Stochastic gradient descent algorithms for strongly convex functions at $O(1/T)$ convergence rates

Shenghuo Zhu
zsh@nec-labs.com

Abstract

With a weighting scheme proportional to t, a traditional stochastic gradient descent (SGD) algorithm achieves a high probability convergence rate of $O(\kappa/T)$ for strongly convex functions, instead of $O(\kappa \ln(T)/T)$. We also prove that an accelerated SGD algorithm also achieves a rate of $O(\kappa/T)$.

1 Introduction

Consider a stochastic optimization problem

$$\min_{x \in \mathcal{X}} \{ f(x) := \mathbb{E}_\xi F(x, \xi) \}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a nonempty bounded closed convex set, ξ is a random variable, F is a smooth convex function, f is a smooth strongly-convex function. The requirement of smoothness simplifies the analysis. If the objective function is nonsmooth but satisfies Lipschitz continuity, stochastic gradient descent algorithms can replace gradients with subgradients, but the analysis has to introduce an additional term in the same order as the variance term. Some nonsmooth cases have been studied in (Lan, 2008) and (Ghadimi & Lan, 2012).

Assume that the domain is bounded, i.e. $\sup_{x,y\in\mathcal{X}} \|x-y\|^2 \le D^2$. Let $G(x,\xi)$ be a stochastic gradient of function f at x with a random variable ξ. Then $g(x) := \mathbb{E}_\xi G(x,\xi)$ is a gradient of $f(x)$. Assume that $\|g(x)-g(y)\|_* \le L\|x-y\|$, where L is known as the Lipschitz constant. We only consider strongly convex function in this note, thus assume that there is $\mu > 0$, such that $f(y) \ge f(x) + \langle g(x), y-x\rangle + \frac{\mu}{2}\|y-x\|^2$. We assume that stochastic gradients are bounded, i.e., there exists $Q > 0$, such that $\sup_\xi \|G(x,\xi) - g(x)\|_* \le Q$.

We are interested in the conditional number κ, which is defined as L/μ. The conditional number, κ, could be as large as $\sqrt{N}$, where N is the number of samples and $T = N$. One reference case is regularized linear classifiers (Smale & Zhou, 2003), where the regularization factor could be as large as $\sqrt{N}$. The other reference case is the conditional number of a $N \times n$ random matrix (Rudelson & Vershynin, 2009), where the smallest singular value is $O(\sqrt{N} - \sqrt{n-1})$. When $\kappa = \Theta(\sqrt{T})$, $O(\kappa/T) = O(1/\sqrt{T})$, which bridges the gap between the convergence rate for strongly convex functions and that for those without strongly convex condition. In this note, we assume $\kappa = O(T)$. We use big-O notation in term of T and κ and hide the factors D^2L, Q^2/L and DQ besides constants.

Notation

Denote $\{1 \cdots T\}$ by $[T]$. Let $\{\xi_t : t \in [T]\}$ be a sequence of independent random variables. Denote $\mathbb{E}_{|t-1}\{\cdot\} := \mathbb{E}\{\cdot|\xi_1, \cdots, \xi_{t-1}\}$. We define $\widetilde{\ln}(T,t) = \sum_{\tau=t+1}^T \frac{1}{\tau}$. Then $\widetilde{\ln}(T,t) \le \frac{1}{t+1} + \ln(T/(t+1))$, and for $t \ge 1$, $\widetilde{\ln}(T,t) \le \ln(T/t)$.

2 Stochastic gradient descent algorithm

Algorithm 1 shows the stochastic gradient descent method. Unlike the conventional averaging by equal weights $w_t = 1/T$, we use a weighting scheme $w_t = \alpha_t \prod_{\tau=t+1}^T (1-\alpha_\tau) = t/(2T(T+1))$, where $\alpha_t = 2/(t+1)$. Theorem 1 shows a convergence rate of $O(\kappa/T)$, assuming that $T > \kappa$. Let $A_t = \|x_t - x_*\|^2$, $B_t = \langle \delta_t, x_{t-1} - x_* \rangle / Q$, $C_t = \|\delta_t\|_*^2/Q^2$,

Algorithm 1 Stochastic gradient descent algorithm

```
Input: initial solution $x_0$, step sizes $\{\gamma_t > 0 : t \in [T]\}$ and averaging factor $\{\alpha_t > 0 : t \in [T]\}$.
for $t \in [T]$ do
    Let sample gradient $\hat{g}_k = G(x_{t-1}, \xi_t)$, where $\xi_t$ is independent from $\{\xi_\tau : \tau \in [t-1]\}$.
    Let $x_t = \arg\min_{x \in \mathcal{X}} \left\{ \langle \hat{g}_t, x \rangle + \frac{1}{2\gamma_t} \|x - x_{t-1}\|^2 \right\}$;
    Set $\bar{x}_t = \bar{x}_{t-1} + \alpha_t (x_t - \bar{x}_{t-1})$;
end for
Output: $\bar{x}_T$.
```

and the coefficients $b_t = O(1)$ and $c_t = O(1/t)$. The informal argument is that the weighting scheme equalizes the variance of each iteration, since $\mathrm{var}(b_t B_t)$ and $c_t C_t$ are $O(1/t)$ assuming that $A_t = O(1/t)$.

Theorem 1. *Assume that the underlying function f is strongly convex, i.e., $\mu > 0$. Let $\kappa = L/\mu$. If $\alpha_t = \frac{2}{t+1}$, $\gamma_t = \frac{2}{\mu(t+2\kappa)}$, then it holds for Algorithm 1 that for $\theta > 0$,*

$$\Pr\{f(\bar{x}_T) - f(x_*) \geq \bar{K}(T) + \sqrt{2\theta}\tilde{K}(T) + \theta\hat{K}(T)\} \leq \exp\{-\theta\}, \tag{1}$$

where

$$\begin{aligned}
\bar{K}(T) &:= \frac{D^2 L}{T} + \frac{2\kappa Q^2}{LT} = O(\kappa/T), \\
\tilde{K}(T) &:= \frac{4DQ(\kappa+1)}{T^{3/2}} + \frac{2\sqrt{2}\kappa Q^2}{LT} + \frac{4\sqrt{2}\kappa^{3/2} Q^2 \sqrt{1+\ln T}}{LT^{3/2}} = O(\kappa/T), \\
\hat{K}(T) &:= \frac{10\kappa Q^2}{LT} = O(\kappa/T).
\end{aligned}$$

Similarly with traditional equal weighting scheme, $w_t = 1/T$, we have a convergence rate of $O(\kappa \ln(T)/T)$ in Proposition 2. Informally, $\mathrm{var}(\sum_t w_t b_t B_t) = \ln(T)/T$ implies a convergence rate of $O(\ln(T)/T)$.

Proposition 2. *Assume that $\mu > 0$. Let $\kappa = L/\mu$. If $\alpha_t = \frac{1}{t}$, $\gamma_t = \frac{1}{\mu(t+\kappa)}$, then for $\theta > 0$,*

$$\Pr\{f(\bar{x}_T) - f(x_*) \geq \bar{K}(T) + \sqrt{2\theta}\tilde{K}(T) + \theta\hat{K}(T)\} \leq \exp\{-\theta\},$$

where

$$\bar{K}(T) := \frac{LD^2}{2T} + \frac{\kappa Q^2}{2LT}(1 + \ln T), \qquad \tilde{K}(T) := \frac{DQ\sqrt{\kappa+1}}{T} + \frac{\kappa Q^2}{LT}\sqrt{1+\ln T}, \qquad \hat{K}(T) := \frac{6\kappa Q^2}{LT}.$$

Proposition 3 shows that if the optimal solution x_* is an interior point, it is possible to simply take the non-averaged solution, x_T. The convergence rate is $O(\kappa^2/T)$. However, if $\kappa = \Theta(\sqrt{T})$, $O(\kappa^2/T)$ means not convergent, just like the non-averaged SGD solution without strongly convex conditions.

Proposition 3. *Assume that $\mu > 0$ and the optimal solution x_* is an interior point. Let $\kappa = L/\mu$. If $\gamma_t = \frac{1}{\mu(t+\kappa)}$, then for $\theta > 0$,*

$$\Pr\{f(x_T) - f(x_*) \geq \bar{K}(T) + \sqrt{2\theta}\tilde{K}(T) + \theta\hat{K}(T)\} \leq \exp\{-\theta\},$$

where

$$\begin{aligned}
\bar{K}(T) &:= \frac{D^2 L(\kappa+1)^2}{2(T+\kappa)^2} + \frac{\kappa^2 Q^2 (T + \kappa(1+\ln T))}{2L(T+\kappa)^2} = O(\kappa^2/T), \\
\tilde{K}(T) &:= \frac{DQ(\kappa+1)^2}{\sqrt{2}(T+\kappa)^{3/2}} + \frac{\kappa^2 Q^2}{2L(T+\kappa)} + \frac{\kappa^2 Q^2 \sqrt{\kappa T(1+\ln(T))}}{2L(T+\kappa)^2} = O(\kappa^2/T), \\
\hat{K}(T) &:= \frac{6\kappa^2 Q^2}{L(T+\kappa)} = O(\kappa^2/T).
\end{aligned}$$

Remark 1. *There are studies on the high probability convergence rate of stochastic algorithm on strongly convex functions, such as (Rakhlin et al., 2012). The convergence rate usefully is $O(\mathrm{polylog}(T)/T)$. Here, we prove a convergence rate of $O(\frac{\kappa}{T})$ with proper weighting scheme.*

3 Accelerated Stochastic Gradient Descent Algorithm

Algorithm 2 Accelerated Stochastic Gradient Descent algorithm

```
Input: $x_0$, $\mu$, $\{\alpha_t \geq 0\}$, $\{\gamma_t > 0\}$;
Let $\bar{x}_0 = x_0$;
for $k \in [T]$ do
    Let $y_{t-1} = \alpha_t x_{t-1} + (1-\alpha_t)\bar{x}_{t-1}$;
    Let $\hat{g}_t = G(y_{t-1}, \xi_t)$, where $\{\xi_t\}$ is a sample;
    Let $x_t = \arg\min_{x \in \mathcal{X}} \left\{ \langle \hat{g}_t - \mu(y_{t-1} - x_{t-1}), x \rangle + \frac{1}{2\gamma_t}\|x - x_{t-1}\|^2 \right\}$;
    Set $\bar{x}_t = \bar{x}_{t-1} + \alpha_t(x_t - \bar{x}_{t-1})$;
end for
Output: $\bar{x}_t$.
```

Algorithm 2 is a stochastic variant of Nesterov's accelerated methods. The convergence rate is also $O(\kappa/T)$. Comparing with Theorem 1, the determinant part in Theorem 4 have a better rate, i.e. $\frac{LD^2}{T^2}$.

Theorem 4. *Assume that $\mu > 0$. If $\alpha_t = \frac{2}{t+1}$, $\gamma_t = \frac{1}{\mu(2\kappa/t + 1/\alpha_t)}$, then for $\theta > 0$,*

$$\Pr\{f(\bar{x}_T) - f(x_*) > \bar{K}(T) + \sqrt{2\theta}\tilde{K}(T) + \theta\hat{K}(T)\} \leq \exp\{-\theta\},$$

where

$$\bar{K}(T) := \frac{2D^2L}{T^2} + \frac{2\kappa Q^2}{LT}, \qquad \tilde{K}(T) := \frac{\sqrt{20\kappa}DQ}{T^{3/2}} + \frac{\sqrt{10}\kappa Q^2}{2LT}, \qquad \hat{K}(T) := \frac{8\kappa Q^2}{LT}.$$

Remark 2. *The paper (Ghadimi & Lan, 2012) has its strongly convex version for AC-SA for sub-Gaussian gradient assumption, but its proof relies on a multi-stage algorithm.*

Although SAGE (Hu et al., 2009) also provided a stochastic algorithm based on Nesterov's method for strongly convexity, the high probability bound was not given in the paper.

4 A note on weighting schemes

In this study, we find the interesting property of weighting scheme with $\alpha_t = \frac{2}{t+1}$, i.e. $w_t = \frac{2t}{T(T+1)}$. The scheme takes advantage of a sequence with variance at the decay rate of $\frac{1}{t}$. Now let informally investigate a sequence with homogeneous variance, say 1. With a constant weighting scheme, $\alpha_t = 1/t$, i.e. $w_t = 1/T$, the averaged variance is $1/T$. With an exponential weighting scheme, $\alpha_1 = 1$, $\alpha_t = \alpha$, i.e. $w_1 = (1-\alpha)^{T-1}$ and $w_t = \alpha(1-\alpha)^{T-t}$, the averaged variance is $\frac{\alpha}{2-\alpha}(1 + (1-\alpha)^{2T-1}) \approx \frac{\alpha}{2-\alpha}$, which is translated to that the number of effective tail samples is a constant $\frac{2}{\alpha} - 1$. With the weighting scheme $\alpha_t = \frac{2}{t+1}$ or $w_t = 2t/(T(T+1))$, the averaged variance is $\frac{2(2T+1)}{3T(T+1)} \approx \frac{4}{3T}$, which is translated to $\frac{3T}{4}$ effective tail samples. This is a trade-off between sample efficiency and recency. To make other trade-offs, We can use a generalized scheme[1], $\alpha_t = \frac{t^r}{\sum_{\tau=1}^{t} \tau^r}$ or $w_t = \frac{t^r}{\sum_{\tau=1}^{T} \tau^r}$. Then the averaged variance is approximately $\frac{(1+r)^2}{(1+2r)T}$.

5 Proofs

The proof strategy is first to construct inequalities from the algorithms in Lemma 6 and 7, then to apply Lemma 5 to derive the probability inequalities.

[1] An alternative scheme is $\alpha_t = \frac{1+r}{t+r}$ or $w_t = \frac{(1+r)\Gamma(t+r;t)}{\Gamma(T+r+1;T)}$, where $\Gamma(T;t) := \Gamma(T)/\Gamma(t)$.

Lemma 5. *Assume that B_t is martingale difference, $w_t \geq 0$, $\tilde{a}_t \geq 0$, $\tilde{c}_t \geq 0$, $a_t \geq 0$, $c_t \geq 0$, $d_t > 0$, $A_0 \leq D^2$, $A_t \geq 0$, and*

$$\begin{aligned} X_t &= w_t(\tilde{a}_t A_{t-1} + 2\tilde{b}_t B_t + \tilde{c}_t C_t), && (2) \\ A_t &\leq d_t(a_t A_{t-1} + 2b_t B_t + c_t C_t), && (3) \\ B_t^2 &\leq A_{t-1} C_t, \\ C_t &\leq 1. \end{aligned}$$

If the following conditions hold

1. *for $u \in (0, \frac{1}{2\hat{R}_T})$,*
$$\mathbb{E}_{|T} \exp(u X_{T+1}) \leq \exp((u\bar{P}_T + \frac{2u^2\tilde{P}_T^2}{1-u\hat{R}_T})A_T + u\bar{R}_T + \frac{2u^2\tilde{R}_T^2}{1-u\hat{R}_T}), \tag{4}$$

2. *for $t \in [T]$,*
$$\begin{aligned} & a_t d_t \bar{P}_t + w_t \tilde{a}_t \leq \bar{P}_{t-1}, \\ & \bar{R}_t + w_t \tilde{c}_t + c_t d_t \bar{P}_t \leq \bar{R}_{t-1}, \\ & a_t d_t \tilde{P}_t^2 + 4(w_t \tilde{b}_t + b_t d_t \bar{P}_t)^2 \leq \tilde{P}_{t-1}^2, \\ & \tilde{R}_t^2 + c_t d_t \tilde{P}_t^2 \leq \tilde{R}_{t-1}^2, \\ & \hat{R}_t \leq \hat{R}_{t-1}, \\ & a_t d_t \tilde{P}_t^2 \hat{R}_t + 4b_t d_t(w_t \tilde{b}_t + b_t d_t \bar{P}_t)\tilde{P}_t^2 \leq \tilde{P}_{t-1}^2 \hat{R}_{t-1}, \\ & a_t d_t \tilde{P}_t^2 \hat{R}_t^2 + 4b_t d_t(w_t \tilde{b}_t + b_t d_t \bar{P}_t)\tilde{P}_t^2 \hat{R}_t + 2b_t^2 d_t^2 \tilde{P}_t^4 \leq \tilde{P}_{t-1}^2 \hat{R}_{t-1}^2, \end{aligned} \tag{5}$$

then for $\theta > 0$,

$$\Pr\{\sum_{t=1}^{T+1} X_t \geq \bar{P}_0 D^2 + \bar{R}_0 + \sqrt{2\theta}(\tilde{P}_0 D^2 + \tilde{R}_0) + 2\theta\hat{R}_0\} \leq \exp\{-\theta\}. \tag{6}$$

Proof. We will prove the following inequality by induction,

$$\mathbb{E}_{|t} \exp(u \sum_{\tau=t+1}^{T+1} X_\tau) \leq \exp((u\bar{P}_t + \frac{2u^2\tilde{P}_t^2}{1-u\hat{R}_t})A_t + u\bar{R}_t + \frac{2u^2\tilde{R}_t^2}{1-u\hat{R}_t}), \quad \forall u \in (0, \frac{1}{2\hat{R}_t}). \tag{7}$$

Eq. 4 implies that Eq. (7) holds for $t = T$. For $u \in (0, \frac{1}{2\hat{R}_{t-1}})$,

$$\mathbb{E}_{|t-1} \exp(u \sum_{\tau=t}^{T+1} X_\tau) \le \mathbb{E}_{|t-1} \exp(uX_t + (u\bar{P}_t + \frac{u^2 \tilde{P}_t^2}{2(1-u\hat{R}_t)})A_t + u\bar{R}_t + \frac{u^2 \tilde{R}_t^2}{2(1-u\hat{R}_t)}) \tag{8}$$

$$\le \mathbb{E}_{|t-1} \exp(uw_t(\tilde{a}_t A_{t-1} + 2\tilde{b}_t B_t + \tilde{c}_t C_t) + (u\bar{P}_t + \frac{u^2 \tilde{P}_t^2}{2(1-u\hat{R}_t)})d_t(a_t A_{t-1} + 2b_t B_t + c_t C_t) + u\bar{R}_t + \frac{u^2 \tilde{R}_t^2}{2(1-u\hat{R}_t)}) \tag{9}$$

$$\le \exp((u(\bar{P}_t d_t a_t + p_t \tilde{a}_t) + \frac{u^2 \tilde{P}_t^2 d_t a_t}{2(1-u\hat{R}_t)})A_{t-1} + u(\bar{R}_t + p_t c_t + \bar{P}_t d_t c_t) + \frac{u^2 \tilde{R}_t^2}{2(1-u\hat{R}_t)} + \frac{u^2 \tilde{P}_t^2 d_t c_t}{2(1-u\hat{R}_t)}) \tag{10}$$

$$\times \mathbb{E}_{|t-1} \exp(2u(w_t \tilde{b}_t + b_t d_t \bar{P}_t + \frac{u b_t d_t \tilde{P}_t^2}{2(1-u\hat{R}_t)})B_t)$$

$$\le \exp((u(\bar{P}_t d_t a_t + p_t \tilde{a}_t) + \frac{u^2 \tilde{P}_t^2 d_t a_t}{2(1-u\hat{R}_t)} + 2u^2(w_t \tilde{b}_t + b_t d_t \bar{P}_t + \frac{u b_t d_t \tilde{P}_t^2}{2(1-u\hat{R}_t)})^2)A_{t-1}$$

$$+ u(\bar{R}_t + w_t \tilde{c}_t + \bar{P}_t d_t c_t) + \frac{u^2 \tilde{R}_t^2}{2(1-u\hat{R}_t)} + \frac{u^2 \tilde{P}_t^2 d_t c_t}{2(1-u\hat{R}_t)}) \tag{11}$$

$$\le \exp((u(\bar{P}_t d_t a_t + p_t \tilde{a}_t) + \frac{u^2 \tilde{P}_t^2 d_t a_t}{2(1-u\hat{R}_t)} + 2u^2(w_t \tilde{b}_t + b_t d_t \bar{P}_t)^2 + \frac{u^3(w_t \tilde{b}_t + b_t d_t \bar{P}_t) b_t d_t \tilde{P}_t^2}{2(1-u\hat{R}_t)} + \frac{2u^4 b_t^2 d_t^2 \tilde{P}_t^4}{2(1-u\hat{R}_t)})A_{t-1}$$

$$+ u(\bar{R}_t + w_t \tilde{c}_t + \bar{P}_t d_t c_t) + \frac{u^2 \tilde{R}_t^2}{2(1-u\hat{R}_t)} + \frac{u^2 \tilde{P}_t^2 d_t c_t}{2(1-u\hat{R}_t)}) \tag{12}$$

$$\le \exp((u\bar{P}_{t-1} + \frac{u^2 \tilde{P}_{t-1}^2}{2(1-u\hat{R}_{t-1})})A_{t-1} + u\bar{R}_{t-1} + \frac{u^2 \tilde{R}_{t-1}^2}{2(1-u\hat{R}_{t-1})}), \tag{13}$$

where Eq. (8) is due to the assumption of induction; Eq. (9) is due to Eq. (2,3); Eq. (10) is due to $C_t \le 1$; Eq. (11) is due to $\mathbb{E}_{|t-1} B_t = 0$, $B_t^2 \le A_{t-1} C_t \le A_{t-1}$, and Hoeffding's lemma, thus $\mathbb{E}_{|t-1} \exp(2vB_t) \le \exp(2v^2 A_{t-1})$; Eq. (12) is due to $\frac{1}{1-u\hat{R}_t} \le \frac{2\hat{R}_{t-1}}{2\hat{R}_{t-1} - \hat{R}_t} \le 2$; Eq. (13) is due to Eqs. (5). Then for $u \in (0, \frac{1}{2\hat{R}_t})$,

$$\mathbb{E} \exp(u \sum_{\tau=1}^{T+1} X_\tau) \le \exp((u\bar{P}_0 + \frac{u^2 \tilde{P}_0^2}{2(1-u\hat{R}_0)})A_0 + u\bar{R}_0 + \frac{u^2 \tilde{R}_0^2}{2(1-u\hat{R}_0)}) \le \exp(u(\bar{P}_0 D^2 + \bar{R}_0) + \frac{u^2(\tilde{P}_0^2 D^2 + \tilde{R}_0^2)}{2(1-2u\hat{R}_0)}).$$

Eq. (6) follows Lemma 8. □

We prove Lemma 6, which is the same as Lemma 7 of (Lan, 2008) except for the strong convexity.

Lemma 6. *Let* $\delta_t = G(x_{t-1}, \xi_t) - g(x_{t-1})$, $A_t = \|x_t - x_*\|^2$, $B_t = \langle \delta_t, x_{t-1} - x_* \rangle / Q$, $C_t = \|\delta_t\|_*^2 / Q^2$. *If* $\gamma_t > 0$ *and* $\gamma_t L < 1$, *it holds for Algorithm 1 that*

$$f(x_t) - f(x_*) \le \frac{1 - \gamma_t \mu}{2\gamma_t} A_{t-1} - \frac{1}{2\gamma_t} A_t - QB_t + \frac{\gamma_t}{2(1-\gamma_t L)} Q^2 C_t.$$

Proof. Let $d_t = x_t - x_{t-1}$.

$$f(x_t) \le f(x_{t-1}) + \langle g(x_{t-1}), d_t \rangle + \frac{L}{2}\|d_t\|^2 \tag{14}$$

$$\le f(x_*) + \langle g(x_{t-1}), x_t - x_* \rangle - \frac{\mu}{2}\|x_{t-1} - x_*\|^2 + \frac{L}{2}\|d_t\|^2 \tag{15}$$

$$= f(x_*) + \langle \hat{g}_t, x_t - x_* \rangle - \frac{\mu}{2}\|x_{t-1} - x_*\|^2 + \frac{L}{2}\|d_t\|^2 - \langle \delta_t, x_t - x_* \rangle$$

$$\le f(x_*) + \frac{1 - \gamma_t \mu}{2\gamma_t}\|x_{t-1} - x_*\|^2 - \frac{1}{2\gamma_t}\|x_t - x_*\|^2 - \frac{1 - \gamma_t L}{2\gamma_t}\|d_t\|^2 - \langle \delta_t, d_t \rangle - \langle \delta_t, x_{t-1} - x_* \rangle \tag{16}$$

$$\le f(x_*) + \frac{1 - \gamma_t \mu}{2\gamma_t}\|x_{t-1} - x_*\|^2 - \frac{1}{2\gamma_t}\|x_t - x_*\|^2 + \frac{\gamma_t}{2(1-\gamma_t L)}\|\delta_t\|_*^2 - \langle \delta_t, x_{t-1} - x_* \rangle . \tag{17}$$

Eq. (14) is due to the Lipschitz continuity of f, Eq. (15) due to the strong convexity of f, Eq. (16) due to the optimality of Step 4. □

Proof of Theorem 1. Because $\gamma_t L = \frac{2\kappa}{t+2\kappa} < 1$, it follows Lemma 6 that

$$f(x_t) - f(x_*) \leq \frac{1-\gamma_t\mu}{2\gamma_t}A_{t-1} - \frac{1}{2\gamma_t}A_t - QB_t + \frac{\gamma_t Q^2}{2(1-\gamma_t L)}$$
$$\leq (t+2\kappa-2)\frac{\mu A_{t-1}}{4} - (t+2\kappa)\frac{\mu A_t}{4} - QB_t + \frac{Q^2}{\mu t}.$$

As $f(x_t) - f(x_*) \geq \frac{\mu}{2}A_t$ it follows Lemma 6 that

$$A_t \leq d_t(a_t A_{t-1} + 2b_t B_t + c_t C_t),$$

where $a_t = \frac{\mu(t+2\kappa-2)}{4}$, $b_t = -\frac{Q}{2}$, $c_t = \frac{Q^2}{\mu t}$ and $d_t = \frac{4}{\mu(t+2\kappa+2)}$. Let $w_t = \alpha_t \prod_{\tau=t+1}^{T}(1-\alpha_\tau) = \frac{2t}{T(T+1)}$. Assume that $\alpha_0 = 0$ and $\gamma_0 = 1$. Then

$$f(\bar{x}_T) - f(x_*) \leq \sum_{t=1}^{T} w_t(f(x_t) - f(x_*)) \leq \sum_{t=1}^{T} w_t\left(\frac{1-\gamma_t\mu}{2\gamma_t}A_{t-1} - \frac{1}{2\gamma_t}A_t - QB_t + \frac{\gamma_t Q^2}{2(1-\gamma_t L)}\right)$$
$$\leq \sum_{t=1}^{T} w_t\left(\frac{1-\gamma_t\mu}{2\gamma_t} - \frac{w_{t-1}}{2w_t\gamma_{t-1}}\right)A_{t-1} - \sum_{t=1}^{T} w_t QB_t + \sum_{t=1}^{T} w_t\frac{\gamma_t Q^2}{2(1-\gamma_t L)}$$
$$\leq \sum_{t=1}^{T} w_t\left(\frac{L}{2t}A_{t-1} - QB_t + \frac{Q^2}{\mu t}\right) \leq \frac{LD^2}{T} + \sum_{t=1}^{T} w_t\left(-QB_t + \frac{Q^2}{\mu t}\right).$$

Note that we use the factor $A_{t-1} \leq D^2$ for simplicity. Let $\tilde{a}_t = 0$, $\tilde{b}_t = b_t$, $\tilde{c}_t = c_t$, $X_{T+1} = \frac{LD^2}{T}$, and

$$\bar{P}_t = 0,$$
$$\bar{R}_t = \frac{LD^2}{T} + \frac{2\kappa Q^2(T-t)}{LT^2},$$
$$\tilde{P}_t^2 = \frac{4Q^2(T-t)(t+2\kappa+2)(t+2\kappa-1)}{T^2(T+1)^2},$$
$$\tilde{R}_t^2 = \frac{Q^4\kappa^2}{L^2T^2(T+1)^2}(8(T-t)(T-t-1) + 32\kappa T\widetilde{\ln}(T,t)),$$
$$\hat{R}_t = \frac{5\kappa Q^2(T-t)}{LT^2}.$$

Given the facts that $\kappa \geq 1$, $(t+2\kappa-2)(t+2\kappa-1) \leq (t+2\kappa+1)(t+2\kappa-2)$, $(T-t+1)-(T-t) = 1$, $(T-t+1)^2 - (T-t)^2 \geq 2(T-t)$, $(T-t+1)^3 - (T-t)^3 \geq 3(T-t)^2$, the proof of Eq. (23) follows from Lemma 5, because for $t \geq 1$,

$$a_t d_t \bar{P}_t + w_t \tilde{a}_t = 0 = \bar{P}_{t-1},$$

$$\bar{R}_t + w_t c_t + c_t d_t \bar{P}_t \leq \bar{R}_t + \frac{2t}{T^2}\frac{Q^2}{\mu t} \leq \bar{R}_{t-1},$$

$$a_t d_t \tilde{P}_t^2 + 4(w_t\tilde{b}_t + b_t d_t \bar{P}_t)^2 \leq \frac{t+2\kappa-2}{t+2\kappa+2}\tilde{P}_t^2 + \frac{4t^2Q^2}{T^2(T+1)^2} \leq \frac{Q^2}{T^2(T+1)^2}(4(T-t)(t+2\kappa+1)(t+2\kappa-2) + 4t^2)$$
$$\leq \tilde{P}_{t-1}^2 - \frac{Q^2}{T^2(T+1)^2}(4(t+2\kappa+1)(t+2\kappa-2) - 4t^2)$$
$$= \tilde{P}_{t-1}^2 - \frac{Q^2}{T^2(T+1)^2}(4(2\kappa-1)t + 16\kappa^2 - 8\kappa - 8) \leq \tilde{P}_{t-1}^2,$$

$$\tilde{R}_t^2 + c_t d_t \tilde{P}_t^2 \leq \tilde{R}_t^2 + \frac{16Q^4(T-t)(t+2\kappa+2)(t+2\kappa-1)}{\mu^2T^2(T+1)^2 t(t+2\kappa+2)}$$
$$\leq \frac{Q^4}{\mu^2T^2(T+1)^2}(8(T-t)(T-t-1) + 32\kappa T\widetilde{\ln}(T,t) + 16(T-t) + \frac{16(2\kappa-1)}{t}) \leq \tilde{R}_{t-1}^2,$$

and

$$\begin{aligned}
a_t d_t \tilde{P}_t^2 \hat{R}_t + 4b_t d_t (w_t \tilde{b}_t + b_t d_t \bar{P}_t)\tilde{P}_t^2 &\leq \frac{4Q^2(T-t)(t+2\kappa-1)(t+2\kappa-2)}{T^4}\hat{R}_t + \frac{32Q^4 t(T-t)(t+2\kappa-1)}{\mu T^6} \\
&\leq \frac{Q^4}{\mu T^6}(20(T-t)^2(t+2\kappa-1)(t+2\kappa-2) + 32t(T-t)(t+2\kappa-1)) \\
&\leq \tilde{P}_{t-1}^2 \hat{R}_{t-1} - \frac{Q^4(T-t)}{\mu T^6}(2 \times 20(t+2\kappa+1)(t+2\kappa-2) - 32t(t+2\kappa-1)) \\
&= \tilde{P}_{t-1}^2 \hat{R}_{t-1} - \frac{Q^4(T-t)}{\mu T^6}(8t^2 - 8t + 16\kappa(6t-5) + 160\kappa^2 - 80) \leq \tilde{P}_{t-1}^2 \hat{R}_{t-1}.
\end{aligned}$$

$$\begin{aligned}
&a_t d_t \tilde{P}_t^2 \hat{R}_t^2 + 4b_t d_t (w_t \tilde{b}_t + b_t d_t \bar{P}_t)\tilde{P}_t^2 \hat{R}_t + 2b_t^2 d_t^2 \tilde{P}_t^4 \\
&\leq \frac{4Q^2(T-t)(t+2\kappa+1)(t+2\kappa-2)}{T^4}\hat{R}_t^2 + \frac{32Q^4 t(T-t)(t+2\kappa-1)}{\mu T^6}\hat{R}_t + \frac{128Q^6(T-t)^2(t+2\kappa-1)^2}{\mu^2 T^8} \\
&\leq \tilde{P}_{t-1}^2 \hat{R}_{t-1}^2 - \frac{Q^6(T-t)^2}{\mu^2 T^8}(3 \times 100(t+2\kappa+1)(t+2\kappa-2) - 160t(t+2\kappa-1) - 128(t+2\kappa-1)^2) \\
&= \tilde{P}_{t-1}^2 \hat{R}_{t-1}^2 - \frac{Q^6(T-t)^2}{\mu^2 T^8}(12(t-1)^2 + 368(t-1)(\kappa-1) + 688(\kappa-1)^2 + 508(t-1) + 1656(\kappa-1) + 368) \\
&\leq \tilde{P}_{t-1}^2 \hat{R}_{t-1}^2.
\end{aligned}$$

□

Proof of Proposition 2. Because $\gamma_t L < 1$, it follows Lemma 6 that

$$\begin{aligned}
f(x_t) - f(x_*) &\leq \frac{1-\gamma_t\mu}{2\gamma_t}A_{t-1} - \frac{1}{2\gamma_t}A_t - QB_t + \frac{\gamma_t Q^2}{2(1-\gamma_t L)} \\
&\leq (L + \mu(2t-1))\frac{A_{t-1}}{2} - (L + 2\mu t)\frac{A_t}{2} - QB_t + \frac{Q^2}{4\mu t}.
\end{aligned}$$

As the strong convexity implies that $f(x_t) - f(x_*) \geq \frac{\mu}{2}A_t$, it follows Lemma 6 that

$$A_t \leq d_t(a_t A_{t-1} + 2b_t B_t + c_t C_t),$$

where $a_t = \frac{\mu(t+\kappa-1)}{2}$, $b_t = -\frac{Q}{2}$, $c_t = \frac{Q^2}{2\mu t}$ and $d_t = \frac{2}{\mu(t+\kappa+1)}$. Let $w_t = \alpha_t \prod_{\tau=t+1}^{T}(1-\alpha_\tau) = \frac{1}{T}$. Assume that $\alpha_0 = 0$ and $\gamma_0 = 1$. Then

$$\begin{aligned}
f(\bar{x}_T) - f(x_*) &\leq \sum_{t=1}^{T} w_t(f(x_t) - f(x_*)) \leq \sum_{t=1}^{T} w_t \left(\frac{1-\gamma_t\mu}{2\gamma_t}A_{t-1} - \frac{1}{2\gamma_t}A_t - QB_t + \frac{\gamma_t Q^2}{2(1-\gamma_t L)}\right) \\
&\leq \sum_{t=1}^{T} w_t \left(\frac{1-\gamma_t\mu}{2\gamma_t} - \frac{w_{t-1}}{2w_t\gamma_{t-1}}\right)A_{t-1} - \sum_{t=1}^{T} w_t Q B_t + \sum_{t=1}^{T} w_t \frac{\gamma_t Q^2}{2(1-\gamma_t L)} \\
&\leq \frac{LA_0}{2T} + \sum_{t=1}^{T} w_t \left(-QB_t + \frac{Q^2}{4\mu t}\right).
\end{aligned}$$

Let $\tilde{a}_t = 0$, $\tilde{b}_t = b_t$, $\tilde{c}_t = c_t$, $X_{T+1} = \frac{LD^2}{2T}$, and

$$\begin{aligned}
\bar{P}_t &= 0, \\
\bar{R}_t &= \frac{Q^2}{2\mu T}\widetilde{\ln}(T,t), \\
\tilde{P}_t^2 &= \frac{Q^2(t+\kappa+1)}{T^2}, \\
\tilde{R}_t^2 &= \frac{Q^4}{\mu^2 T^2}\widetilde{\ln}(T,t), \\
\hat{R}_t &= \frac{3Q^2}{\mu T}.
\end{aligned}$$

The proof follows from Lemma 5, because for $k \geq 1$,

$$\begin{aligned}
&\bar{P}_t d_t a_t + p_t \tilde{a}_t = 0 = \bar{P}_{t-1}, \\
&\bar{R}_t + w_t c_t + \bar{P}_t d_t c_t \leq \frac{Q^2}{2\mu T} \ln \frac{T}{t} + \frac{1}{T} \frac{Q^2}{2\mu t} \leq \bar{R}_{t-1}, \\
&\tilde{P}_t^2 d_t a_t + 4(w_t + \bar{P}_t d_t)^2 b_t^2 \leq \frac{Q^2(\kappa + t + 1)}{T^2} \frac{t + \kappa - 1}{t + \kappa + 1} + \frac{Q^2}{T^2} = \frac{Q^2(t + \kappa)}{T^2(t + \kappa + 1)} = \tilde{P}_{t-1}^2, \\
&\tilde{R}_t^2 + \tilde{P}_t^2 d_t c_t \leq \frac{Q^4}{\mu^2 T^2} \ln \frac{T}{t} + \frac{2Q^2}{\mu T^2} \frac{Q^2}{2\mu t} \leq \tilde{R}_{t-1}^2,
\end{aligned}$$

and

$$\begin{aligned}
&a_t d_t \tilde{P}_t^2 \hat{R}_t + 4 b_t d_t (w_t \tilde{b}_t + b_t d_t \bar{P}_t) \tilde{P}_t^2 \leq \frac{Q^2(t + \kappa - 1)(t + \kappa + 1)}{T^2(t + \kappa + 1)} \hat{R}_t + \frac{2Q^4(t + \kappa + 1)}{\mu T^3 (t + \kappa + 1)} \leq \frac{Q^2(t + \kappa)}{T^2} \hat{R}_{t-1}. \\
&a_t d_t \tilde{P}_t^2 \hat{R}_t^2 + 4 b_t d_t (w_t \tilde{b}_t + b_t d_t \bar{P}_t) \tilde{P}_t^2 \hat{R}_t + 2 b_t^2 d_t^2 \tilde{P}_t^4 \\
&\qquad \leq \frac{Q^2(t + \kappa - 1)(t + \kappa + 1)}{T^2(t + \kappa + 1)} \hat{R}_t^2 + \frac{2Q^4(t + \kappa + 1)}{\mu T^3(t + \kappa + 1)} \hat{R}_t + \frac{2Q^6(t + \kappa + 1)^2}{\mu^2 T^4 (t + \kappa + 1)^2} \leq \frac{Q^2(t + \kappa)}{T^2} \hat{R}_{t-1}^2.
\end{aligned}$$

□

Proof of Proposition 3. Because $\gamma_t L < 1$, it follows Lemma 6 that

$$\begin{aligned}
f(x_t) - f(x_*) &\leq \frac{1 - \gamma_t \mu}{2\gamma_t} A_{t-1} - \frac{1}{2\gamma_t} A_t - Q B_t + \frac{\gamma_t Q^2}{2(1 - \gamma_t L)} \\
&\leq (L + \mu(t-1)) \frac{A_{t-1}}{2} - (L + \mu t) \frac{A_t}{2} - Q B_t + \frac{Q^2}{2\mu t}.
\end{aligned}$$

As the strong convexity implies that $f(x_t) - f(x_*) \geq \frac{\mu}{2} A_t$, it follows Lemma 6 that

$$A_t \leq d_t (a_t A_{t-1} + 2 b_t B_t + c_t C_t),$$

where $a_t = \frac{\mu(t+\kappa-1)}{2}$, $b_t = -\frac{Q}{2}$, $c_t = \frac{Q^2}{2\mu t}$ and $d_t = \frac{2}{\mu(t+\kappa+1)}$. Because the solution is an interior point, we have

$$f(x_T) - f(x_*) \leq \frac{L}{2} A_T.$$

Let $w_t = 0$, $X_{T+1} = \frac{L}{2} A_T$, and

$$\begin{aligned}
\bar{P}_t &= \frac{L(t + \kappa)(t + \kappa + 1)}{2(T + \kappa)(T + \kappa + 1)}, \\
\bar{R}_t &= \frac{\kappa^2 Q^2}{2L(T + \kappa)(T + \kappa + 1)} (T - t + \kappa \widetilde{\ln}(T, t)), \\
\tilde{P}_t^2 &= \frac{Q^2 \kappa^2 (T - t)(t + \kappa)(t + \kappa + 1)}{2(T + \kappa)^2 (T + \kappa + 1)^2}, \\
\tilde{R}_t^2 &= \frac{\kappa^4 Q^4}{4L^2 (T + \kappa)^2 (T + \kappa + 1)^2} ((T - t)(T - t - 1) + \kappa T \widetilde{\ln}(T, t)), \\
\hat{R}_t &= \frac{2\kappa^2 Q^2 (T - t)}{L(T + \kappa)(T + \kappa + 1)}.
\end{aligned}$$

The proof follows from Lemma 5, because

$$
\begin{aligned}
&\bar{P}_t d_t a_t = \frac{L(t+\kappa)(t+\kappa-1)}{2(T+\kappa)(T+\kappa+1)} = \bar{P}_{t-1}, \\
&\bar{R}_t + \bar{P}_t d_t c_t \le \bar{R}_t + \frac{L(t+\kappa)(t+\kappa+1)}{2(T+\kappa)(T+\kappa+1)} \frac{2}{\mu(t+\kappa+1)} \frac{Q^2}{2\mu t} \\
&\qquad \le \frac{\kappa^2 Q^2}{2L(T+\kappa)(T+\kappa+1)}(T - t + \kappa \widetilde{\ln}(T,t) + \frac{t+\kappa}{t}) \le \bar{R}_{t-1}, \\
&\tilde{P}_t^2 d_t a_t + \bar{P}_t^2 d_t^2 b_t^2 \le \frac{t+\kappa-1}{t+\kappa+1}\tilde{P}_t^2 + \frac{\kappa^2 Q^2 (t+\kappa)^2}{4(T+\kappa)^2(T+\kappa+1)^2} \\
&\qquad \le \tilde{P}_{t-1}^2 - \frac{\kappa^2 Q^2}{(T+\kappa)^2(T+\kappa+1)^2}(\frac{1}{2}(t+\kappa-1)(t+\kappa) - \frac{1}{4}(t+\kappa)^2) \\
&\qquad \le \tilde{P}_{t-1}^2 - \frac{\kappa^2 Q^2}{(T+\kappa)^2(T+\kappa+1)^2}(\frac{1}{4}(t+\kappa)(t+\kappa-2)) \le \tilde{P}_{t-1}^2, \qquad [t \ge 1 \text{ and } \kappa \ge 1] \\
&\tilde{R}_t^2 + \tilde{P}_t^2 d_t c_t \le \tilde{R}_t^2 + \frac{Q^4 \kappa^2 (T-t)(t+\kappa)}{2\mu^2 (T+\kappa)^2 (T+\kappa+1)^2 t} \\
&\qquad \le \frac{Q^4 \kappa^2}{4\mu^2 (T+\kappa)^2 (T+\kappa+1)^2}((T-t)(T-t-1) + \kappa T \widetilde{\ln}(T,t) + 2(T-t) + \frac{(T-t)\kappa}{t}) \le \tilde{R}_{t-1}^2,
\end{aligned}
$$

and

$$
\begin{aligned}
&a_t d_t \tilde{P}_t^2 \hat{R}_t + 4 b_t^2 d_t^2 \bar{P}_t \tilde{P}_t^2 \\
&\quad \le \frac{Q^2 \kappa^2}{(T+\kappa)^2 (T+\kappa+1)^2} \left(\frac{1}{2}(T-t)(t+\kappa)(t+\kappa-1)\hat{R}_t + (T-t)(t+\kappa)\frac{LQ^2(t+\kappa)}{\mu^2 (T+\kappa)(T+\kappa+1)} \right) \\
&\quad \le \tilde{P}_{t-1}^2 \hat{R}_{t-1} - \frac{Q^4 \kappa^4}{L(T+\kappa)^3 (T+\kappa+1)^3} \big(2(T-t)(t+\kappa)(t+\kappa-1) - (T-t)(t+\kappa)^2 \big) \\
&\quad \le \tilde{P}_{t-1}^2 \hat{R}_{t-1} - \frac{Q^4 \kappa^4}{L(T+\kappa)^3 (T+\kappa+1)^3 (T-t)}(t+\kappa)(t+\kappa-2) \le \tilde{P}_{t-1}^2 \hat{R}_{t-1}. \\
&a_t d_t \tilde{P}_t^2 \hat{R}_t^2 + 4 b_t^2 d_t^2 \bar{P}_t \tilde{P}_t^2 \hat{R}_t + 2 b_t^2 d_t^2 \tilde{P}_t^4 \\
&\quad \le \frac{Q^2 \kappa^2}{(T+\kappa)^2 (T+\kappa+1)^2}(\frac{1}{2}(T-t)(t+\kappa)(t+\kappa-1)\hat{R}_t^2 + (T-t)(t+\kappa)\frac{LQ^2(t+\kappa)}{\mu^2 (T+\kappa)(T+\kappa+1)}\hat{R}_t \\
&\qquad + \frac{Q^4 \kappa^2 (T-t)^2 (t+\kappa)^2}{4\mu^2 (T+\kappa)^2 (T+\kappa+1)^2}) \\
&\quad \le \tilde{P}_{t-1}^2 \hat{R}_{t-1}^2 - \frac{Q^6 \kappa^6}{L^2 (T+\kappa)^4 (T+\kappa+1)^4}(6(T-t)^2(t+\kappa)(t+\kappa-1) - 2(T-t)^2(t+\kappa) - \frac{1}{4}(T-t)^2(t+\kappa)^2) \\
&\quad \le \tilde{P}_{t-1}^2 \hat{R}_{t-1}^2 - \frac{Q^6 \kappa^6 (T-t)^2 (t+\kappa)}{L^2 (T+\kappa)^4 (T+\kappa+1)^4}(\frac{15}{4}(t+\kappa) - 6) \le \tilde{P}_{t-1}^2 \hat{R}_{t-1}^2.
\end{aligned}
$$

□

Similar to Lemma 9 of (Lan, 2008), we have the following lemma for Algorithm 2 with the consideration of strongly convex cases.

Lemma 7. *Let $\delta_t = G(y_{t-1}, \xi_t) - g(y_{t-1})$, $A_t = \|x_t - x_*\|^2$, $B_t = \langle \delta_t, x_{t-1} - x_* \rangle / Q$, $C_t = \|\delta_t\|_*^2 / Q^2$. If $0 < \alpha_t < 1$, $\gamma_t > 0$ and $\gamma_t(\alpha_t L + \mu) < 1$, it holds for Algorithm 2 that*

$$
f(\bar{x}_t) - f(x_*) \le (1-\alpha_t)(f(\bar{x}_{t-1}) - f(x_*)) + \frac{\alpha_t(1-\gamma_t \mu)}{2\gamma_t} A_{t-1} - \frac{\alpha_t}{2\gamma_t} A_t - \alpha_t Q B_t + \frac{\alpha_t \gamma_t}{2(1 - \alpha_t \gamma_t L - \gamma_t \mu)} Q^2 C_t.
$$

Proof. Let $d_t = x_t - x_{t-1}$ and $v_t = x_{t-1} + \gamma_t \mu (y_{t-1} - x_{t-1})$. Note that $\bar{x}_t - y_{t-1} = \alpha_t d_t$.

$$
\begin{aligned}
f(\bar{x}_t) &\le f(y_{t-1}) + \langle g(y_{t-1}), \bar{x}_t - y_{t-1} \rangle + \frac{L}{2}\|\bar{x}_t - y_{t-1}\|^2 && (18)\\
&= (1-\alpha_t)[f(y_{t-1}) + \langle g(y_{t-1}), \bar{x}_{t-1} - y_{t-1} \rangle] + \alpha_t[f(y_{t-1}) + \langle g(y_{t-1}), x_t - y_{t-1}\rangle] + \frac{\alpha_t^2 L}{2}\|d_t\|^2 \\
&\le (1-\alpha_t) f(\bar{x}_{t-1}) + \alpha_t f(x_*) + \alpha_t \langle g(y_{t-1}), x_t - x_* \rangle - \frac{\alpha_t \mu}{2}\|y_{t-1} - x_*\|^2 + \frac{\alpha_t^2 L}{2}\|d_t\|^2 && (19)\\
&= (1-\alpha_t) f(\bar{x}_{t-1}) + \alpha_t f(x_*) + \alpha_t \langle \hat{g}_t, x_t - x_* \rangle - \frac{\alpha_t \mu}{2}\|y_{t-1} - x_*\|^2 + \frac{\alpha_t^2 L}{2}\|d_t\|^2 - \alpha_t \langle \delta_t, x_t - x_* \rangle \\
&\le (1-\alpha_t) f(\bar{x}_{t-1}) + \alpha_t f(x_*) + \frac{\alpha_t}{\gamma_t} \langle x_t - v_t, x_* - x_t \rangle - \frac{\alpha_t \mu}{2}\|y_{t-1} - x_*\|^2 + \frac{\alpha_t^2 L}{2}\|d_t\|^2 - \alpha_t \langle \delta_t, x_t - x_* \rangle && (20)\\
&= (1-\alpha_t) f(\bar{x}_{t-1}) + \alpha_t f(x_*) + \frac{\alpha_t(1-\gamma_t\mu)}{2\gamma_t}\|x_{t-1} - x_*\|^2 - \frac{\alpha_t}{2\gamma_t}\|x_t - x_*\|^2 - \frac{\alpha_t\mu}{2}\|y_{t-1} - x_t\|^2 \\
&\quad - \frac{\alpha_t(1-\gamma_t\mu - \alpha_t\gamma_t L)}{2\gamma_t}\|d_t\|^2 - \alpha_t \langle \delta_t, d_t \rangle - \alpha_t \langle \delta_t, x_{t-1} - x_* \rangle \\
&\le (1-\alpha_t) f(\bar{x}_{t-1}) + \alpha_t f(x_*) + \frac{\alpha_t(1-\gamma_t\mu)}{2\gamma_t}\|x_{t-1} - x_*\|^2 - \frac{\alpha_t}{2\gamma_t}\|x_t - x_*\|^2 \\
&\quad + \frac{\alpha_t\gamma_t}{2(1-\gamma_t\mu - \alpha_t\gamma_t L)}\|\delta_t\|_*^2 - \alpha_t \langle \delta_t, x_{t-1} - x_* \rangle . && (21)
\end{aligned}
$$

Eq. (24) is due to the Lipschitz continuity of f, Eq. (25) due to the strong convexity of f, Eq. (20) due to the optimality of Step 6. □

Proof of Theorem 4. Let $\lambda_t = \prod_{\tau=t+1}^{T}(1-\alpha_t) = \frac{t(t+1)}{T(T+1)}$. We have and

$$
\frac{\lambda_t\alpha_t(1-\gamma_t\mu)}{\gamma_t} - \frac{\lambda_{t-1}\alpha_{t-1}}{\gamma_{t-1}} = \frac{2t}{T(T+1)}(\frac{2L}{t} + \frac{\mu(t+1)}{2} - \mu) - \frac{2(t-1)}{T(T+1)}(\frac{2L}{t-1} + \frac{\mu t}{2}) = 0, \quad \forall t > 1.
$$

Let $a_t = \frac{\mu(4\kappa + t(t-1))}{2t}$, $b_t = -\frac{Q}{2}$, $c_t = \frac{Q^2}{\mu t}$, and $d_t = \frac{2t}{\mu(4\kappa + t(t+1))}$. Summing up the inequality in Lemma 7 weighted by λ_t, we have

$$
\begin{aligned}
f(\bar{x}_t) - f(x_*) &\le \frac{\lambda_1\alpha_1(1-\gamma_1\mu)}{2\gamma_1}A_0 - \frac{\lambda_t\alpha_t}{2\gamma_t}A_t + \sum_{\tau=1}^{t}\lambda_\tau\alpha_\tau(-QB_t + \frac{\gamma_\tau}{2(1-\alpha_\tau\gamma_\tau L - \gamma_\tau\mu)}Q^2C_\tau) \\
&\le \frac{2L}{T(T+1)}A_0 - \frac{2t}{T(T+1)}\frac{A_t}{d_t} + \sum_{\tau=1}^{t}\frac{2\tau}{T(T+1)}\left(2b_\tau B_\tau + c_\tau C_\tau\right).
\end{aligned} \tag{22}
$$

Let $\tilde{A}_t := \frac{d_t}{t}\left\{LA_0 + \sum_{\tau=1}^{t}\left(2\tau b_\tau B_t + \tau c_t C_t\right)\right\}$. Because $f(\bar{x}_t) - f(x_*) \ge 0$, we have

$$
\frac{t}{d_t}A_t \le \frac{t}{d_t}\tilde{A}_t = \frac{t-1}{d_{t-1}}\tilde{A}_{t-1} + 2tb_tB_t + tc_tC_t = ta_t\tilde{A}_{t-1} + 2tb_tB_t + tc_tC_t
$$

Then

$$
A_t \le \tilde{A}_t = d_t(a_t\tilde{A}_{t-1} + 2b_tB_t + c_tC_t). \tag{23}
$$

Given Eq. (22) and Eq. (23), letting $w_t = \frac{2t}{T(T+1)}$, $\tilde{a}_t = 0$, $\tilde{b}_t = b_t$, $\tilde{c}_t = c_t$, $X_{T+1} = \frac{2LD^2}{T(T+1)}$, and

$$\begin{aligned}
\bar{P}_t &= 0,\\
\bar{R}_t &= \frac{2LD^2}{T^2} + \frac{2\kappa Q^2(T-t)}{LT^2},\\
\tilde{P}_t^2 &= \frac{5Q^2(T-t)(t(t+1)+4\kappa)}{T^4},\\
\tilde{R}_t^2 &= \frac{5\kappa^2 Q^4(T-t)(T-t-1)}{2L^2T^4},\\
\hat{R}_t &= \frac{4\kappa Q^2(T-t)}{LT^2},
\end{aligned}$$

the proof follows from Lemma 5, because

$$\begin{aligned}
&a_t d_t \bar{P}_t + w_t \tilde{a}_t = 0 = \bar{P}_{t-1},\\
&\bar{R}_t + w_t \tilde{c}_t + c_t d_t \bar{P}_t \leq \bar{R}_t + \frac{2t}{T^2}\frac{Q^2}{\mu t} \leq \bar{R}_{t-1},\\
&a_t d_t \tilde{P}_t^2 + 4(w_t \tilde{b}_t + b_t d_t \bar{P}_t)^2 \leq \frac{t(t-1)+4\kappa}{t(t+1)+4\kappa}\tilde{P}_t^2 + \frac{4t^2Q^2}{T^4}\\
&\quad \leq \frac{Q^2}{T^4}(6(t(t-1)+4\kappa)(T-t)+4t^2) \leq \tilde{P}_{t-1}^2 - \frac{Q^2}{T^4}(5(t(t-1)+4\kappa)-4t^2)\\
&\quad \leq \tilde{P}_{t-1}^2 - \frac{Q^2}{T^4}(t^2-5t+20\kappa) \leq \tilde{P}_{t-1}^2 - \frac{Q^2}{T^4}(3t) \leq \tilde{P}_{t-1}^2,\\
&\tilde{R}_t^2 + c_t d_t \tilde{P}_t^2 \leq \tilde{R}_t^2 + \frac{5Q^4(T-t)}{\mu^2 T^4} \leq \tilde{R}_{t-1}^2,
\end{aligned}$$

and

$$\begin{aligned}
&a_t d_t \tilde{P}_t^2 \hat{R}_t + 4 b_t d_t (w_t \tilde{b}_t + b_t d_t \bar{P}_t)\tilde{P}_t^2 \leq \frac{Q^2(t(t-1)+4\kappa)(T-t)}{T^4}\hat{R}_t + \frac{4t^2Q^4(T-t)}{\mu T^6}\\
&\quad \leq \frac{Q^4}{\mu T^6}(4(t(t-1)+4\kappa)(T-t)^2 + 4t^2(T-t))\\
&\quad \leq \tilde{P}_{t-1}^2\hat{R}_{t-1} - \frac{Q^4(T-t)}{\mu T^6}(2\times 4(t(t-1)+4\kappa) - 4t^2)\\
&\quad = \tilde{P}_{t-1}^2\hat{R}_{t-1} - \frac{Q^4(T-t)}{\mu T^6}(4t^2-8t+32\kappa) \leq \tilde{P}_{t-1}^2\hat{R}_{t-1} - \frac{Q^4(T-t)}{\mu T^6}(14t) \leq \tilde{P}_{t-1}^2\hat{R}_{t-1}\\
&a_t d_t \tilde{P}_t^2 \hat{R}_t^2 + 4 b_t d_t (w_t \tilde{b}_t + b_t d_t \bar{P}_t)\tilde{P}_t^2 \hat{R}_t + 2 b_t^2 d_t^2 \tilde{P}_t^4\\
&\quad \leq \frac{5Q^2(t(t-1)+4\kappa)(T-t)}{T^4}\hat{R}_t^2 + \frac{20t^2Q^4(T-t)}{\mu T^6}\hat{R}_t + \frac{100t^2Q^6(T-t)^2}{\mu^2 T^8}\\
&\quad \leq \frac{Q^6}{\mu^2 T^8}(80(t(t-1)+4\kappa)(T-t)^3 + 80t^2(T-t)^2 + 100t^2(T-t)^2)\\
&\quad \leq \tilde{P}_{t-1}^2\hat{R}_{t-1}^2 - \frac{Q^6(T-t)^2}{\mu^2 T^8}(3\times 80(t(t-1)+4\kappa) - 80t^2 - 100t^2)\\
&\quad = \tilde{P}_{t-1}^2\hat{R}_{t-1}^2 - \frac{Q^6(T-t)^2}{\mu^2 T^8}(60t^2-240t+960\kappa) \leq \tilde{P}_{t-1}^2\hat{R}_{t-1}^2 - \frac{(T-t)^2Q^6}{\mu^2 T^8}(240t) \leq \tilde{P}_{t-1}^2\hat{R}_{t-1}^2.
\end{aligned}$$

□

Supporting lemma

We use part of the proof of Lemma 8 in (Birgé & Massart, 1998).

Lemma 8. *Let $B > 0$ and $\sigma > 0$. If the log-moment generating function satisfies*

$$\log \mathbb{E} \exp\{uZ\} \leq \frac{\sigma^2 u^2}{2(1-uB)} \quad \textit{for all } 0 \leq u < 1/B,$$

then

$$\Pr\{Z \geq \epsilon\} \leq \exp\{-\frac{\epsilon^2}{2\sigma^2 + 2\epsilon B}\} \quad \textit{for all } \epsilon \geq 0, \tag{24}$$

and

$$\Pr\{Z \geq \sqrt{2\theta\sigma^2} + \theta B\} \leq \exp\{-\theta\} \quad \textit{for all } \theta \geq 0. \tag{25}$$

Proof. It follows Markov's inequality that

$$\Pr\{Z \geq \epsilon\} \leq \inf_u \mathbb{E} \exp\{-u\epsilon + uZ\} = \exp\{-h(\epsilon)\},$$

where $h(\epsilon) := \sup_u u\epsilon - \frac{\sigma^2 u^2}{2(1-uB)}$. Also, the supremum is achieved for

$$\epsilon = \frac{\sigma^2 u}{1-uB} + \frac{\sigma^2 u^2 B}{2(1-uB)^2} = \frac{\sigma^2 u}{2(1-uB)} + \frac{\sigma^2 u}{2(1-uB)^2},$$

i.e. $u = B^{-1}[1 - \sigma(2\epsilon B + \sigma^2)^{-1/2}] < 1/B$. Then we prove Eq. (24), as

$$h(\epsilon) = \frac{\epsilon^2}{\epsilon B + \sigma^2 + \sigma^2(1 + 2\epsilon B/\sigma^2)^{1/2}} \geq \frac{\epsilon^2}{2\epsilon B + 2\sigma^2}.$$

Let

$$\theta := \frac{\sigma^2 u^2}{2(1-uB)^2} = h(\epsilon).$$

Then we prove Eq. (25), as

$$\sqrt{2\theta\sigma^2} + \theta B = \frac{\sigma^2 u}{(1-uB)} + \frac{\sigma^2 u^2 B}{2(1-uB)^2} = \epsilon.$$

□

References

Birgé, L., & Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, *4*, 329–375.

Ghadimi, S., & Lan, G. (2012). Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: a generic algorithmic framework. *Optimization-online.*

Hu, C., Kwok, J. T., & Pan, W. (2009). Accelerated gradient methods for stochastic optimization and online learning. *NIPS'09: Neural Information Processing Systems.*

Lan, G. (2008). Efficient methods for stochastic composite optimization. *SIAM Journal on Optimization.*

Rakhlin, A., Shamir, O., & Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. *ICML 2012.*

Rudelson, M., & Vershynin, R. (2009). Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, *62*, 1707–1739.

Smale, S., & Zhou, D.-X. (2003). Estimating the approximation error in learning theory. *Anal. Appl. (Singap.)*, *1*, 17–41.