Faster Accelerated First-order Methods for Convex Optimization with Strongly Convex Function Constraints

Zhenwei Lin
School of Information Management and Engineering
Shanghai University of Finance and Economics

Joint work with

Qi Deng
ANTAI, SJTU

NeurIPS, 2024

November 14, 2024

Outline

Problem Setting

Convex optimization with strongly convex function constraints:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & G(\mathbf{x}) \leq \mathbf{0} \end{aligned} \tag{1}$$

- f is a proper and convex continuous (possibly non-smooth) function of which the prox-mapping is easy to compute
- $G = [g_1, \ldots, g_m]^\top : \mathbb{R}^n \to \mathbb{R}^m$; g_i is Lipschitz smooth
- g_i is strongly-convex function with modulus μ_i.

Related Work and Existing Problem

Complexities of First-order Methods:

- $\mathcal{O}(1/\varepsilon)$ if $f(\mathbf{x})$ is generally-convex, regardless the strong convexity of constraint functions
- $\mathcal{O}(1/\sqrt{\varepsilon})$ if the problem is strongly-convex-concave (Lin et al. (2020))

Existing Problem:

- Best $\mathcal{O}(1/\sqrt{\varepsilon})$ can be achieved if we assume $\langle \mathbf{y}, G(\mathbf{x}) \rangle$ has a uniform convexity modulus for any feasible $\mathbf{y}$ (Juditsky et al. (2011); Hamedani and Aybat (2021)).

Our contribution:

- A procedure to estimate the lower bound of strong convexity of Lagrangian function.

Core Assumption and Examples

Assumption 1 (Nontrivial Solution (Informal))

1. $\forall \mathbf{x}_0^* \in \operatorname{argmin}_{\mathbf{x}} f(\mathbf{x})$, *we have* $g_i(\mathbf{x}_0^*) > 0$
2. $\mathit{dist}(\partial f(\mathbf{x}^*), \mathbf{0}) \geq r > 0$

Sparse Learning

$$
\begin{aligned}
\min \quad & f(\mathbf{x}) := \sum_{i=1}^{B} p_i \|\mathbf{x}_{(i)}\| \\
\text{s.t.} \quad & \mathbf{x} = \mathbf{x}_{(1)} \times \ldots \times \mathbf{x}_{(B)}, \Longrightarrow r = \min_{1 \leq i \leq B} \{p_i\} \\
& \mathbf{x}_{(i)} \in \mathbb{R}^{n_i}, 1 \leq i \leq B \\
& G(\mathbf{x}) \leq \mathbf{0}.
\end{aligned}
$$

Linear Objective function

$$f(\mathbf{x}) = \mathbf{c}^T \mathbf{x} \Longrightarrow r = \|\mathbf{c}\|$$

Outline

How to estimate strong convexity

Key intuition: Apply the subdifferential separation property to bound the dual variables.

- Using the KKT condition

$$\mathbf{0} \in \partial f(\mathbf{x}^*) + \nabla G(\mathbf{x}^*)\mathbf{y}^*.$$

- It follows from $\operatorname{dist}(\partial f(\mathbf{x}^*), \mathbf{0}) \geq r > 0$ that

$$r \leq \|\nabla G(\mathbf{x}^*)\mathbf{y}^*\| \leq \|\nabla G(\mathbf{x}^*)\| \cdot \|\mathbf{y}^*\| \leq \|\mathbf{y}^*\|_1 \|\nabla G(\mathbf{x}^*)\|,$$

- $\|\nabla G(\mathbf{x}^*)\| - \|\nabla G(\mathbf{x})\| \leq \|\nabla G(\mathbf{x}^*) - \nabla G(\mathbf{x})\| \leq L_X \|\mathbf{x} - \mathbf{x}^*\|$
  1. $\frac{1}{2}\|\hat{\mathbf{x}} - \mathbf{x}^*\|^2 \leq \beta$

$$\|\mathbf{y}^*\|_1 \geq h_1(\hat{\mathbf{x}}, \beta) := r\left[\|\nabla G(\hat{\mathbf{x}})\| + L_X\sqrt{2\beta}\right]^{-1}. \tag{2}$$

  2. $\frac{\|\mathbf{y}\|_1 \underline{\mu}}{2}\|\hat{\mathbf{x}} - \mathbf{x}^*\|^2 \leq \frac{(\mathbf{y}^*)^T \boldsymbol{\mu}}{2}\|\hat{\mathbf{x}} - \mathbf{x}^*\|^2 \leq \beta$:

$$\|\mathbf{y}^*\|_1 \geq h_2(\hat{\mathbf{x}}, \beta) := \left[\frac{L_X}{r}\sqrt{\frac{\beta}{2\underline{\mu}}} + \sqrt{\frac{L_X^2\beta}{2\underline{\mu}r^2} + \frac{\|\nabla G(\hat{\mathbf{x}})\|}{r}}\right]^{-2}. \tag{3}$$

Accelerated Primal-Dual Algorithm (APD) [1]

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}) + \langle G(\mathbf{x}), \mathbf{y}\rangle \tag{4}$$

If the strongly convex modulus ρ is known, then we

1. $\mathbf{z}_k \leftarrow (1+\sigma_{k-1}/\sigma_k)G(\mathbf{x}_k) - (\sigma_{k-1}/\sigma_k)G(\mathbf{x}_{k-1})$
2. $\mathbf{y}_{k+1} \leftarrow \mathrm{proj}_{\mathcal{Y}}\{\mathbf{y}_k + \sigma_k \mathbf{z}_k\}$
3. $\mathbf{x}_{k+1} \leftarrow \mathrm{argmin}_{\mathbf{x}\in\mathcal{X}}\{f(\mathbf{x}) + \langle \nabla G(\mathbf{x}_k)\mathbf{y}_{k+1}, \mathbf{x}\rangle + \frac{1}{2\tau_k}\|\mathbf{x}-\mathbf{x}_{k-1}\|^2\}$
4. $\tau_{k+1} \leftarrow \tau_k/\sqrt{1+\rho\tau_k},\ \sigma_{k+1} \leftarrow \tau_0\sigma_0/\tau_{k+1}$

[1] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. SIAM Journal on Optimization, 2021

Algorithm 1 APD with Progressive Strong Convexity Estimation (APDPro)

Require: $\tau_0 > 0, \sigma_0 > 0, \mathbf{x}_0 \in \mathcal{X}, \mathbf{y}_0 \in \mathbf{y}, \rho_0 > 0, N > 0$

Initialize: $(\mathbf{x}_{-1}, \mathbf{y}_{-1}) \leftarrow (\mathbf{x}_0, \mathbf{y}_0)\,, \bar{\mathbf{x}}_0 \leftarrow \mathbf{x}_0,\ \sigma_{-1} \leftarrow \sigma_0, \gamma_0 \leftarrow \sigma_0/\tau_0, T_0 = 0$

Set $\Delta_{XY} = \frac{1}{2\tau_0} D_X^2 + \frac{1}{2\sigma_0} D_Y^2$

for $k = 0, 1, \ldots, N$ **do**

$\mathcal{Y}_k \leftarrow \left\{\mathbf{y} \in \mathbb{R}_+^m \mid \|\mathbf{y}\|_1 \cdot \underline{\mu} \geq \rho_k\right\} \bigcap \mathcal{Y}$

$\mathbf{z}_k \leftarrow (1 + \sigma_{k-1}/\sigma_k) G(\mathbf{x}_k) - (\sigma_{k-1}/\sigma_k) G(\mathbf{x}_{k-1})$

$\mathbf{y}_{k+1} \leftarrow \operatorname{argmin}_{\mathbf{y} \in \mathcal{Y}_k} -\langle \mathbf{y}, \mathbf{z}_k \rangle + \frac{1}{2\sigma_k} \|\mathbf{y} - \mathbf{y}_k\|^2$

$\mathbf{x}_{k+1} \leftarrow \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \langle \nabla G(\mathbf{x}_k) \mathbf{y}_{k+1}, \mathbf{x} \rangle + \frac{1}{2\tau_k} \|\mathbf{x} - \mathbf{x}_{k-1}\|^2$

$t_k \leftarrow \sigma_k/\sigma_0,\ \bar{\mathbf{x}}_{k+1} \leftarrow (T_k \bar{\mathbf{x}}_k + t_k \mathbf{x}_{k+1})/(T_k + t_k),\ T_{k+1} \leftarrow T_k + t_k$

$\rho_{k+1} \leftarrow \textsc{Improve}(\mathbf{x}_k,\ \bar{\mathbf{x}}_k,\ \frac{\sigma_0 \tau_{k-1} \Delta_{XY}}{\sigma_{k-1}},\ \frac{\Delta_{XY}}{T_k},\ \rho_k)$

$\tau_{k+1} \leftarrow \tau_k / \sqrt{1 + \rho_{k+1} \tau_{k+1}},\ \sigma_{k+1} \leftarrow \tau_0 \sigma_0 / \tau_{k+1}$

end for

Output: $\mathbf{x}_{N+1}, \mathbf{y}_{N+1}, (\rho_{N+1}$ for restart)

procedure $\textsc{Improve}(\mathbf{x},\ \bar{\mathbf{x}},\ \beta,\ \bar{\beta},\ \rho_{\text{old}})$

$\hat{\rho} = \underline{\mu} \cdot \max\left\{ r\left[\|\nabla G(\mathbf{x})\| + L_X \sqrt{2\beta}\right]^{-1}, \left[\frac{L_X}{r}\sqrt{\frac{\bar{\beta}}{2\underline{\mu}}} + \sqrt{\frac{L_X^2 \bar{\beta}}{2\underline{\mu} r^2} + \frac{\|\nabla G(\bar{\mathbf{x}})\|}{r}}\right]^{-2} \right\}$

Set $\rho_{\text{new}} = \max\{\rho_{\text{old}}, \hat{\rho}\}$

return ρ_{new}

end procedure

Convergence Analysis for APDPro

Suppose that

$$\tau_0^{-1} \geq L_{XY} + L_G^2 \sigma_0, \tag{5}$$

then we have

$$f(\bar{\mathbf{x}}_K) - f(\mathbf{x}^*) \leq \frac{6}{6 + \tau_0 \tilde{\rho}_K (K+1) K} \Big(\frac{1}{2\tau_0} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{D_Y^2}{2\sigma_0} \Big)$$

$$\|[G(\bar{\mathbf{x}}_K)]_+\| \leq \frac{6}{c^* \big(6 + \tau_0 \tilde{\rho}_K (K+1) K\big)} \Big(\frac{1}{2\tau_0} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{D_Y^2}{2\sigma_0} \Big)$$

$$\frac{1}{2} \|\mathbf{x}_K - \mathbf{x}^*\|^2 \leq \frac{3\sigma_0}{\hat{\rho}_K^2 \tau_0^2 K^2 + 9\sigma_0/\tau_0} \Big(\frac{1}{2\tau_0} \|\mathbf{x}^* - \mathbf{x}_0\|^2 + \frac{1}{2\sigma_0} \|\mathbf{y}^* - \mathbf{y}_0\|^2 \Big),$$

where

$$\hat{\rho}_1 = 3\sqrt{\rho_1/\tau_0}, \quad \hat{\rho}_{k+1} = \frac{\sqrt{\hat{\rho}_k^2 k^2 + (3\rho_{k+1}\hat{\rho}_k)k}}{k+1}, \quad \tilde{\rho}_k = 2 \sum_{s=0}^{k} \frac{\hat{\rho}_s s}{k(k+1)} \tag{6}$$

Reference I

Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31 (2):1299–1329, 2021.

Anatoli Juditsky, Arkadi Nemirovski, et al. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning*, 30(9):149–183, 2011.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020.

Q & A

Thank you