Non-accelerated solvers: \(\kappa_s = L_s/\mu\) for \(\mu\)-strongly convex objectives (where \(L_s = \hat{L} + \|A\|_2^2/\zeta_s\) is the Lipschitz constant of \(f_{\gamma_s}\)). When the solver is stochastic, we use \(\kappa_s = L_{m,s}/\mu\), where \(L_{m,s} = \max_i \tilde{L}_i + \|A\|_2^2/\zeta_s\) (Schmidt, Roux, and Bach 2013).

**Proof of Lemma 1**

**Lemma 1.** For both non-accelerated solvers and accelerated solvers, if \(T_1\) is large enough such that \(\rho_1 \leq \hat{\rho}\), where \(\hat{\rho} \in (0, 1)\), then \(\rho_s \leq \hat{\rho}\) for all \(s > 1\).

**Proof.** To prove this result, we use induction.

- **Non-accelerated solvers:**
  Base step: Since we assume that \(T_1\) is large enough such that \(\rho_1 \leq \hat{\rho}\), then property holds for \(s = 1\).
  Inductive step: Assume that \(\rho_s \leq \hat{\rho}\). Consider \(\rho_{s+1}\). By the definition of \(\kappa_s\) and \(\gamma_{s+1} = \gamma_s/\tau\), we have \(\kappa_{s+1} \leq \tau \kappa_s\). Recall that the non-accelerated solvers take \(T_s = a \kappa_s \phi(\rho_s) + b \phi(\rho_s) + c\) iterations to achieve error reduction factor \(\rho_s\) at stage \(s\).

  With the assumptions on \(\phi(\rho_s)\) and \(a, b, c\), we have \(\phi(\rho_s) \geq \phi(\hat{\rho})\) and \(T_s \geq a \kappa_s \phi(\hat{\rho}) + b \phi(\hat{\rho}) + c\).

  By \(T_{s+1} = \tau T_s\), we have
  \[
  T_{s+1} = a \kappa_{s+1} \phi(\rho_{s+1}) + b \phi(\rho_{s+1}) + c \\
  \geq a \kappa_s \phi(\hat{\rho}) + b \tau \phi(\hat{\rho}) + c \tau \\
  \geq a \kappa_{s+1} \phi(\hat{\rho}) + b \tau \phi(\hat{\rho}) + c \tau \\
  \geq a \kappa_{s+1} \phi(\hat{\rho}) + b \phi(\hat{\rho}) + c
  \]

  Hence, we have \(\rho_{s+1} \leq \hat{\rho}\).

- **Accelerated solvers:** The first part of the proof is identical to that for non-accelerated solvers. By \(T_{s+1} = \sqrt{\tau} T_s\), we have
  \[
  T_{s+1} = a \sqrt{\tau} \kappa_s \phi(\rho_{s+1}) + b \sqrt{\tau} \phi(\rho_{s+1}) + c \sqrt{\tau} \\
  \geq a \sqrt{\tau} \kappa_{s+1} \phi(\hat{\rho}) + b \sqrt{\tau} \phi(\hat{\rho}) + c \sqrt{\tau} \\
  \geq a \sqrt{\tau} \kappa_{s+1} \phi(\hat{\rho}) + b \phi(\hat{\rho}) + c
  \]

  Hence, we have \(\rho_{s+1} \leq \hat{\rho}\).

**Proof of Lemma 2**

**Lemma 2.** \(\hat{P}_s(x) - \hat{P}_s(x^*_s) - \gamma_s D_u \leq P(x) - P(x^*) \leq \hat{P}_s(x) - \hat{P}_s(x^*_s) + \gamma_s D_u\).

**Proof.** Following immediately from (2.7) of (Nesterov 2005), we have \(P(x) - P(x^*) \leq \hat{P}_s(x) - \hat{P}_s(x^*_s) + \gamma_s D_u\) for any \(x \in \mathbb{R}^d\).

From (Nesterov 2005), \(P(x) \leq \hat{P}_s(x) + \gamma_s D_u\). Thus, \(-\hat{P}_s(x^*_s) \leq -P(x^*_s) + \gamma_s D_u \leq -P(x^*) + \gamma_s D_u\). Combining with the fact in (Nesterov 2005) that \(\hat{P}_s(x) \leq P(x)\), we have \(P(x) - \hat{P}_s(x^*_s) \leq P(x) - P(x^*) + \gamma_s D_u\).

Thus, we have \(\hat{P}_s(x) - \hat{P}_s(x^*_s) - \gamma_s D_u \leq P(x) - P(x^*) \leq \hat{P}_s(x) - \hat{P}_s(x^*_s) + \gamma_s D_u\).
Proof of Theorem 1

Lemma 3. If $\gamma_s$ is monotonically decreasing with $s$, then for any $s \geq 2$ and $x \in \mathbb{R}^d$, 
$$\tilde{P}_s(x) - \tilde{P}_s(x^*_s) \leq \tilde{P}_{s-1}(x) - \tilde{P}_{s-1}(x^*_s) + (\gamma_{s-1} - \gamma_s)D_u.$$ 

Proof. From Lemma 9 in (Ouyang and Gray 2012), we have $\tilde{P}_{s-1}(x) \leq \tilde{P}_s(x) \leq \tilde{P}_{s-1}(x) + (\gamma_{s-1} - \gamma_s)D_u$. Result follows by combining the two parts of the inequality.

Proof. (of Theorem 1) With $\gamma_s = \frac{2}{s \tau^s}$, we have

$$
\mathbb{E}P(\tilde{x}_s) - P(x^*) 
\leq \mathbb{E}P_s(\tilde{x}_s) - \tilde{P}_s(x^*_s) + \gamma_s D_u \quad \text{(by Lemma 2)} 
\leq \rho_s(\mathbb{E}P_s(\tilde{x}_{s-1}) - \tilde{P}_s(x^*_s)) + \gamma_s D_u \quad \text{(by Assumption 2)} 
\leq \rho_s(\mathbb{E}P_{s-1}(\tilde{x}_{s-1}) - \tilde{P}_{s-1}(x^*_{s-1}) + (\gamma_{s-1} - \gamma_s)D_u) + \gamma_s D_u \quad \text{(by Lemma 3)} 
= \rho_s(\mathbb{E}P_{s-1}(\tilde{x}_{s-1}) - \tilde{P}_{s-1}(x^*_s)) + \rho_s(\gamma_{s-1} - \gamma_s)D_u + \gamma_s D_u 
\leq \rho_s\rho_{s-1}(\mathbb{E}P_{s-1}(\tilde{x}_{s-2}) - \tilde{P}_{s-1}(x^*_{s-1})) + \rho_s(\gamma_{s-1} - \gamma_s)D_u + \gamma_s D_u \quad \text{(by Assumption 2)}
$$

where in the second-to-last inequality, we apply Lemma 3 and Assumption 2 recursively the same as second and third inequality, and use $\gamma_{s-1} - \gamma_s = \frac{2}{s \tau^s} - \frac{2}{(s-1) \tau^{s-1}}$. In the last inequality, we use Lemma 2. Moreover, note that $\{\beta_s\}$ is monotonically decreasing as follows.

$$
\beta_s - \beta_{s-1} = \left(\sum_{i=1}^{s-1} \frac{s-1}{\tau^i} \prod_{j=1}^{s-1} \rho_j + \frac{1}{\tau^{s-1}} + \prod_{i=1}^{s} \rho_i\right) - \left(\sum_{i=1}^{s-2} \frac{s-1}{\tau^i} \prod_{j=1}^{s-1} \rho_j + \frac{1}{\tau^{s-2}} + \prod_{i=1}^{s} \rho_i\right)
\leq \left(\sum_{i=1}^{s-2} \frac{s-1}{\tau^i} \prod_{j=1}^{s-1} \rho_j + \frac{1}{\tau^{s-1}} + \prod_{i=1}^{s} \rho_i\right)(\rho_s - 1)
\langle 0.
$$

Hence, from (1), $\mathbb{E}P(\tilde{x}_s) - P(x^*)$ converges to zero. We now find out how fast $\{\beta_s\}$ decays. Let $\hat{\rho} = \frac{1}{\tau^s}$, we obtain

$$
\beta_s = \sum_{i=1}^{s-1} \frac{s-1}{\tau^i} \prod_{j=1}^{s-1} \rho_j + \frac{1}{\tau^{s-1}} + \prod_{i=1}^{s} \rho_i
\leq \sum_{i=1}^{s-1} \frac{s-1}{\tau^i} \hat{\rho}^{s-i} + \frac{1}{\tau^{s-1}} + \hat{\rho}^s \quad \text{(by Lemma 1)}
= \sum_{i=1}^{s-1} \frac{s-1}{\tau^{2s-i}} + \frac{1}{\tau^{s-1}} + \frac{1}{\tau^{2s}}
= \frac{1}{\tau^s} - \frac{1}{\tau^{2s-i}} + \frac{1}{\tau^{s-1}} + \frac{1}{\tau^{2s}}
\leq \frac{1 + \tau}{\tau^s},
$$

and

$$
T = \sum_{i=1}^{s} T_i = T_1 \sum_{i=1}^{s} \tau^{i-1} = \frac{\tau^s - 1}{\tau - 1} T_1.
$$
These imply $s = O \left( \log(T) \right)$ and $\beta_s = O \left( \frac{1}{T} \right)$. From (1), we obtain
\[ \mathbb{E} P(\hat{x}_s) - P(x^*) \leq \left( \prod_{i=1}^{s} \rho_i \right) (P(\hat{x}_0) - P(x^*)) + O \left( \frac{\gamma_1 D_u}{T} \right). \]
\[ \square \]

**Proof of Theorem 2**

*Proof.* The first part of the proof is identical to that for Theorem 1. Here, as $T_s = \sqrt{T}_{s-1}$, we have
\[ T = \sum_{i=1}^{s} T_i = T_1 \sum_{i=1}^{s} \sqrt{\tau}^{i-1} = \frac{\sqrt{T} - 1}{\sqrt{\tau} - 1} T_1. \] (5)

Hence, $s = O \left( \log(T) \right)$, $\beta_s = O \left( \frac{1}{T^2} \right)$, and (1) yields
\[ \mathbb{E} P(\hat{x}_s) - P(x^*) \leq \left( \prod_{i=1}^{s} \rho_i \right) (P(\hat{x}_0) - P(x^*)) + O \left( \frac{\gamma_1 D_u}{T^2} \right). \]
\[ \square \]

**Proposition 6.** If we require $\rho_1 \leq 1/\tau$, the rate will be slowed to $O(\log(T)/T)$; if $\rho_1 \leq 1/\sqrt{\tau}$, it degrades further to $O(1/\sqrt{T})$. On the other hand, if $\rho_1 \leq 1/\tau^c$ with $c > 2$, the rate remains at $O(1/T)$.

*Proof.* Following (2) and (4),
- if $\hat{\rho} = \frac{1}{\tau}$, then it leads to $\beta_s \leq \frac{s(2\tau-1)+1}{\tau^2} = O(\log(T)/T)$.
- if $\hat{\rho} = \frac{1}{\sqrt{\tau}}$, then $\beta_s \leq \frac{2\tau+1}{\sqrt{\tau}} = O(1/\sqrt{T})$.
- if $\hat{\rho} = \frac{1}{\tau^c}$ with $c > 2$, then $\beta_s \leq \frac{c+1}{\tau^c} = O(1/T)$.
\[ \square \]

**Proposition 7.** If we require $\rho_1 \leq 1/\tau$, the rate will be slowed to $O(\log(T)/T^2)$; if $\rho_1 \leq 1/(\sqrt{T})$, it degrades further to $O(1/T^2)$. On the other hand, if $\rho_1 \leq 1/(\tau^c)$ with $c > 2$, the rate remains at $O(1/T^2)$.

*Proof.* Following (2) and (5),
- if $\hat{\rho} = \frac{1}{\tau}$, then it leads to $\beta_s \leq \frac{s(2\tau-1)+1}{\tau^2} = O(\log(T)/T^2)$.
- if $\hat{\rho} = \frac{1}{\sqrt{\tau}}$, then $\beta_s \leq \frac{2\tau+1}{\sqrt{\tau}} = O(1/T)$.
- if $\hat{\rho} = \frac{1}{\tau^c}$ with $c > 2$, then $\beta_s \leq \frac{c+1}{\tau^c} = O(1/T^2)$.
\[ \square \]

**Proof of Theorem 3**

In this section, $x^*_s$ denotes the optimal solution to $H_s(x)$.

Note that there are two cases regarding condition number $\kappa_s$. If $\frac{\lambda_2}{2\tau} \|x\|^2$ is added to $\tilde{f}_{x^*_s}$, $\kappa_s = (L_s + \lambda_s)/\lambda_s$ for batch solvers and $\kappa_s = (L_{m,s} + \lambda_s)/\lambda_s$ for stochastic solvers, or if $\frac{\lambda_2}{2\tau} \|x\|^2$ is added to $\tau$, $\kappa_s = L_s/\lambda_s$ for batch solvers and $\kappa_s = L_{m,s}/\lambda_s$ for stochastic solvers.

**Lemma 4.** For any $x \in \mathbb{R}^d$,
\[ P(x) - P(x^*) \leq H_s(x) - H_s(x^*_s) + \gamma_s D_u + \frac{\lambda_s}{2} \|x^*\|^2, \]

*Proof.* As $\hat{P}_s(x) \leq P(x) \leq \hat{P}_s(x) + \gamma_s D_u$ by (2.7) of (Nesterov 2005), we have $P(x) \leq H_s(x) + \gamma_s D_u$, and also $H_s(x^*_s) = \hat{P}_s(x^*_s) + \frac{\lambda_2}{2\tau} \|x^*\|^2 \leq \min_x P(x) + \frac{\lambda_2}{2\tau} \|x\|^2 \leq P(x^*) + \frac{\lambda_2}{2\tau} \|x^*\|^2$. Result follows on combining the two inequalities.
\[ \square \]

**Lemma 5.** For any $x \in \mathbb{R}^d$, $H_s(x) - H_s(x^*_s) \leq P(x) - P(x^*) + \gamma_s D_u + \frac{\lambda_s}{2} \|x\|^2$.

*Proof.* Since $\hat{P}_s(x) \leq P(x)$, we have $H_s(x) \leq P(x) + \frac{\lambda_2}{2\tau} \|x\|^2$. Moreover, since $P(x) \leq H_s(x) + \gamma_s D_u$, and so $P(x^*) \leq H_s(x^*_s) + \gamma_s D_u$. Result follows on combining the two inequalities.
\[ \square \]
Lemma 6. If $\gamma_s$ and $\lambda_s$ are monotonically decreasing with $s$, then for any $s \geq 2$ and $x \in \mathbb{R}^d$,

$$H_s(x) - H_s(x^*_s) \leq H_{s-1}(x) - H_{s-1}(x^*_{s-1}) + (\gamma_{s-1} - \gamma_s)D_u + \frac{1}{2}(\lambda_{s-1} - \lambda_s)||x^*_s||^2,$$

Proof. From Lemma 9 in (Ouyang and Gray 2012), we have $\tilde{P}_{s-1}(x) \leq \tilde{P}_s(x) \leq \tilde{P}_{s-1}(x) + (\gamma_{s-1} - \gamma_s)D_u$. Since $\lambda_{s-1} > \lambda_s$, then

$$H_s(x) \leq H_{s-1}(x) + (\gamma_{s-1} - \gamma_s)D_u.$$

Moreover, $\tilde{P}_{s-1}(x) \leq \tilde{P}_s(x)$ implies $H_{s-1}(x) + \frac{1}{2}(\lambda_s - \lambda_{s-1})||x||^2 \leq H_s(x)$. Thus,

$$H_{s-1}(x^*_{s-1}) \leq H_s(x^*_s) + \frac{1}{2}(\lambda_{s-1} - \lambda_s)||x^*_s||^2.$$

Result follows on combining the two inequalities. \qed

Lemma 7. For both non-accelerated solvers and accelerated solvers, if $T_1$ is large enough such that $\rho_1 \leq \tilde{\rho}$, where $\tilde{\rho} \in (0, 1)$, then $\rho_s \leq \tilde{\rho}$ for all $s > 1$.

Proof. The proof is similar to the one of Lemma 1. We consider induction.

- Non-accelerated solvers:
  Base step: Since we assume that $T_1$ is large enough such that $\rho_1 \leq \tilde{\rho}$, then property holds for $s = 1$.
  Inductive step: Assume that $\rho_s \leq \tilde{\rho}$. Consider $\rho_{s+1}$. By the definition of $\kappa_s$, $\gamma_{s+1} = \gamma_s/\tau$ and $\lambda_{s+1} = \lambda_s/\tau$, we have $\kappa_{s+1} \leq \tau^2\kappa_s$. Recall that the non-accelerated solvers take $T_s = a\kappa_s\phi(\rho_s) + b\phi(\rho_s) + c$ iterations to achieve error reduction factor $\rho_s$ at stage $s$. With the assumptions on $\phi(\rho_s)$ and $a, b, c$, we have $\phi(\rho_s) \geq \phi(\tilde{\rho})$ and $T_s \geq a\kappa_s\phi(\tilde{\rho}) + b\phi(\tilde{\rho}) + c$. By $T_{s+1} = \tau^2T_s$, we have

$$T_{s+1} = a\kappa_{s+1}\phi(\rho_{s+1}) + b\phi(\rho_{s+1}) + c \geq a\tau^2\kappa_s\phi(\rho_s) + b\tau^2\phi(\rho_s) + c\tau^2 \geq a\kappa_{s+1}\phi(\tilde{\rho}) + b\tau^2\phi(\tilde{\rho}) + c\tau^2 \geq a\kappa_{s+1}\phi(\tilde{\rho}) + b\phi(\tilde{\rho}) + c$$

Hence, we have $\rho_{s+1} \leq \tilde{\rho}$.

- Accelerated solvers: The first part of the proof is identical to that for non-accelerated solvers. By $T_{s+1} = \tau T_s$, we have

$$T_{s+1} = a\sqrt{\kappa_{s+1}}\phi(\rho_{s+1}) + b\phi(\rho_{s+1}) + c \geq a\sqrt{\tau^2\kappa_s}\phi(\rho_s) + b\phi(\rho_s) + c \tau \geq a\sqrt{\kappa_{s+1}}\phi(\tilde{\rho}) + b\phi(\tilde{\rho}) + c \tau \geq a\sqrt{\kappa_{s+1}}\phi(\tilde{\rho}) + b\phi(\tilde{\rho}) + c$$

Hence, we have $\rho_{s+1} \leq \tilde{\rho}$. \qed
Proof. (of Theorem 3) With $\gamma_s = \frac{2}{\tau + 1}, \lambda_s = \frac{\lambda_1}{\tau + 1}$, we have
\[
\mathbb{E} P(\tilde{x}_s) - P(x^*)
\leq \mathbb{E} H_s(\tilde{x}_s) - H_s(x_s^*) + \gamma_s D_u + \frac{\lambda_s}{2} \|x^\ast\|^2_2
\] (by Lemma 4)
\leq \rho_s (\mathbb{E} H_s(\tilde{x}_{s-1}) - H_s(x_{s-1}^*)) + \gamma_s D_u + \frac{\lambda_s}{2} \|x^\ast\|^2_2
\] (by Assumption 3)
\leq \rho_s \left( \mathbb{E} H_{s-1}(\tilde{x}_{s-1}) - H_{s-1}(x_{s-1}^*) + (\gamma_{s-1} - \gamma_s) D_u + (\lambda_{s-1} - \lambda_s) \frac{1}{2} \|x^\ast\|^2_2 \right)
+ \gamma_s D_u + \frac{\lambda_s}{2} \|x^\ast\|^2_2
\] (by Lemma 6)
\leq \rho_s \rho_{s-1} \left( \mathbb{E} H_{s-2}(\tilde{x}_{s-2}) - H_{s-2}(x_{s-2}^*) \right) + \rho_s (\gamma_{s-1} - \gamma_s) D_u + \gamma_s D_u
+ \rho_s (\lambda_{s-1} - \lambda_s) \frac{1}{2} \|x^\ast\|^2_2 + \frac{\lambda_s}{2} \|x^\ast\|^2_2
\] (by Assumption 3)
\leq \left( \prod_{i=1}^s \rho_i \right) \left( H_1(\tilde{x}_0) - H_1(x_0^*) \right) + \left( \sum_{i=1}^{s-1} \frac{\tau - 1}{\tau^i} \prod_{j=i+1}^s \rho_j + \frac{1}{\tau^{s-1}} \right) \gamma_1 D_u
+ \left( \sum_{i=1}^{s-1} \frac{\tau - 1}{\tau^i} \prod_{j=i+1}^s \rho_j + \frac{1}{\tau^{s-1}} \right) \lambda_1 R^2
\] (by Lemma 6 and Assumption 3 recursively the same as second and third inequality, and use $\gamma_{s-1} - \gamma_s = \frac{\tau - 1}{\tau^1} \gamma_1$ and $\lambda_{s-1} - \lambda_s = \frac{\tau - 1}{\tau^1} \lambda_1$, and apply assumption $\|x^\ast\|_2 \leq R$ and $\|x_s^\ast\|_s \leq R$ for all $s$. In the last inequality, we use Lemma 5. By the proof of Theorem 1 and Lemma 7 with $\bar{\rho} = \frac{1}{\tau^2}$, we have $\beta_s, \alpha_s \leq \frac{1}{\tau^3}$. And
\[
= \left( \prod_{i=1}^s \rho_i \right) \left( P(\tilde{x}_0) - P(x^*) \right) + \lambda_1 \frac{1}{2} \|\tilde{x}_0\|^2_2
+ \left( \sum_{i=1}^{s-1} \frac{\tau - 1}{\tau^i} \prod_{j=i+1}^s \rho_j + \frac{1}{\tau^{s-1}} \right) \gamma_1 D_u
+ \left( \sum_{i=1}^{s-1} \frac{\tau - 1}{\tau^i} \prod_{j=i+1}^s \rho_j + \frac{1}{\tau^{s-1}} \right) \lambda_1 R^2,
\] (6)

where in the second-to-last inequality, we apply Lemma 6 and Assumption 3 recursively the same as second and third inequality, and use $\gamma_{s-1} - \gamma_s = \frac{\tau - 1}{\tau^1} \gamma_1$ and $\lambda_{s-1} - \lambda_s = \frac{\tau - 1}{\tau^1} \lambda_1$, and apply assumption $\|x^\ast\|_2 \leq R$ and $\|x_s^\ast\|_s \leq R$ for all $s$. In the last inequality, we use Lemma 5. By the proof of Theorem 1 and Lemma 7 with $\bar{\rho} = \frac{1}{\tau^2}$, we have $\beta_s, \alpha_s \leq \frac{1}{\tau^3}$. And
\[
T = \sum_{i=1}^s T_i = T_1 \sum_{i=1}^s \tau^{2(i-1)} = \frac{2^s - 1}{2^{2i-1}} T_1,
\] (7)

which implies that $s = O \left( \log(T) \right)$ and $\beta_s, \alpha_s = O \left( \frac{1}{\sqrt{T}} \right)$. Then, we obtain
\[
\mathbb{E} P(\tilde{x}_s) - P(x^*) \leq \left( \prod_{i=1}^s \rho_i \right) \left( P(\tilde{x}_0) - P(x^*) \right) + \lambda_1 \frac{1}{2} \|\tilde{x}_0\|^2_2
+ O \left( \frac{\lambda_1 R^2}{\sqrt{T}} \right) + O \left( \frac{\gamma_1 D_u}{\sqrt{T}} \right).
\]

For the convergence rate of accelerated solvers, the first part of the proof is identical to that for non-accelerated solvers. Here,
as \( T_s = \tau T_{s-1} \), we have
\[
T = \sum_{i=1}^{s} T_i = T_1 \sum_{i=1}^{s} \tau^{i-1} = \frac{\tau^s - 1}{\tau - 1} T_1
\]
(8)
Hence, \( s = O(\log(T)) \), \( \beta_s, \alpha_s = O\left(\frac{1}{s}\right) \), and (6) yields
\[
\mathbb{E} P(\tilde{x}_s) - P(x^*) \leq \left( \prod_{i=1}^{s} \rho_i \right) \left( P(\tilde{x}_0) - P(x^*) + \frac{\lambda_1}{T} \|x_0\|^2 \right) + O\left( \frac{\lambda_1 R^2}{T} \right) + O\left( \frac{21 D_u}{T} \right).
\]
\[\square\]

**Proposition 8.** For non-accelerated solvers, if we require \( \rho_1 \leq 1/\tau \), the rate will be slowed to \( O(\log(T)/\sqrt{T}) \); if \( \rho_1 \leq 1/\sqrt{\tau} \), it degrades further to \( O(1/T^{1/4}) \). On the other hand, if \( \rho_1 \leq 1/\tau^c \) with \( c > 2 \), the rate remains at \( O(1/\sqrt{T}) \).

For accelerated solvers, if we require \( \rho_1 \leq 1/\tau \), the rate will be slowed to \( O(\log(T)/T) \); if \( \rho_1 \leq 1/\sqrt{\tau} \), it degrades further to \( O(1/\sqrt{T}) \). On the other hand, if \( \rho_1 \leq 1/\tau^c \) with \( c > 2 \), the rate remains at \( O(1/T) \).

**Proof.** Following the proof of Proposition 6 and 7 with (7) and (8). \[\square\]

**Convergence Factors of Example Algorithms**

- Proximal Gradient descent (Nesterov 2013): \( O(\kappa_s \phi(\rho_s)) = 4\kappa_s \log(1/\rho_s) \)
- Accelerated Proximal Gradient descent (Nesterov 2004; Schmidt, Roux, and Bach 2011): \( O(\sqrt{\kappa_s} \phi(\rho_s)) = \sqrt{\kappa_s} \log(2/\rho_s) \)
- Proximal SVRG (Xiao and Zhang 2014): \( O(\kappa_s \phi(\rho_s)) = \frac{\theta}{(1-4\theta)\rho_s - 4\theta} (\kappa_s + 4) \)
- Accelerated Proximal SVRG (Nitanda 2014): \( O(\kappa_s \phi(\rho_s)) = \sqrt{\kappa_s} \log(\frac{1-\rho_s}{1-\gamma\rho_s}) \)
- SAGA (Defazio, Bach, and Lacoste-Julien 2014): \( O(\kappa_s \phi(\rho_s)) = \frac{3n}{p} \left( \frac{3n_s}{n} + 1 \right) \)
- MISO (Mairal 2013): \( O(\kappa_s \phi(\rho_s)) = \frac{n_s}{\rho_s} \)

where \( \theta \in (0, 0.25) \) and satisfies \( (1 - 4\theta)\rho_s - 4\theta > 0 \), and \( p \in (0, 1) \) and satisfies \( \rho_s > \frac{\theta(2+p)}{1-p} \). The convergence rate for SAGA and MISO on strongly convex problems are derived from each convergence rate on general convex problems with some mathematical transformations.

For SAGA:
\[
\mathbb{E} \hat{P}_s(\tilde{x}_s) - \hat{P}_s(x^*_s) \leq \frac{3n}{T_s} \left( \frac{3L_{m,s}}{2n} \|x_{s-1} - x_s^*\|^2 + \tilde{f}_{\gamma_s}(\tilde{x}_{s-1}) - \nabla \tilde{f}_{\gamma_s}(x_s^*)^T(x_{s-1} - x_s^*) - \tilde{f}_{\gamma_s}(x_s^*) \right) \quad \text{(by Defazio, Bach, and Lacoste-Julien 2014)}
\leq \frac{3n}{T_s} \left( \frac{3L_{m,s}}{n\mu} + 1 \right) \left( \hat{P}_s(\tilde{x}_{s-1}) - \hat{P}_s(x^*_s) \right)
\]
where second inequality come from \( \frac{1}{2} \|x_{s-1} - x_s^*\|^2 \leq \hat{P}_s(\tilde{x}_{s-1}) - \hat{P}_s(x_s^*) \) and \( -\nabla \tilde{f}_{\gamma_s}(x_s^*)^T(x_{s-1} - x_s^*) \leq \tau(\tilde{x}_{s-1} - r(x_s^*)) \).

For MISO:
\[
\mathbb{E} \hat{P}_s(\tilde{x}_s) - \hat{P}_s(x^*_s) \leq \frac{nL_{m,s}}{2T_s} \|\tilde{x}_{s-1} - x_s^*\|^2 \quad \text{(by Mairal 2013)}
\leq \frac{nL_{m,s}}{T_s \mu} \left( \hat{P}_s(\tilde{x}_{s-1}) - \hat{P}_s(x_s^*) \right)
\]

**References**


