u}^0$, $\overline{\boldsymbol{\omega}}^0 = \boldsymbol{\omega}^0$, $\overline{\boldsymbol{\lambda}}^0 = \boldsymbol{\lambda}^0$, $k = 0$
- **Repeat** (until some convergence criterion is met)

$$\tilde{\mathbf{u}}^k := r_k\mathbf{u}^k + (1-r_k)\overline{\mathbf{u}}^k \qquad \text{(Interpolation)}$$

Sample $\tilde{\boldsymbol{\xi}}^k \sim \nu$ and define $\tilde{\mathbf{v}}^k \triangleq \nabla_{\mathbf{u}} F(\mathbf{u}, \tilde{\boldsymbol{\xi}}^k)|_{\mathbf{u}=\tilde{\mathbf{u}}^k}$

$$\begin{aligned} \boldsymbol{\omega}^{k+1} &:= \arg\min_{\boldsymbol{\omega}\in\mathbf{dom}\, h} L_k^\varrho(\mathbf{u}^k, \boldsymbol{\omega}, \boldsymbol{\lambda}^k) \\ \mathbf{u}^{k+1} &:= \arg\min_{\mathbf{u}\in\mathbf{dom}\, g} L_k^\varrho(\mathbf{u}, \boldsymbol{\omega}^{k+1}, \boldsymbol{\lambda}^k) \end{aligned} \qquad \text{(Alternating Update)}$$

$$\boldsymbol{\lambda}^{k+1} := \boldsymbol{\lambda}^k - \varrho(\mathbf{A}\mathbf{u}^{k+1} - \boldsymbol{\omega}^{k+1}) \qquad \text{(Multiplier Update)}$$

Algorithm II: A Stochastic ADMM Algorithm (Cont'd)

$$(\overline{\boldsymbol{\omega}}^{k+1}, \overline{\mathbf{u}}^{k+1}, \overline{\boldsymbol{\lambda}}^{k+1}) := r_k(\boldsymbol{\omega}^{k+1}, \mathbf{u}^{k+1}, \boldsymbol{\lambda}^{k+1}) + (1 - r_k)(\overline{\boldsymbol{\omega}}^{k}, \overline{\mathbf{u}}^{k}, \overline{\boldsymbol{\lambda}}^{k}) \qquad \text{(Averging)}$$

Algorithm II: A Stochastic ADMM Algorithm (Cont'd)

$$(\overline{\boldsymbol{\omega}}^{k+1}, \overline{\mathbf{u}}^{k+1}, \overline{\boldsymbol{\lambda}}^{k+1}) := r_k(\boldsymbol{\omega}^{k+1}, \mathbf{u}^{k+1}, \boldsymbol{\lambda}^{k+1}) + (1 - r_k)(\overline{\boldsymbol{\omega}}^{k}, \overline{\mathbf{u}}^{k}, \overline{\boldsymbol{\lambda}}^{k}) \qquad \text{(Averging)}$$

$$k := k + 1$$

Algorithm II: A Stochastic ADMM Algorithm (Cont'd)

$$(\overline{\boldsymbol{\omega}}^{k+1}, \overline{\mathbf{u}}^{k+1}, \overline{\boldsymbol{\lambda}}^{k+1}) := r_k(\boldsymbol{\omega}^{k+1}, \mathbf{u}^{k+1}, \boldsymbol{\lambda}^{k+1}) + (1 - r_k)(\overline{\boldsymbol{\omega}}^{k}, \overline{\mathbf{u}}^{k}, \overline{\boldsymbol{\lambda}}^{k}) \qquad \text{(Averging)}$$

$$k := k + 1$$

- **Output**: $(\overline{\mathbf{u}}^{k}, \overline{\boldsymbol{\omega}}^{k}, \overline{\boldsymbol{\lambda}}^{k})$

Connection to Algorithm I

To see the connection, we choose any penalty parameter $\varrho > 0$,

Connection to Algorithm I

To see the connection, we choose any penalty parameter $\varrho > 0$,

$$r_k = \frac{1}{k+1},$$

Connection to Algorithm I

To see the connection, we choose any penalty parameter $\varrho > 0$,

$$r_k = \frac{1}{k+1}, \quad \eta_k^{-1} = L + 2\sigma(k+1)^{3/2} + c\varrho B^2(k+1),$$

Connection to Algorithm I

To see the connection, we choose any penalty parameter $\varrho > 0$,

$$r_k = \frac{1}{k+1}, \quad \eta_k^{-1} = L + 2\sigma(k+1)^{3/2} + c\varrho B^2(k+1),$$
$$a = \rho B^2/(3L^{1/3}\sigma^{2/3} + c\varrho B^2),$$

Connection to Algorithm I

To see the connection, we choose any penalty parameter $\varrho > 0$,

$$r_k = \frac{1}{k+1}, \quad \eta_k^{-1} = L + 2\sigma(k+1)^{3/2} + c\varrho B^2(k+1),$$
$$a = \rho B^2/(3L^{1/3}\sigma^{2/3} + c\varrho B^2),$$
$$\mathbf{W}^k = a\mathbf{I} - (\eta_k/r_k)\varrho\mathbf{A}^T\mathbf{A} \succeq 0. \qquad \text{(Pre-conditioning)}$$

Connection to Algorithm I

To see the connection, we choose any penalty parameter $\varrho > 0$,

$$r_k = \frac{1}{k+1}, \quad \eta_k^{-1} = L + 2\sigma(k+1)^{3/2} + c\varrho B^2(k+1),$$
$$a = \rho B^2/(3L^{1/3}\sigma^{2/3} + c\varrho B^2),$$
$$\mathbf{W}^k = a\mathbf{I} - (\eta_k/r_k)\varrho\mathbf{A}^T\mathbf{A} \succeq 0. \qquad \text{(Pre-conditioning)}$$

Using variable substitution and Moreau's identity,

Algorithm II is equivalent to Algorithm I with unit extrapolation parameter.

Numerical Experiments: Setup

	Abbrev.	Algorithms
Ours	OTPDHG	Algorithm I & Multi-Comp. Ext.
	OSADMM	Algorithm II & Multi-Comp. Ext.
Benchmarks	ESADMM	Algorithm 1, Lin et al. (2018)
	SADMM	Algorithm 2, Suzuki (2013)
	ASG-PA	Algorithm 1, Zhong-Kwok (2014)
	TPDHG	Algorithm 1, Z.-Cevher (2018)
	FOBOS	Section 2, Duchi-Singer (2009)

Numerical Experiments: Setup

	Abbrev.	Algorithms
Ours	OTPDHG	Algorithm I & Multi-Comp. Ext.
	OSADMM	Algorithm II & Multi-Comp. Ext.
Benchmarks	ESADMM	Algorithm 1, Lin et al. (2018)
	SADMM	Algorithm 2, Suzuki (2013)
	ASG-PA	Algorithm 1, Zhong-Kwok (2014)
	TPDHG	Algorithm 1, Z.-Cevher (2018)
	FOBOS	Section 2, Duchi-Singer (2009)

▷ Tasks: I) Graph-Guided Fused Logistic Regression
II) Sparse Overlapping Group Lasso

Numerical Experiments: Setup

	Abbrev.	Algorithms
Ours	OTPDHG	Algorithm I & Multi-Comp. Ext.
	OSADMM	Algorithm II & Multi-Comp. Ext.
Benchmarks	ESADMM	Algorithm 1, Lin et al. (2018)
	SADMM	Algorithm 2, Suzuki (2013)
	ASG-PA	Algorithm 1, Zhong-Kwok (2014)
	TPDHG	Algorithm 1, Z.-Cevher (2018)
	FOBOS	Section 2, Duchi-Singer (2009)

▷ Tasks: I) Graph-Guided Fused Logistic Regression
II) Sparse Overlapping Group Lasso

▷ Parameter setting:

- Algorithm I: $\rho = 1 \times 10^{-3}$ and $\rho' = 1 \times 10^{-5}$
- Algorithm II: $\varrho = \rho'/B$ and $c = 6 \times 10^{-2}$

Numerical Experiments: Setup

	Abbrev.	Algorithms
Ours	OTPDHG	Algorithm I & Multi-Comp. Ext.
	OSADMM	Algorithm II & Multi-Comp. Ext.
Benchmarks	ESADMM	Algorithm 1, Lin et al. (2018)
	SADMM	Algorithm 2, Suzuki (2013)
	ASG-PA	Algorithm 1, Zhong-Kwok (2014)
	TPDHG	Algorithm 1, Z.-Cevher (2018)
	FOBOS	Section 2, Duchi-Singer (2009)

▷ Tasks: I) Graph-Guided Fused Logistic Regression
II) Sparse Overlapping Group Lasso

▷ Parameter setting:

- Algorithm I: $\rho = 1\times 10^{-3}$ and $\rho' = 1\times 10^{-5}$
- Algorithm II: $\varrho = \rho'/B$ and $c = 6\times 10^{-2}$

▷ Comparison criterion: $P(\overline{\mathbf{x}}^k) - P^*$

Graph-Guided Fused Logistic Regression (GLR)

$$P_{\mathrm{GLR}}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} \log(1 + \exp(-b_i \mathbf{a}_i^T \mathbf{x})) + \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \|\mathbf{F}\mathbf{x}\|_1$$

Graph-Guided Fused Logistic Regression (GLR)

$$P_{\text{GLR}}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} \log(1 + \exp(-b_i \mathbf{a}_i^T \mathbf{x})) + \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \|\mathbf{F}\mathbf{x}\|_1$$

- $\lambda_1 = \lambda_2 = 1/\sqrt{n}$.

Graph-Guided Fused Logistic Regression (GLR)

$$P_{\mathrm{GLR}}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} \log(1+\exp(-b_i\mathbf{a}_i^T\mathbf{x})) + \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \|\mathbf{F}\mathbf{x}\|_1$$

- $\lambda_1 = \lambda_2 = 1/\sqrt{n}$.
- Relations among features $\{x_i\}_{i=1}^d$ represented by $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$.

Graph-Guided Fused Logistic Regression (GLR)

$$P_{\mathrm{GLR}}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} \log(1 + \exp(-b_i \mathbf{a}_i^T \mathbf{x})) + \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \|\mathbf{F}\mathbf{x}\|_1$$

- $\lambda_1 = \lambda_2 = 1/\sqrt{n}$.
- Relations among features $\{x_i\}_{i=1}^d$ represented by $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$.
- $\mathcal{G}$ encoded by the matrix $\mathbf{F}$ (of size $|\mathcal{E}| \times d$)
  - Let $\pi : \mathcal{E} \to [|\mathcal{E}|]$ be any bijection.
  - For any edge $(i,j) \in \mathcal{E}$ $(i < j)$, $\mathbf{F}_{\pi(i,j),i} = w(i,j)$, $\mathbf{F}_{\pi(i,j),j} = -w(i,j)$ and $\mathbf{F}_{\pi(i,j),s} = 0$ for all $s \in [d] \setminus \{i,j\}$.

Graph-Guided Fused Logistic Regression (GLR)

$$P_{\text{GLR}}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} \log(1 + \exp(-b_i \mathbf{a}_i^T \mathbf{x})) + \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \|\mathbf{F}\mathbf{x}\|_1$$

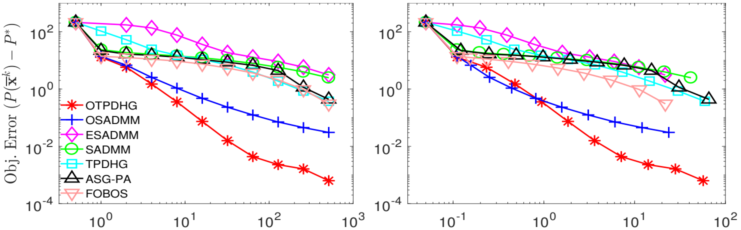
- $\lambda_1 = \lambda_2 = 1/\sqrt{n}$.
- Relations among features $\{x_i\}_{i=1}^{d}$ represented by $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$.
- $\mathcal{G}$ encoded by the matrix $\mathbf{F}$ (of size $|\mathcal{E}| \times d$)
  - Let $\pi : \mathcal{E} \to [|\mathcal{E}|]$ be any bijection.
  - For any edge $(i,j) \in \mathcal{E}$ $(i < j)$, $\mathbf{F}_{\pi(i,j),i} = w(i,j)$, $\mathbf{F}_{\pi(i,j),j} = -w(i,j)$ and $\mathbf{F}_{\pi(i,j),s} = 0$ for all $s \in [d] \setminus \{i,j\}$.
- Stochastic gradient
  - Uniformly randomly sample $\mathcal{B}_k$ from $[n]$.
  - $\mathbf{v}^k = (1/|\mathcal{B}_k|)\sum_{i \in \mathcal{B}_k} \nabla \ell_i^{\text{LR}}(\mathbf{x}^k)$.

Graph-Guided Fused Logistic Regression (GLR)

$$P_{\mathrm{GLR}}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} \log(1+\exp(-b_i \mathbf{a}_i^T \mathbf{x})) + \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \|\mathbf{F}\mathbf{x}\|_1$$

- $\lambda_1 = \lambda_2 = 1/\sqrt{n}$.
- Relations among features $\{x_i\}_{i=1}^{d}$ represented by $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$.
- $\mathcal{G}$ encoded by the matrix $\mathbf{F}$ (of size $|\mathcal{E}| \times d$)
  - Let $\pi : \mathcal{E} \to [|\mathcal{E}|]$ be any bijection.
  - For any edge $(i,j) \in \mathcal{E}$ $(i < j)$, $\mathbf{F}_{\pi(i,j),i} = w(i,j)$, $\mathbf{F}_{\pi(i,j),j} = -w(i,j)$ and $\mathbf{F}_{\pi(i,j),s} = 0$ for all $s \in [d] \setminus \{i,j\}$.
- Stochastic gradient
  - Uniformly randomly sample $\mathcal{B}_k$ from $[n]$.
  - $\mathbf{v}^k = (1/|\mathcal{B}_k|)\sum_{i\in\mathcal{B}_k} \nabla \ell_i^{\mathrm{LR}}(\mathbf{x}^k)$.
- Plots averaged from ten independent runs.

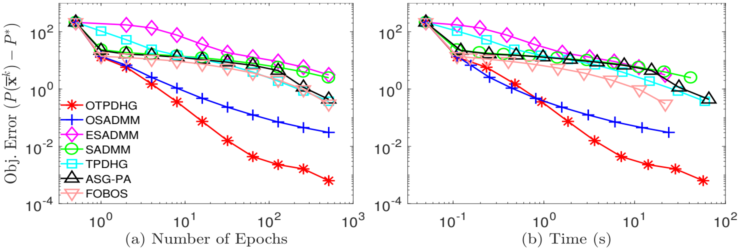
a9a



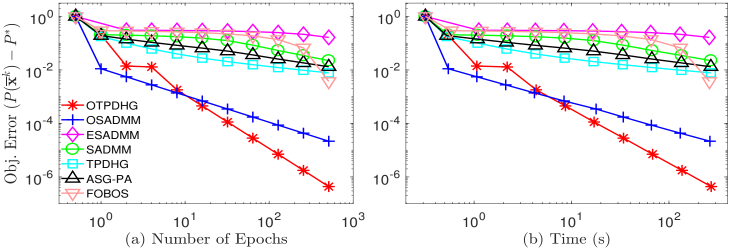
(a) Number of Epochs

(b) Time (s)

a9a


Obj. Error $(P(\overline{\mathbf{x}}^k) - P^*)$
OTPDHG
OSADMM
ESADMM
SADMM
TPDHG
ASG-PA
FOBOS
(a) Number of Epochs
(b) Time (s)

covtype


Obj. Error $(P(\overline{\mathbf{x}}^k) - P^*)$
OTPDHG
OSADMM
ESADMM
SADMM
TPDHG
ASG-PA
FOBOS
(a) Number of Epochs
(b) Time (s)

Sparse Overlapping Group Lasso (OGL)

$$P_{\mathrm{GLR}}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n}(\mathbf{a}_i^T\mathbf{x} - b_i)^2/2 + \lambda_0\|\mathbf{x}\|_1 + \sum_{i=1}^{p}\lambda_i\|\mathbf{x}_{\mathcal{G}_i}\|$$

Sparse Overlapping Group Lasso (OGL)

$$P_{\text{GLR}}(\mathbf{x}) = n^{-1}\textstyle\sum_{i=1}^{n}(\mathbf{a}_i^T\mathbf{x} - b_i)^2/2 + \lambda_0\|\mathbf{x}\|_1 + \textstyle\sum_{i=1}^{p}\lambda_i\|\mathbf{x}_{\mathcal{G}_i}\|$$

- $\lambda_0 = \lambda_1 = \ldots = \lambda_p = 1.$

Sparse Overlapping Group Lasso (OGL)

$$P_{\mathrm{GLR}}(\mathbf{x}) = n^{-1}\textstyle\sum_{i=1}^{n}(\mathbf{a}_i^T\mathbf{x} - b_i)^2/2 + \lambda_0\|\mathbf{x}\|_1 + \textstyle\sum_{i=1}^{p}\lambda_i\|\mathbf{x}_{\mathcal{G}_i}\|$$

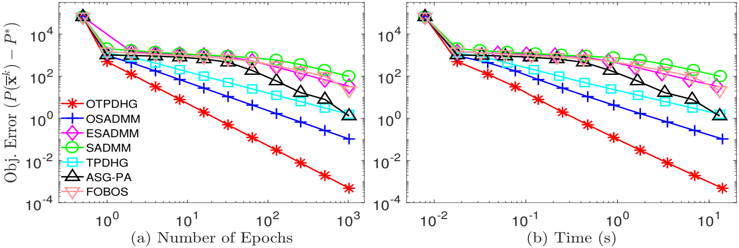
- $\lambda_0 = \lambda_1 = \ldots = \lambda_p = 1$.
- $\mathbf{x}_{\mathcal{G}_i}$ denotes the subvector of $\mathbf{x}$ indexed by $\mathcal{G}_i \subseteq [d]$.

Sparse Overlapping Group Lasso (OGL)

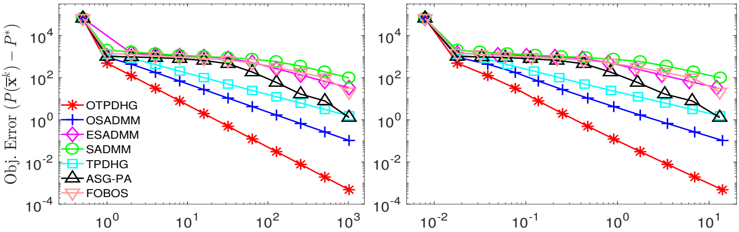
$$P_{\text{GLR}}(\mathbf{x}) = n^{-1}\textstyle\sum_{i=1}^{n}(\mathbf{a}_i^T\mathbf{x} - b_i)^2/2 + \lambda_0\|\mathbf{x}\|_1 + \textstyle\sum_{i=1}^{p}\lambda_i\|\mathbf{x}_{\mathcal{G}_i}\|$$

- $\lambda_0 = \lambda_1 = \ldots = \lambda_p = 1$.
- $\mathbf{x}_{\mathcal{G}_i}$ denotes the subvector of $\mathbf{x}$ indexed by $\mathcal{G}_i \subseteq [d]$.
- Solved by multi-composite extensions of Algorithms I and II.

cpu

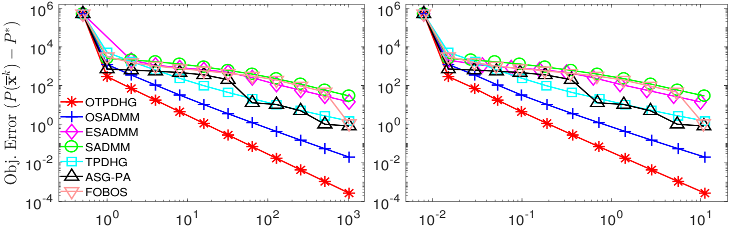

Obj. Error $(P(\bar{\mathbf{x}}^k) - P^*)$
OTPDHG
OSADMM
ESADMM
SADMM
TPDHG
ASG-PA
FOBOS
(a) Number of Epochs
(b) Time (s)

cpu



(a) Number of Epochs (b) Time (s)

abalone



(a) Number of Epochs (b) Time (s)

Future work

- Consider strongly convex f.
- Extend to non-Euclidean geometry.
- Consider randomized matrix-vector product $\mathbf{Ax}$ and $\mathbf{A}^T\mathbf{y}$.
- Extend to non-bilinear structure.

Thank you!

References

[AS14] Samaneh Azadi and Suvrit Sra. “Towards an Optimal Stochastic Alternating Direction Method of Multipliers”. In: *Proc. ICML*. Beijing, China, 2014, pp. 620–628.

[CLO14] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. “Optimal Primal-Dual Methods for a Class of Saddle Point Problems”. In: *SIAM J. Optim.* 24.4 (2014), pp. 1779–1814.

[DS09] John Duchi and Yoram Singer. “Efficient Online and Batch Learning Using Forward Backward Splitting”. In: *J. Mach. Learn. Res.* 10 (2009), pp. 2899–2934.

[Lan12] Guanghui Lan. “An Optimal Method for Stochastic Composite Optimization”. In: *Math. Program.* 133.1-2 (2012), pp. 365–397.

[Lin18] Tianyi Lin et al. “Stochastic Primal-Dual Proximal ExtraGradient descent for compositely regularized optimization”. In: *Neurocomput.* 273 (2018), pp. 516 –525.

[Ouy13] Hua Ouyang et al. “Stochastic Alternating Direction Method of Multipliers”. In: *Proc. ICML*. Atlanta, USA, 2013, pp. 80–88.

[Suz13] Taiji Suzuki. “Dual Averaging and Proximal Gradient Descent for Online Alternating Direction Multiplier Method”. In: *Proc. ICML*. Atlanta, USA, 2013, pp. 392–400.

[ZC18] Renbo Zhao and Volkan Cevher. “Stochastic Three-Composite Convex Minimization with a Linear Operator”. In: *Proc. AISTATS*. Lanzarote, Spain, 2018.

[ZK14] Wenliang Zhong and James Kwok. “Accelerated Stochastic Gradient Method for Composite Regularization”. In: *Proc. AISTATS*. Reykjavik, Iceland, 2014, pp. 1086–1094.