Lecture 1.

Inexact Nesterov's accelerated proximal gradient method

Yangyang Xu

xuy21@rpi.edu

Motivated by applications where only approximate gradients are available, we study an inexact variant of Nesterov's accelerated proximal gradient (APG) method given in [1]. Though inexact APG has been explored in literature, e.g., [2, 3], Nesterov's APG achieves accelerated convergence for both convex and strongly convex problems in a unified framework.

1.1 Problem Setting

We consider the following structured problem

$$x^* = \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \phi(x) := f(x) + \Psi(x), \tag{1.1}$$

where f is convex and L_f-smooth on $\mathbb{R}^n$, and $\Psi(x)$ is closed and μ-strongly convex with $\mu > 0$. We assume that for any $\eta > 0$, the proximal mapping of $\eta\Psi$ can be computed exactly, i.e.,

$$\mathbf{prox}_{\eta\Psi}(u) := \operatorname*{arg\,min}_{x} \frac{1}{2}\|x-u\|^2 + \eta\Psi(x) \tag{1.2}$$

can be computed for each $u \in \mathbb{R}^n$. However, an exact gradient of f is not available; instead its approximate gradient at any x, denoted as $\tilde{\nabla} f(x)$, can be accessed. For each $x \in \mathbb{R}^n$, we denote

$$e(x) = \tilde{\nabla} f(x) - \nabla f(x) \tag{1.3}$$

as the error of the approximate gradient.

1.2 Inexact APG with line search

With the inexact gradient oracle, we give the inexact variant of Nesterov's APG in Algorithm 1 for solving (1.1). Notice that we assume to know the strong convexity constant $\mu > 0$. Nevertheless, by performing line search, we do not need to know the smoothness constant L_f. Except using inexact gradients, other settings, including the line search strategy and choice of coefficients a_k and A_k, are the same as those in [1]. A key difference lies at the step to obtain a near-stationary solution.

1.3 Convergence Analysis

We follow the analysis in [1] but extend a few of its key results to accommodate the usage of inexact gradients. First, we show the while-loop for line search must stop within a finite number of steps. Based on the next lemma, we assume $L_{\min} \le L_f$ and $L_0 \le L_f$ without loss of generality.

Lemma 1. For $\xi = \tilde{\nabla} f(z) - \tilde{\nabla} f(y) - L(z-y)$, when $L \ge L_f$, it must hold that

$$\langle \xi, y - z\rangle \ge \frac{1}{L}\|\xi\|^2 - 3\|e(y)-e(z)\|\cdot\|y-z\| - \frac{1}{L}\|e(y)-e(z)\|^2.$$

Algorithm 1: Inexact accelerated proximal gradient method with line search for (1.1)

Input: $\varepsilon > 0$, $x_0 \in \mathrm{dom}(\Psi)$, $\gamma_u > 1$, $\gamma_d > 1$, $L_{\min} > 0$, and a positive sequence $\{\tau_k\}_{k\ge 0}$.
Set: $A_0 = 0$, $\psi_0(x) = \frac{1}{2}\|x - x_0\|^2$, $L_0 \ge L_{\min}$, and $v_0 = x_0$.
for $k = 0, 1, \ldots$ **do**
  Let $L = L_k$. Set `FLAG` = false.
  while *FLAG* == *false* **do**
    Find $a > 0$ such that $\frac{a^2}{A_k + a} = \frac{2(1+\mu A_k)}{L}$; let $y = \frac{A_k x_k + a v_k}{A_k + a}$.
    Set $z = \mathbf{prox}_{\Psi/L}\left(y - \frac{1}{L}\tilde\nabla f(y)\right)$ and let $\xi = \tilde\nabla f(z) - \tilde\nabla f(y) - L(z - y)$.
    /* Suppose $\|e(y)\| \le \tau_k$ and $\|e(z)\| \le \tau_k$ */
    if $\langle \xi, y - z\rangle \ge \frac{1}{L}\|\xi\|^2 - 6\tau_k\|y - z\| - \frac{4\tau_k^2}{L}$ **then**
      `FLAG` = true.
    else
      $L = L \cdot \gamma_u$.
  Let $y_k = y$, $x_{k+1} = z$, $M_k = L$, $L_{k+1} = \max\{L_{\min}, L/\gamma_d\}$, $a_{k+1} = a$, and $A_{k+1} = A_k + a$.
  Set $v_{k+1} = \arg\min_x \psi_{k+1}(x) := \psi_k(x) + a_{k+1}\left(f(x_{k+1}) + \langle \tilde\nabla f(x_{k+1}), x - x_{k+1}\rangle + \Psi(x)\right)$.
  if $\|\tilde\nabla f(z) - \tilde\nabla f(y) - L(z - y)\| \le \frac{\varepsilon}{2}$ **then**
    Return z and Stop.

Consequently, the while-loop in Algorithm 1 must exit after at most $\left\lceil \log_{\gamma_u} \frac{L_f}{L_{\min}} \right\rceil + 1$ steps.

Proof. By $\xi = \tilde\nabla f(z) - \tilde\nabla f(y) - L(z - y)$, it holds

$$
\begin{aligned}
&\langle \xi,\, y - z\rangle \\
&= L\|y - z\|^2 - \left\langle \tilde\nabla f(y) - \tilde\nabla f(z),\, y - z\right\rangle \\
&= \frac{1}{L}\left(\|\xi\|^2 + 2L\langle \tilde\nabla f(y) - \tilde\nabla f(z),\, y - z\rangle - \|\tilde\nabla f(y) - \tilde\nabla f(z)\|^2\right) - \left\langle \tilde\nabla f(y) - \tilde\nabla f(z),\, y - z\right\rangle \\
&= \frac{1}{L}\|\xi\|^2 + \left\langle \tilde\nabla f(y) - \tilde\nabla f(z),\, y - z\right\rangle - \frac{1}{L}\|\tilde\nabla f(y) - \tilde\nabla f(z)\|^2.
\end{aligned}
$$

Hence, using the definition of $e(\cdot)$ in (1.3), we have

$$
\begin{aligned}
\langle \xi,\, y - z\rangle = {}& \frac{1}{L}\|\xi\|^2 + \langle \nabla f(y) - \nabla f(z),\, y - z\rangle - \frac{1}{L}\|\nabla f(y) - \nabla f(z)\|^2 \\
& + \langle e(y) - e(z),\, y - z\rangle - \frac{1}{L}\|e(y) - e(z)\|^2 - \frac{2}{L}\langle \nabla f(y) - \nabla f(z),\, e(y) - e(z)\rangle \\
\ge {}& \frac{1}{L}\|\xi\|^2 + \langle \nabla f(y) - \nabla f(z),\, y - z\rangle - \frac{1}{L}\|\nabla f(y) - \nabla f(z)\|^2 \\
& - \left(1 + \tfrac{2L_f}{L}\right)\|e(y) - e(z)\| \cdot \|y - z\| - \tfrac{1}{L}\|e(y) - e(z)\|^2,
\end{aligned} \tag{1.4}
$$

where the inequality follows from $\langle \nabla f(y) - \nabla f(z),\, e(y) - e(z)\rangle \le L_f\|e(y) - e(z)\| \cdot \|y - z\|$ by the L_f-smoothness of f and the Cauchy-Schwarz inequality.

Now when $L \ge L_f$, we have $1 + \frac{2L_f}{L} \le 3$ and from the convexity and L_f-smoothness of f that

$$
\langle \nabla f(y) - \nabla f(z),\, y - z\rangle - \frac{1}{L}\|\nabla f(y) - \nabla f(z)\|^2 \ge \langle \nabla f(y) - \nabla f(z),\, y - z\rangle - \frac{1}{L_f}\|\nabla f(y) - \nabla f(z)\|^2 \ge 0.
$$

Hence, when $L \ge L_f$, the inequality in (1.4) implies the desired result. $\square$

The next lemma extends one key relation in [1] to the inexact case.

Lemma 2. Let $\delta_0 = 0$ and define $\{\delta_k\}_{k\ge 1}$ by

$$\delta_{k+1} = \delta_k + A_{k+1}\,\varepsilon_k + A_k\tau_k\|x_k - x_{k+1}\|, \ \forall\, k \ge 0, \tag{1.5}$$

with

$$\varepsilon_k = 6\tau_k\|y_k - x_{k+1}\| + \frac{4\tau_k^2}{M_k}, \ \forall\, k \ge 0. \tag{1.6}$$

Also, let $\psi_k^* = \min_x \psi_k(x)$ for each $k \ge 0$. Then it holds that

$$A_k\phi(x_k) \le \psi_k^* + \delta_k, \ \forall\, k \ge 0. \tag{$\mathcal{R}_k^1$}$$

Proof. Since $A_0 = 0$, $\delta_0 = 0$, and $\psi_0^* = 0$, the inequality in $(\mathcal{R}_k^1)$ clearly holds for $k = 0$. Now suppose it holds at k. We show it also holds for $k+1$ as follows.

By the $(1+\mu A_k)$–strong convexity of ψ_k, we have

$$\psi_k(x) \ge \psi_k^* + \frac{1+\mu A_k}{2}\|x - v_k\|^2 \ge A_k\phi(x_k) - \delta_k + \frac{1+\mu A_k}{2}\|x - v_k\|^2. \tag{1.7}$$

Hence, for any $g_{k+1} \in \partial\Psi(x_{k+1})$, it holds

$$\begin{aligned}
\psi_{k+1}^* &= \min_x \Big\{\psi_k(x) + a_{k+1}\big[f(x_{k+1}) + \langle\tilde\nabla f(x_{k+1}), x - x_{k+1}\rangle + \Psi(x)\big]\Big\}\\
&\overset{(1.7)}{\ge} \min_x \Big\{A_k\phi(x_k) - \delta_k + \frac{1+\mu A_k}{2}\|x - v_k\|^2\\
&\qquad\qquad + a_{k+1}\big[\phi(x_{k+1}) + \langle\tilde\nabla f(x_{k+1}) + g_{k+1}, x - x_{k+1}\rangle\big]\Big\}\\
&\ge \min_x \Big\{A_k\big[\phi(x_{k+1}) + \langle\nabla f(x_{k+1}) + g_{k+1}, x_k - x_{k+1}\rangle\big] - \delta_k + \frac{1+\mu A_k}{2}\|x - v_k\|^2\\
&\qquad\qquad + a_{k+1}\big[\phi(x_{k+1}) + \langle\tilde\nabla f(x_{k+1}) + g_{k+1}, x - x_{k+1}\rangle\big]\Big\}\\
&= \min_x \Big\{A_{k+1}\phi(x_{k+1}) + \langle\nabla f(x_{k+1}) + g_{k+1}, A_k(x_k - x_{k+1})\rangle - \delta_k + \frac{1+\mu A_k}{2}\|x - v_k\|^2\\
&\qquad\qquad + a_{k+1}\langle\tilde\nabla f(x_{k+1}) + g_{k+1}, x - x_{k+1}\rangle\Big\}\\
&= \min_x \Big\{A_{k+1}\phi(x_{k+1}) - \langle e(x_{k+1}), A_k(x_k - x_{k+1})\rangle - \delta_k + \frac{1+\mu A_k}{2}\|x - v_k\|^2\\
&\qquad + \langle\tilde\nabla f(x_{k+1}) + g_{k+1}, A_k(x_k - x_{k+1})\rangle + a_{k+1}\langle\tilde\nabla f(x_{k+1}) + g_{k+1}, x - x_{k+1}\rangle\Big\}\\
&= \min_x \Big\{A_{k+1}\phi(x_{k+1}) - \langle e(x_{k+1}), A_k(x_k - x_{k+1})\rangle - \delta_k + \frac{1+\mu A_k}{2}\|x - v_k\|^2\\
&\qquad + A_{k+1}\langle\tilde\nabla f(x_{k+1}) + g_{k+1}, y_k - x_{k+1}\rangle + a_{k+1}\langle\tilde\nabla f(x_{k+1}) + g_{k+1}, x - v_k\rangle\Big\},
\end{aligned} \tag{1.8}$$

where the second inequality follows from the convexity of ϕ, and the last equality holds by the relation $A_k x_k = A_{k+1}y_k - a_{k+1}v_k$.

Notice that the minimum in (1.8) is achieved at $v_k - \frac{a_{k+1}}{1+\mu A_k}(\tilde\nabla f(x_{k+1}) + g_{k+1})$. Thus,

$$\begin{aligned}
\psi_{k+1}^* \ge{}& A_{k+1}\phi(x_{k+1}) - \langle e(x_{k+1}), A_k(x_k - x_{k+1})\rangle + A_{k+1}\langle\tilde\nabla f(x_{k+1}) + g_{k+1}, y_k - x_{k+1}\rangle - \delta_k\\
&- \frac{a_{k+1}^2}{2(1+\mu A_k)}\|\tilde\nabla f(x_{k+1}) + g_{k+1}\|^2.
\end{aligned} \tag{1.9}$$

Now choose $g_{k+1} = -\tilde\nabla f(y_k) - M_k(x_{k+1} - y_k) \in \partial\Psi(x_{k+1})$ in the above inequality and use the exit condition for the first while-loop in Algorithm 1, i.e., $\langle\tilde\nabla f(x_{k+1}) + g_{k+1}, y_k - x_{k+1}\rangle \ge \frac{1}{M_k}\|\tilde\nabla f(x_{k+1}) + g_{k+1}\|^2 - \varepsilon_k$, to obtain from (1.9) and the Cauchy-Schwarz inequality that

$$\psi_{k+1}^* \ge A_{k+1}\phi(x_{k+1}) - A_k\|e(x_{k+1})\|\|x_k - x_{k+1}\| - A_{k+1}\varepsilon_k - \delta_k$$

$$+\left(\frac{A_{k+1}}{M_k}-\frac{a_{k+1}^2}{2(1+\mu A_k)}\right)\|\tilde{\nabla} f(x_{k+1})+g_{k+1}\|^2$$
$$\geq A_{k+1}\phi(x_{k+1})-\delta_{k+1},$$

where the last inequality follows from the definition of δ_{k+1}, $\|e(x_{k+1})\|\leq\tau_k$, and the choice of a_{k+1}. Hence, we complete the induction and finish the proof. □

The lemma below extends another key relation in [1] to the inexact case.

Lemma 3. Let $B_0(x)\equiv 0, \forall\, x\in\mathbb{R}^n$ and define the sequence of functions $\{B_k(\cdot)\}_{k\geq 1}$ by

$$B_{k+1}(x)=B_k(x)+a_{k+1}\langle e(x_{k+1}), x-x_{k+1}\rangle,\ \forall\, k\geq 0. \tag{1.10}$$

Then it holds

$$\psi_k(x)\leq A_k\phi(x)+\frac{1}{2}\|x-x_0\|^2+B_k(x), \forall\, x\in \mathrm{dom}(\Psi),\ \forall\, k\geq 0. \tag{$\mathcal{R}_k^2$}$$

Proof. By the definition of $B_0(\cdot)$, the choice of $\psi_0(\cdot)$, and $A_0=0$, the inequality in ($\mathcal{R}_k^2$) clearly holds for $k=0$. Below we prove it by induction. Suppose it holds for some k. Then we have

$$\begin{aligned}
\psi_{k+1}(x) &= \psi_k(x)+a_{k+1}\Big[f(x_{k+1})+\langle\tilde{\nabla} f(x_{k+1}),\, x-x_{k+1}\rangle+\Psi(x)\Big]\\
&\leq A_k\phi(x)+\frac{1}{2}\|x-x_0\|^2+B_k(x)+a_{k+1}[f(x_{k+1})+\langle\nabla f(x_{k+1}),\, x-x_{k+1}\rangle+\Psi(x)]\\
&\quad + a_{k+1}\langle e(x_{k+1}),\, x-x_{k+1}\rangle\\
&\leq A_k\phi(x)+\frac{1}{2}\|x-x_0\|^2+B_k(x)+a_{k+1}[f(x)+\Psi(x)]+a_{k+1}\langle e(x_{k+1}),\, x-x_{k+1}\rangle\\
&= A_{k+1}\phi(x)+\frac{1}{2}\|x-x_0\|^2+B_{k+1}(x),
\end{aligned}$$

where the first inequality follows from the induction assumption, and the second one is by the convexity of f. Hence, we finish the induction and complete the proof. □

With ($\mathcal{R}_k^1$) and ($\mathcal{R}_k^2$), we are ready to establish the convergence rate result.

Theorem 1 (Convergence Rate with Error Terms)**.** For each $k\geq 0$, it holds

$$\frac{1+\mu A_k}{2}\|x^*-v_k\|^2\leq\frac{1}{2}\|x^*-x_0\|^2+B_k(x^*)+\delta_k \tag{1.11}$$

and

$$\phi(x_k)-\phi(x^*)\leq\frac{1}{A_k}\left[\frac{1}{2}\|x^*-x_0\|^2+B_k(x^*)+\delta_k\right]. \tag{1.12}$$

Proof. By ($\mathcal{R}_k^1$) and ($\mathcal{R}_k^2$), we have

$$\begin{aligned}
A_k\phi(x_k)+\frac{1+\mu A_k}{2}\|x-v_k\|^2 &\overset{(\mathcal{R}_k^1)}{\leq} \psi_k^*+\delta_k+\frac{1+\mu A_k}{2}\|x-v_k\|^2\leq\psi_k(x)+\delta_k\\
&\overset{(\mathcal{R}_k^2)}{\leq} A_k\phi(x)+\frac{1}{2}\|x-x_0\|^2+B_k(x)+\delta_k,
\end{aligned}$$

where the second inequality follows from the $(1+\mu A_k)$-strong convexity of ψ_k. Letting $x=x^*$ in the above inequality gives the desired results. □

To fully determine the convergence rate, we still need to bound the error terms $B_k(x^*)$ and δ_k. To establish such bounds, we assume the following condition on the gradient error.

Assumption 1. For each iteration $k \geq 0$ in Algorithm 1, the gradient error bound satisfies

$$\tau_k \leq \min\left\{\frac{\varepsilon}{2}, \quad \frac{\mu \cdot \theta^{k+1}}{A_{k+1}}\right\}, \tag{1.13}$$

for some $\theta \in [0, 1)$.

> **Remark 1.** Notice that if $\theta = 0$ in (1.13), there is no gradient error, and Algorithm 1 reduces to the exact version of Nesterov's APG.

The setting of a_k and A_k is the same as that in [1]. Hence, we directly have the following result from Lemma 8 in [1].

Lemma 4. The coefficient sequence $\{A_k\}$ from Algorithm 1 satisfies

$$A_k \geq \max\left\{\frac{k^2}{2\gamma_u L_f}, \frac{2}{\gamma_u L_f}\left(1 + \sqrt{\frac{\mu}{2\gamma_u L_f}}\right)^{2(k-1)}\right\}, \forall\, k \geq 1. \tag{1.14}$$

Under Assumption 1, we have from (1.10) and the Cauchy-Schwarz inequality that

$$B_{k+1}(x^*) \leq B_k(x^*) + a_{k+1}\|e(x_{k+1})\| \cdot \|x^* - x_{k+1}\| \leq B_k(x^*) + \mu\, \theta^{k+1}\, \|x^* - x_{k+1}\|$$

where the second inequality follows from (1.13) and $a_{k+1} \leq A_{k+1}$. Recursively applying the above inequality gives

$$B_k(x^*) \leq \mu \sum_{t=1}^{k} \theta^t\, \|x^* - x_t\|, \qquad \forall\, k \geq 1. \tag{1.15}$$

In addition, by the definition of δ_k in (1.5), it follows

$$\begin{aligned}
\delta_{k+1} &= \delta_k + A_k\tau_k\|x_k - x_{k+1}\| + A_{k+1}\left(6\tau_k\|y_k - x_{k+1}\| + \frac{4\tau_k^2}{M_k}\right) \\
&\leq \delta_k + \mu\, \theta^{k+1}\|x_k - x_{k+1}\| + 6\mu\, \theta^{k+1}\|y_k - x_{k+1}\| + \frac{2\mu^2\gamma_u L_f \theta^{2k+2}}{M_k},
\end{aligned}$$

where we have used $A_{k+1} \geq \frac{2}{\gamma_u L_f}$ from (1.14) to obtain the inequality. Hence, applying the above inequality recursively and noting $M_k \geq L_{\min}$, we have

$$\delta_{k+1} \leq \mu \sum_{t=0}^{k}\left(\theta^{t+1}\|x_t - x_{t+1}\| + 6\theta^{t+1}\|y_t - x_{t+1}\| + \frac{2\mu\gamma_u L_f \theta^{2t+2}}{L_{\min}}\right), \ \forall\, k \geq 0. \tag{1.16}$$

1.3.1 Bound on iterates

We show that all involved points in Algorithm 1 are bounded.

Lemma 5. For all $k \geq 0$, it holds that

$$\|y_k - x^*\| \leq R, \qquad \|v_k - x^*\| \leq R, \qquad \|x_k - x^*\| \leq R, \tag{1.17}$$

where

$$\begin{aligned}
R = \max\Bigg\{ & 2\|x^* - x_0\| + \sqrt{\frac{16\theta^2\mu^2\gamma_u L_f}{L_{\min}(1-\theta^2)}}, \frac{120\theta\mu}{1-\theta}, \frac{60\theta\gamma_u L_f}{(1-\theta)}, \\
& \frac{\sqrt{2\gamma_u L_f}\|x^* - x_0\|}{\sqrt{\mu}} + 16\theta\gamma_u L_f + \sqrt{\frac{8\theta^2\mu(\gamma_u L_f)^2}{L_{\min}(1-\theta^2)}}\Bigg\}.
\end{aligned} \tag{1.18}$$

Proof. We prove the result by induction. Notice $y_0 = v_0 = x_0$. Hence, (1.17) holds for $k = 0$. Suppose that for some $k \geq 0$, it holds

$$\|y_t - x^*\| \leq R,\ \|v_t - x^*\| \leq R,\ \|x_t - x^*\| \leq R,\ \forall\, t = 0, \ldots, k. \tag{1.19}$$

In the following, we show (1.17) also holds for $k+1$.

First, by the μ-strong convexity of ϕ, it holds $\frac{\mu}{2}\|x_{k+1} - x^*\|^2 \leq \phi(x_{k+1}) - \phi(x^*)$. Hence, from (1.12), (1.15), and (1.16), it follows

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leq \frac{1}{\mu A_{k+1}}\left[\|x^* - x_0\|^2 + 2B_{k+1}(x^*) + 2\delta_{k+1}\right] \\
&\leq \frac{1}{\mu A_{k+1}}\|x^* - x_0\|^2 + \frac{2}{A_{k+1}}\sum_{t=1}^{k+1}\theta^t\|x^* - x_t\| \\
&\quad + \frac{2}{A_{k+1}}\sum_{t=0}^{k}\left(\theta^{t+1}\|x_t - x_{t+1}\| + 6\theta^{t+1}\|y_t - x_{t+1}\| + \frac{2\mu\gamma_u L_f\theta^{2t+2}}{L_{\min}}\right) \\
&\leq \frac{1}{\mu A_{k+1}}\|x^* - x_0\|^2 + \frac{2}{A_{k+1}}\left(\frac{\theta R}{1-\theta} + \theta^{k+1}\|x^* - x_{k+1}\|\right) \\
&\quad + \frac{2}{A_{k+1}}\left(\frac{2\theta R}{1-\theta} + \theta^{k+1}\|x^* - x_{k+1}\| + \frac{12\theta R}{1-\theta} + 6\theta^{k+1}\|x^* - x_{k+1}\| + \frac{2\mu\gamma_u L_f\theta^2}{L_{\min}(1-\theta^2)}\right) \\
&= \frac{16\theta^{k+1}}{A_{k+1}}\|x^* - x_{k+1}\| + \frac{1}{\mu A_{k+1}}\|x^* - x_0\|^2 + \frac{2}{A_{k+1}}\left(\frac{15\theta R}{1-\theta} + \frac{2\mu\gamma_u L_f\theta^2}{L_{\min}(1-\theta^2)}\right).
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
\|x_{k+1} - x^*\| &\leq \sqrt{\frac{1}{\mu A_{k+1}}\|x^* - x_0\|^2 + \frac{2}{A_{k+1}}\left(\frac{15\theta R}{1-\theta} + \frac{2\mu\gamma_u L_f\theta^2}{L_{\min}(1-\theta^2)}\right) + \left(\frac{8\theta^{k+1}}{A_{k+1}}\right)^2} + \frac{8\theta^{k+1}}{A_{k+1}} \\
&\leq \frac{\|x^* - x_0\|}{\sqrt{\mu A_{k+1}}} + \frac{16\theta^{k+1}}{A_{k+1}} + \sqrt{\frac{4\mu\gamma_u L_f\theta^2}{A_{k+1}L_{\min}(1-\theta^2)}} + \sqrt{\frac{30\theta R}{A_{k+1}(1-\theta)}} \\
&\leq \frac{\sqrt{\gamma_u L_f}\|x^* - x_0\|}{\sqrt{2\mu}} + 8\theta\gamma_u L_f + \sqrt{\frac{2\theta^2\mu(\gamma_u L_f)^2}{L_{\min}(1-\theta^2)}} + \sqrt{\frac{15\theta\gamma_u L_f R}{(1-\theta)}} \\
&\leq R,
\end{aligned} \tag{1.20}$$

where the third inequality holds because $A_{k+1} \geq \frac{2}{\gamma_u L_f}, \forall\, k \geq 0$ and $\theta \in (0,1)$, and the last one follows from the choice of R.

Second, plugging (1.15) and (1.16) into (1.11) gives

$$\begin{aligned}
\frac{1 + \mu A_{k+1}}{2}\|x^* - v_{k+1}\|^2 &\leq \frac{1}{2}\|x^* - x_0\|^2 + \mu\sum_{t=1}^{k+1}\theta^t\|x^* - x_t\| \\
&\quad + \mu\sum_{t=0}^{k}\left(\theta^{t+1}\|x_t - x_{t+1}\| + 6\theta^{t+1}\|y_t - x_{t+1}\| + \frac{2\mu\gamma_u L_f\theta^{2t+2}}{L_{\min}}\right) \\
&\leq \frac{1}{2}\|x^* - x_0\|^2 + \mu\left(\frac{\theta R}{1-\theta} + \frac{2\theta R}{1-\theta} + \frac{12\theta R}{1-\theta} + \frac{2\mu\gamma_u L_f\theta^2}{L_{\min}(1-\theta^2)}\right)
\end{aligned} \tag{1.21}$$

where the second inequality follows from (1.19), (1.20), and the triangle inequality. Hence, by

$1 + \mu A_{k+1} > 1$, it follows

$$\|x^* - v_{k+1}\| \le \|x^* - x_0\| + \sqrt{\frac{4\theta^2\mu^2\gamma_u L_f}{L_{\min}(1-\theta^2)}} + \sqrt{\frac{30\theta\mu R}{1-\theta}} \le R,$$

where the second inequality holds by the choice of R.

Third, y_{k+1} is a convex combination of x_{k+1} and v_{k+1}, and thus $\|x^* - y_{k+1}\| \le R$. This finishes the induction and completes the proof. □

By Lemma 5, we have from (1.15) that

$$B_k(x^*) \le \frac{\theta\mu R}{1-\theta}, \ \forall\, k \ge 0, \tag{1.22}$$

and from (1.16) that

$$\delta_k \le \frac{14\theta\mu R}{1-\theta} + \frac{2\theta^2\mu^2\gamma_u L_f}{L_{\min}(1-\theta^2)}, \ \forall\, k \ge 0. \tag{1.23}$$

Therefore, plugging (1.22) and (1.23) into (1.12) gives

$$\phi(x_k) - \phi(x^*) \le \frac{E_{0,R}}{A_k} := \frac{1}{A_k}\left[\frac{1}{2}\|x^* - x_0\|^2 + \frac{15\theta\mu R}{1-\theta} + \frac{2\theta^2\mu^2\gamma_u L_f}{L_{\min}(1-\theta^2)}\right], \ \forall\, k \ge 1. \tag{1.24}$$

1.3.2 Iteration Complexity

We have from (1.24) with the μ-strong convexity of ϕ and from (1.21) with $\theta \in (0,1)$ that

$$\|x_k - x^*\| \le \sqrt{\frac{2E_{0,R}}{\mu A_k}}, \qquad \|v_k - x^*\| \le \sqrt{\frac{2E_{0,R}}{\mu A_k}}, \qquad \forall\, k \ge 1. \tag{1.25}$$

Since y_k is a convex combination of x_k and v_k, it also holds that

$$\|y_k - x^*\| \le \sqrt{\frac{2E_{0,R}}{\mu A_k}}, \qquad \forall\, k \ge 1. \tag{1.26}$$

Hence, by the triangle inequality and L_f-smoothness of f, it follows that

$$\begin{aligned}
&\|\tilde\nabla f(x_{k+1}) - \tilde\nabla f(y_k) - M_k(x_{k+1} - y_k)\| \\
&\le (L_f + M_k)\|x_{k+1} - y_k\| + \|e(x_{k+1})\| + \|e(y_k)\| \\
&\le (L_f + M_k)\|x_{k+1} - y_k\| + 2\tau_k \\
&\le (L_f + \gamma_u L_f)\left(\sqrt{\frac{2E_{0,R}}{\mu A_{k+1}}} + \sqrt{\frac{2E_{0,R}}{\mu A_k}}\right) + 2\tau_k,
\end{aligned} \tag{1.27}$$

where the last inequality holds by (1.25) and (1.26) and $M_k \le \gamma_u L_f$. Therefore, Algorithm 1 will return x_{k+1} as the output if

$$2(L_f + \gamma_u L_f)\sqrt{\frac{2E_{0,R}}{\mu A_k}} + 2\tau_k \le \frac{\varepsilon}{2}. \tag{1.28}$$

On the other side, from the update of x_{k+1}, it holds

$$0 \in \tilde\nabla f(y_k) + M_k(x_{k+1} - y_k) + \partial\Psi(x_{k+1}).$$

Hence, $\nabla f(x_{k+1}) - \tilde\nabla f(y_k) - M_k(x_{k+1} - y_k) \in \partial\phi(x_{k+1})$, and thus

$$\begin{aligned}
\mathrm{dist}\big(0, \partial\phi(x_{k+1})\big) &\le \|\nabla f(x_{k+1}) - \tilde\nabla f(y_k) - M_k(x_{k+1} - y_k)\| \\
&\le \|\tilde\nabla f(x_{k+1}) - \tilde\nabla f(y_k) - M_k(x_{k+1} - y_k)\| + \|e(x_{k+1})\|
\end{aligned}$$

$$\leq \|\tilde{\nabla} f(x_{k+1}) - \tilde{\nabla} f(y_k) - M_k(x_{k+1} - y_k)\| + \tau_k. \tag{1.29}$$

Therefore, when Algorithm 1 stops, x_{k+1} is an ε-stationary solution of ϕ, since $\tau_k \leq \frac{\varepsilon}{2}$ from (1.13).

Summarizing the above analysis, we give the iteration complexity result of Algorithm 1 to produce an ε-stationary solution as follows.

Theorem 2 (Iteration Complexity). Given $\varepsilon > 0$, under Assumption 1, Algorithm 1 will produce an ε-stationary solution of ϕ within k iterations, where

$$k = \max\left\{\left\lceil \frac{\log \frac{4\mu\gamma_u L_f}{\varepsilon}}{\log \frac{1}{\theta}} \right\rceil, \quad \left\lceil \frac{\log 8\sqrt{\frac{\gamma_u L_f E_{0,R}}{\mu}} \frac{(1+\gamma_u)L_f}{\varepsilon}}{\log\left(1 + \sqrt{\frac{\mu}{2\gamma_u L_f}}\right)} \right\rceil + 1\right\}. \tag{1.30}$$

In particular, if exact gradients of f are used in the algorithm, i.e., $\theta = 0$ in (1.3), the total number of iterations reduces to

$$k = \left\lceil \frac{\log 8\sqrt{\frac{\gamma_u L_f \|x^* - x_0\|^2}{2\mu}} \frac{(1+\gamma_u)L_f}{\varepsilon}}{\log\left(1 + \sqrt{\frac{\mu}{2\gamma_u L_f}}\right)} \right\rceil + 1. \tag{1.31}$$

Proof. It suffices to solve the inequality in (1.28) for k. From the condition on k in (1.30), we have $\theta^{k+1} \leq \frac{\varepsilon}{4\mu\gamma_u L_f}$, and thus by (1.13), it follows

$$2\tau_k \leq \frac{\varepsilon}{2\gamma_u L_f A_{k+1}} \leq \frac{\varepsilon}{4}, \tag{1.32}$$

where the second inequality is obtained by using $A_{k+1} \geq \frac{2}{\gamma_u L_f}$ from (1.14).

In addition, we have from (1.30) that

$$\left(1 + \sqrt{\frac{\mu}{2\gamma_u L_f}}\right)^{k-1} \geq 8\sqrt{\frac{\gamma_u L_f E_{0,R}}{\mu}} \frac{(1+\gamma_u)L_f}{\varepsilon}.$$

Hence, it follows from (1.14) that

$$\sqrt{A_k} \geq 8\sqrt{\frac{2}{\gamma_u L_f}}\sqrt{\frac{\gamma_u L_f E_{0,R}}{\mu}} \frac{(1+\gamma_u)L_f}{\varepsilon},$$

and thus

$$2(1+\gamma_u)L_f\sqrt{\frac{2E_{0,R}}{\mu A_k}} \leq \frac{\varepsilon}{4}. \tag{1.33}$$

Now adding (1.32) and (1.33) gives (1.28).

Finally, notice that when $\theta = 0$, the inequality in (1.32) holds for any k because $\tau_k = 0$, and from (1.24), it follows $E_{0,R} = \frac{1}{2}\|x^* - x_0\|^2$. Thus we have (1.31) from (1.30) and complete the proof. □

Remark 2. Because each while-loop must exit in at most $\left\lceil \log_{\gamma_u} \frac{L_f}{L_{\min}} \right\rceil + 1$ steps, the total number of inexact gradient evaluations will be $k\left\lceil \log_{\gamma_u} \frac{L_f}{L_{\min}} \right\rceil + k$, where k is given in (1.30). Suppose $\frac{\mu}{L_f} \ll 1$. The complexity is $\mathcal{O}\left(\sqrt{\frac{L_f}{\mu}} \log \frac{L_f}{\mu\varepsilon}\right)$ in either case of using exact or inexact gradients of f.

Recommended Resources

Articles

[1] Yu Nesterov. "Gradient methods for minimizing composite functions". In: *Mathematical programming* 140.1 (2013), pp. 125–161. (pp. 1, 2, 4, 5)

[2] Mark Schmidt, Nicolas Roux, and Francis Bach. "Convergence rates of inexact proximal-gradient methods for convex optimization". In: *Advances in neural information processing systems* 24 (2011). (p. 1)

[3] Qihang Lin and Yangyang Xu. "Reducing the complexity of two classes of optimization problems by inexact accelerated proximal gradient method". In: *SIAM Journal on Optimization* 33.1 (2023), pp. 1–35. (p. 1)