A. Proof of Lemma 1

In this section, we prove the results on the error generated when solving the subproblem (3.2) inexacty by Procedure 1. Before proving Lemma 1, we will first prove a simpler case in Lemma 3, where the subproblem iterator \( S \) is the proximal gradient step.

**Lemma 3.** Take Assumption 1. Suppose in Procedure 1, we choose \( S \) as the proximal gradient step with step size \( \gamma = \eta \frac{\lambda_{\min}(M)}{\lambda_{\max}(M)} \), and is repeat it \( p \) times, where \( p \geq 1 \). Then, \( w_{t+1} = w_{t+1}^p \) is an approximate solution to (3.2) that satisfies

\[
0 \in \nabla \psi(w_{t+1}) + \frac{1}{\eta} M(w_{t+1} - w_t) + \nabla \hat{t} + M \varepsilon_{t+1}^p,
\]

(A.1)

\[
\|\varepsilon_{t+1}^p\|_M \leq \frac{c(p)}{\eta} \|w_{t+1} - w_t\|_M,
\]

(A.2)

where

\[
c(p) = (\kappa(M) + 1)\kappa(M) \frac{\tau^p + \tau^{p-1}}{1 - \tau^p},
\]

and \( \tau = \sqrt{1 - \kappa^2(M)} < 1 \).

**Proof of Lemma 3.** The optimization problem in (3.2) is of the form

\[
\min_{y \in \mathbb{R}^d} h_1(y) + h_2(y),
\]

(A.3)

for \( h_1(y) = \psi(y) \) and \( h_2(y) = \frac{1}{2\eta} \|y - w_t\|_M^2 + \| \hat{\nabla} \). With our choice of \( S \) as the proximal gradient descent step, the iterations in Procedure 1 are

\[
w_{t+1}^0 = w_t,
\]

\[
w_{t+1}^i = \text{prox}_{\gamma h_i}(w_{t+1}^i - \gamma \nabla h_2(w_{t+1}^i)),
\]

\[
w_{t+1} = w_{t+1}^p,
\]

where \( i = 0, 1, ..., p - 1 \). From the definition of \( \text{prox}_{\gamma h_i} \), we have

\[
0 \in \partial h_1(w_{t+1}^p) + \nabla h_2(w_{t+1}^p) + \frac{1}{\gamma} (w_{t+1}^p - w_{t+1}^{p-1}).
\]

Compare this with (A.1) gives

\[
M \varepsilon_{t+1}^p = \frac{1}{\gamma} (w_{t+1}^p - w_{t+1}^{p-1}) + \nabla h_2(w_{t+1}^p) - \nabla h_2(w_{t+1}).
\]

To bound the right hand side, let \( w_{t+1}^* \) be the solution of (A.3), \( \alpha = \frac{\lambda_{\min}(M)}{\eta} \), and \( \beta = \frac{\lambda_{\max}(M)}{\eta} \). Then \( h_1(y) \) is convex and \( h_2(y) \) is \( \alpha \)-strongly convex and \( \beta \)-Lipschitz differentiable. Consequently, Prop. 26.16(ii) of (Bauschke et al., 2017) gives

\[
\|w_{t+1} - w_{t+1}^*\| \leq \tau^i \|w_{t+1}^0 - w_{t+1}^i\|, \quad \forall i = 0, 1, ..., p,
\]

where \( \tau = \sqrt{1 - \gamma(2\alpha - \gamma^2)} \).

Let \( a_i = \|w_{t+1}^i - w_{t+1}^*\| \). Then, \( a_i \leq \tau^i a_0 \). We can derive

\[
\|M \varepsilon_{t+1}^p\| \leq \frac{1}{\gamma} (\frac{1}{\gamma} + \beta) \|w_{t+1}^p - w_{t+1}^{p-1}\|
\]

\[
\leq \left( \frac{1}{\gamma} + \beta \right) (a_p + a_{p-1}) \leq \left( \frac{1}{\gamma} + \beta \right) (\tau^p + \tau^{p-1}) a_0.
\]

On the other hand, we have

\[
\|w_{t+1} - w_t\| \geq a_0 - a_p \geq (1 - \tau^p) a_0.
\]

Combining these two equations yields

\[
\|M \varepsilon_{t+1}^p\| \leq b(p) \|w_{t+1} - w_t\|,
\]

(A.4)

where

\[
b(p) = \left( \frac{1}{\gamma} + \frac{\lambda_{\max}(M)}{\eta} \right) \frac{\tau^p + \tau^{p-1}}{1 - \tau^p}.
\]

(A.5)

Finally, let the eigenvalues of \( M \) be \( 0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_d \), with orthonormal eigenvectors \( v_1, v_2, ..., v_d \). Let \( \varepsilon_{t+1}^p \) and \( w_{t+1} - w_t \) be decomposed by

\[
\varepsilon_{t+1}^p = \sum_{i=1}^d \alpha_i v_i,
\]

\[
w_{t+1} - w_t = \sum_{i=1}^d \beta_i v_i,
\]

then

\[
\|\varepsilon_{t+1}^p\|_M = \sqrt{\sum_{i=1}^d \lambda_i \alpha_i^2} \leq \sqrt{\frac{1}{\lambda_{\min}(M)} \sum_{i=1}^d \lambda_i^2 \alpha_i^2},
\]

\[
= \sqrt{\frac{1}{\lambda_{\min}(M)} \|M \varepsilon_{t+1}^p\|},
\]

\[
\|w_{t+1} - w_t\| = \sqrt{\sum_{i=1}^d \beta_i^2} \leq \sqrt{\frac{1}{\lambda_{\min}(M)} \sum_{i=1}^d \lambda_i \beta_i^2},
\]

\[
= \sqrt{\frac{1}{\lambda_{\min}(M)} \|w_{t+1} - w_t\|_M}.
\]

Combine these two inequalities with (A.4), we arrive at

\[
\|\varepsilon_{t+1}^p\|_M \leq c(p) \|w_{t+1} - w_t\|_M,
\]

(A.6)

where

\[
c(p) = \frac{1}{\lambda_{\min}(M)} b(p) = \frac{1}{\lambda_{\min}(M)} \frac{1}{\lambda_{\min}(M)} \frac{\tau^p + \tau^{p-1}}{1 - \tau^p}.
\]

\[\square\]
Now, we are ready to prove Lemma 1, the techniques are similar to the proof of Lemma 3.

**Proof of Lemma 1.** We want to find \( c(p) \) such that

\[
0 \in \partial \psi(w_{t+1}) + \frac{1}{\eta} M(w_{t+1} - w_t) + \nabla \epsilon_t + M \xi_t^{p},
\]

(A.7)

\[
\| \xi_t^{p} \|_M \leq \frac{c(p)}{\eta} \| w_{t+1} - w_t \|_M,
\]

(A.8)

Take \( i = r - 1 \) and \( j = p_0 - 1 \), then the optimality condition of the problem in line 5 of Algorithm 3 is

\[
0 \in \partial \psi(u^{(r-1,p_0)}_{t+1}) + \frac{1}{\gamma} \left( w^{(r-1,p_0)}_{t+1} - u^{(r-1,p_0)}_{t+1} \right) + \nabla h_2(u^{(r-1,p_0)}_{t+1}),
\]

compare this with (A.7), we have

\[
M \xi_t^{p} = \frac{1}{\gamma} \left( w^{(r-1,p_0)}_{t+1} - u^{(r-1,p_0)}_{t+1} \right) + \nabla h_2(u^{(r-1,p_0)}_{t+1}) - \frac{1}{\eta} M(w_{t+1} - w_t) - \nabla \epsilon_t
\]

\[
= \frac{1}{\gamma} \left( u^{(r-1,p_0)}_{t+1} - u^{(r-1,p_0)}_t \right) + \frac{1}{\eta} M(u^{(r-1,p_0)}_{t+1} - w_t)
\]

where

\[
u^{(r-1,p_0)}_{t+1} = w^{(r-1,p_0)}_{t+1} + \frac{\theta_{p_0} - 1}{\theta_{p_0} - 1} \left( w^{(r-1,p_0)}_{t+1} - u^{(r-1,p_0)}_{t+1} \right)
\]

As a result,

\[
\| M \xi_t^{p} \| \leq \frac{1}{\gamma} \| u^{(r-1,p_0)}_{t+1} - u^{(r-1,p_0)}_t \| + \frac{1}{\eta} \| M(w_{t+1} - w_t) - \nabla \epsilon_t \|
\]

(A.9)

\[
\leq \frac{1}{\gamma} \| u^{(r-1,p_0)}_{t+1} - u^{(r-1,p_0)}_t \| + \frac{1}{\eta} \| M \| \| w_{t+1} - w_t \|
\]

(A.10)

Let the solution of (3.2) be \( u^*_{t+1} \). By Theorem 4.4 of (Beck & Teboulle, 2009), for any \( 0 \leq i \leq r - 1 \) and \( 0 \leq j \leq p_0 \) we have

\[
\Psi(w^{(i,j)}_{t+1}) - \Psi(u^*_{t+1}) \leq \frac{2\lambda_{\text{max}}(M) \| u^{(i,0)}_{t+1} - u^*_{t+1} \|^2}{\eta^2}.
\]

On the other hand, the strong convexity of \( \Psi = h_1 + h_2 \) gives

\[
\Psi(w^{(i,j)}_{t+1}) - \Psi(u^*_{t+1}) \geq \frac{\lambda_{\text{min}}(M)}{2\eta} \| w^{(i,j)}_{t+1} - u^*_{t+1} \|^2.
\]

Therefore,

\[
\| w^{(i,j)}_{t+1} - u^*_{t+1} \| \leq \sqrt{\frac{4\kappa(M)}{\eta^2}} \| w^{(i,0)}_{t+1} - u^*_{t+1} \|. \tag{A.11}
\]

Now, let us use (A.11) repeatedly to bound the right hand side of (A.10). For example, the first term can be bounded as

\[
\| w^{(r-1,p_0)}_{t+1} - w^{(r-1,p_0-1)}_{t+1} \| \leq \frac{1}{\gamma} \| w^{(r-1,p_0)}_{t+1} - w^*_{t+1} \|
\]

Similarly, the rest of the terms can be bounded as follows,

\[
\| w^{(r-1,p_0-1)}_{t+1} - w^*_{t+1} \| \leq \frac{1}{\gamma} \| w^{(0)}_{t+1} - w^*_{t+1} \|
\]

where in the first and third estimate we have used \( \frac{\theta_{p_0} - 2}{\theta_{p_0} - 1} \leq \frac{\theta_{p_0} - 2}{\theta_{p_0} - 1} \leq 1 \).
where
\[ \frac{\theta_{p_0-2}}{\theta_{p_0-1}} < 1. \]
On the other hand, we have
\[
\|w_{t+1} - w_t\| = \|w_{t+1}^{(r-1,p_0)} - w_t^{(0,0)}\|
\geq \|w_{t+1}^{(0,0)} - w_{t+1}\| - \|w_{t+1}^{(r-1,p_0)} - w_{t+1}^*\|
\geq (1 - \frac{4\kappa(M)}{p_0^2}) \|w_{t+1}^{(0,0)} - w_{t+1}^*\|.
\]
As a result, taking \( \gamma = \frac{\lambda_{\max}(M)}{\eta}, \)
\( w_{t+1}^{(0,0)} = w_t, \) \( w_{t+1}^{(r-1,p_0)} = w_{t+1} \)
and \( \tau = \left( \frac{4\kappa(M)}{p_0^2} \right)^{\frac{1}{p_0}} \) yields
\[
\|M\varepsilon_{t+1}^p\|_M \leq 2\frac{\lambda_{\max}(M)}{\eta} \frac{b(p)}{1 - \tau^p} \|w_{t+1} - w_t\|,
\]
where
\[
b(p) = \tau^{p_0} \left( \left( \frac{4\kappa(M)}{p_0 - 1} \right)^{\frac{1}{2}} + \left( \frac{4\kappa(M)}{p_0 - 2} \right)^{\frac{1}{2}} \right) + \tau^p + \tau^{p_0} \left( \frac{4\kappa(M)}{p_0 - 1} \right)^{\frac{1}{2}}.
\]
(A.12)
Similarly to the end of proof of Lemma 3, we have
\[
\|M\varepsilon_{t+1}^p\|_M \leq 2\frac{\kappa(M)}{\eta} \frac{b(p)}{1 - \tau^p} \|w_{t+1} - w_t\|_M.
\]
Now, let us choose \( p_0 \) such that \( \tau = \left( \frac{4\kappa(M)}{p_0^2} \right)^{\frac{1}{p_0}} \) is minimized, a simple calculation yields
\[
p_0^* = 2e\sqrt{\kappa(M)}.
\]
In order for \( p_0 \) to be an integer, we can take
\[
p_0 = \left[ 2e\sqrt{\kappa(M)} \right],
\]
then
\[
\tau = \left( \frac{4\kappa(M)}{p_0^2} \right)^{\frac{1}{p_0}} \leq \left( \frac{1}{c^2} \right)^{\frac{1}{2(2\sqrt{\kappa(M)} + 1)}} \leq \left( \frac{1}{c^2} \right)^{\frac{1}{2(2\sqrt{\kappa(M)} + 1)}}
\leq \exp \left( -\frac{1}{2e\sqrt{\kappa(M)} + 1} \right).
\]
Finally, Let us show that \( b(p) \) in (A.12) can be bounded by \( 7\tau^p \), and the desired bound (A.8) on \( \|\varepsilon_{t+1}^p\|_M \) follows.
First, we have
\[
\tau^{-p_0} \left( \frac{4\kappa(M)}{p_0 - 1} \right)^{\frac{1}{2}} = \left( \frac{p_0}{p_0 - 1} \right)^{\frac{1}{p_0}},
\]
and
\[
p_0 = \left[ 2e\sqrt{\kappa(M)} \right] \geq [2e] = 6.
\]
On the other hand, a simple calculation shows that \( \left( \frac{p_0}{p_0 - 1} \right)^{\frac{1}{p_0}} \) is decreasing in \( p_0 \), therefore
\[
\tau^{-p_0} \left( \frac{4\kappa(M)}{p_0 - 1} \right)^{\frac{1}{2}} \leq \left( \frac{6}{5} \right)^{\frac{1}{p_0}} < 2,
\]
Similarly, one can show that
\[
\tau^{-p_0} \left( \frac{4\kappa(M)}{p_0 - 2} \right)^{\frac{1}{2}} \leq \left( \frac{6}{5} \right)^{\frac{1}{p_0}} < 2.
\]
Combining these two inequalities with (B.2) yields
\[
b(p) \leq 7\tau^p.
\]
\[\square\]

B. Proof of Theorem 1

In this section, we proceed to establish the convergence of inexact preconditioned SVRG as in Algorithm 1. The proof is similar to that of Theorem D.1 of (Allen-Zhu, 2018).

Before proving Theorem 1, let us first prove several lemmas.

First, the inexact optimality condition (4.1) gives the following descent:

**Lemma 4.** Under Assumption 1, suppose that (4.1) holds. Then, for any \( u \in \mathbb{R}^d \) we have
\[
\langle \nabla_t, w_t - u \rangle + \psi(w_{t+1}) - \psi(u)
\leq \langle \nabla_t, w_t - w_{t+1} \rangle + \frac{\|u - w_t\|^2_M}{2\eta}
- \frac{1}{2\eta} \|w_{t+1} - w_t\|^2_M
- \frac{1}{\eta} \|w_{t+1} - w_t\|^2_M
+ \langle M\varepsilon_{t+1}^p, u - w_{t+1} \rangle.
\]

**Proof.** First, let us rewrite the left hand side as
\[
\langle \nabla_t, w_t - u \rangle + \psi(w_{t+1}) - \psi(u)
= \langle \nabla_t, w_t - w_{t+1} \rangle + \langle \nabla_t, w_{t+1} - u \rangle + \psi(w_{t+1}) - \psi(u).
\]
By (4.1) and the definition of subdifferential we have
\[
\psi(u) \geq \psi(w_{t+1}) - \langle \nabla_t, w_{t+1} - w_t \rangle + M\varepsilon_{t+1}^p, u - w_{t+1} \rangle.
\]
Combining these two gives
\[
\langle \nabla_t, w_t - u \rangle + \psi(w_{t+1}) - \psi(u)
\leq \langle \nabla_t, w_t - w_{t+1} \rangle
+ \langle \frac{1}{\eta} M(w_{t+1} - w_t) + M\varepsilon_{t+1}^p, u - w_{t+1} \rangle
= \langle \nabla_t, w_t - w_{t+1} \rangle + \frac{\|u - w_t\|^2_M}{2\eta}
- \frac{1}{2\eta} \|w_{t+1} - w_t\|^2_M
- \frac{1}{2\eta} \|w_{t+1} - w_t\|^2_M
+ \langle M\varepsilon_{t+1}^p, u - w_{t+1} \rangle,
\]
where in the last equality we have applied
\[
\langle a - b, c - a \rangle_M = -\frac{1}{2} \|a - b\|^2_M - \frac{1}{2} \|a - c\|^2_M + \frac{1}{2} \|b - c\|^2_M.
\]
\[\square\]
Based on lemma 4, we have

**Lemma 5.** Under Assumption 1, if the iterator $S$ in Procedure 1 is proximal gradient descent or FISTA with restart, then, for any $a > 0$, $\eta \leq \frac{1-2c\eta}{2\eta_L^2}$, and $u \in \mathbb{R}^d$ we have

$$
\mathbb{E}[F(w_{t+1}) - F(u)] \\
\leq \mathbb{E}[\eta \nabla_t - \nabla f(w_t)]^2_{M-1} + \frac{1 - \eta \sigma_M}{2\eta} \|u - w_t\|^2_M \\
- \left( \frac{1}{2\eta} - \frac{c(p)}{2\eta a} \right) \|u - w_{t+1}\|^2_M.
$$

**Proof.** We have

$$
\mathbb{E}[F(w_{t+1}) - F(u)] \\
= \mathbb{E}[f(w_{t+1}) - f(u) + \psi(w_{t+1}) - \psi(u)] \\
\leq \mathbb{E}[f(w_t) + \nabla f(w_t), w_{t+1} - w_t] \\
+ \frac{L_M}{2} \|w_t - w_{t+1}\|^2_M - f(u) + \psi(w_{t+1}) - \psi(u)] \\
\mathbb{E}[(\nabla f(w_t), w_t - u) - \frac{\sigma_M}{2} \|u - w_t\|^2_M \\
+ \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L_M}{2} \|w_t - w_{t+1}\|^2_M \\
+ \psi(w_{t+1}) - \psi(u)] \\
= \mathbb{E}[(\nabla_t, w_t - u) - \frac{\sigma_M}{2} \|u - w_t\|^2_M \\
+ \langle \nabla f(w_t), w_{t+1} - w_t \rangle \\
+ \frac{L_M}{2} \|w_t - w_{t+1}\|^2_M + \psi(w_{t+1}) - \psi(u)],
$$

(B.1)

where the first and second inequality are due to the strong convexity and smoothness under $\| \cdot \|_M$ in Assumption 1, respectively. The last equality is due to $\mathbb{E}[\nabla_t] = \nabla f(w_t)$.

On the other hand, recall that Lemma 4 gives

$$
\langle \nabla_t, w_t - u \rangle + \psi(w_{t+1}) - \psi(u) \\
\langle \nabla_t, w_t - w_{t+1} \rangle + \frac{1 - \eta \sigma_M}{2\eta} \|u - w_t\|^2_M \\
- \left( \frac{1}{2\eta} - \frac{c(p)}{2\eta a} \right) \|u - w_{t+1}\|^2_M + \langle M \epsilon^p_{t+1}, u - w_{t+1} \rangle,
$$

For the last term we can apply Cauchy-Schwartz as follows,

$$
\langle M \epsilon^p_{t+1}, u - w_{t+1} \rangle \leq \epsilon^p_{t+1} \|u - w_{t+1}\|_M,
$$

from Lemma 3 and Lemma 1 we know that

$$
\|\epsilon^p_{t+1}\|_M \leq \frac{c(p)}{\eta} \|w_{t+1} - w_t\|_M.
$$

Therefore, by Young’s inequality, we have for any $a > 0$ that

$$
\langle M \epsilon^p_{t+1}, u - w_{t+1} \rangle \\
\leq \frac{c(p) a}{2\eta} \|w_{t+1} - w_t\|^2_M + \frac{c(p)}{2a\eta} \|u - w_{t+1}\|^2_M.
$$

Applying this to Lemma 4 yields

$$
\langle \nabla_t, w_t - u \rangle + \psi(w_{t+1}) - \psi(u) \\
\langle \nabla_t, w_t - w_{t+1} \rangle + \frac{1 - \eta \sigma_M}{2\eta} \|u - w_t\|^2_M \\
- \left( \frac{1}{2\eta} - \frac{c(p)}{2\eta a} \right) \|u - w_{t+1}\|^2_M \\
- \left( \frac{1}{2\eta} - \frac{c(p) a}{2\eta} \right) \|u - w_{t+1}\|^2_M
$$

(B.2)

where in the second inequality we have applied

$$
\langle u_1, u_2 \rangle = \langle M^{-\frac{1}{2}} u_1, M^{\frac{1}{2}} u_2 \rangle \leq \|u_1\|_{M^{-\frac{1}{2}}} \|u_2\|_M \\
\leq \frac{1}{2b} \|u_1\|^2_{M^{-\frac{1}{2}}} + \frac{b}{2} \|u_2\|^2_M
$$

for any $b > 0$.

Finally, since $\eta \leq \frac{1-2c\eta}{2\eta_L^2}$, we have $\frac{\eta}{2(1-c(p)a-\eta L_f^2)} \leq \eta$, which gives the desired result.

\[ \square \]

**Lemma 6.** Under Assumption 1, we have

$$
\mathbb{E}[\|\nabla_t - \nabla f(w_t)\|^2_{M-1}] \leq \langle L_f^M \rangle^2 \|w_0 - w_t\|^2_M.
$$

Acceleration of SVRG and Katyusha X by Inexact Preconditioning
Proof. We have

\[
E[\|\nabla \ell - \nabla f(w_t)\|^2_{M^{-1}}] = E[\|\nabla f(w_0) + \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_0) - \nabla f(w_t)\|^2_{M^{-1}}]
\]

where in the first inequality, we have applied \(E[\frac{1}{N} \nabla f] = 0\) and in the second inequality follows from Assumption 1.

Lemma 7. (Fact 2.3 of (Allen-Zhu, 2018)). Let \(C_1, C_2, \ldots\) be a sequence of numbers, and \(N \sim \text{Geom}(p)\), then

1. \(E_N [C_N - C_{N+1}] = \frac{p}{1-p} E_N [C_0 - C_N]\), and

2. \(E_N [C_N] = (1-p) E \left[ C_{N+1}\right] + p C_0\).

Lemma 8. Under Assumption 1, if \(\eta \leq \min\{\frac{1-2c(p)a}{2\sqrt{mL_f^2}}, \frac{1}{2\sqrt{mL_f^2}}\}\) and \(m \geq 2\), then for any \(u \in \mathbb{R}^d\) we have

\[
E[F(w_{D+1}) - F(u)] \leq E[-\frac{1}{4m\eta} \|w_{D+1} - w_0\|^2_M + \frac{(w_0 - w_{D+1}, w_0 - u)_M}{m\eta} - \left(\frac{\sigma^2}{4} - \frac{c(p)}{2a\eta}\right) \|w_{D+1} - u\|^2_M].
\]

Proof. By Lemmas 5 and 6, we know that

\[
E[F(w_{t+1}) - F(u)] = E[\eta (L_f^M)^2 \|w_0 - w_t\|^2_M + \frac{1 - \eta \sigma^2}{2\eta} \|w_0 - u\|^2_M
\]

where the first equality follows from the item 1 of Lemma 7 with \(C_N = \|u - w_N\|^2_M\), the second inequality follows from item 2 with \(C_N = \|w_0 - w_{D+1}\|^2_M\), item 2 with \(C_N = \|u - w_0\|^2_M\), and item 1 with \(C_N = \|u - w_D\|^2_M\), then third inequality makes use of \(m \geq 2\) and the fourth inequality makes use of \(\eta \leq \frac{1}{2\sqrt{mL_f^2}}\).

Now, let us proceed to prove Theorem 1. With Lemma 8, it can be proved in a similar way as Theorem 3 of (Hannah et al., 2018b).

Proof of Theorem 1. Without loss of generality, we can as-
According to Lemma 8, for any \( u \in \mathbb{R}^d \), and \( \eta \leq \min\{\frac{1 - 2c(p)\alpha}{2L_f^M}, \frac{1}{2\sqrt{mL_f}}\} \) we have
\[
\mathbb{E}[F(x^{j+1}) - F(u)] \\
\leq \mathbb{E}[\frac{1}{4\eta m}(\|x^{j+1} - x^j\|_M^2 + \frac{1}{2\eta m}\|x^j - u\|_M^2) \\
+ \frac{1}{m\eta}(\|x^j - u\|_M^2 - (\sigma_f^M/4 - c(p)/2\eta))\|x^{j+1} - u\|_M^2],
\]
or equivalently,
\[
\mathbb{E}[F(x^{j+1}) - F(u)] \\
\leq \mathbb{E}[\frac{1}{4\eta m}\|x^{j+1} - x^j\|_M^2 + \frac{1}{2\eta m}\|x^j - u\|_M^2] \\
- \frac{1}{2\eta m}\|x^{j+1} - u\|_M^2 - (\sigma_f^M/4 - c(p)/2\eta)\|x^{j+1} - u\|_M^2].
\]

In the following proof, we will omit \( \mathbb{E} \).

Setting \( u = x^* = 0 \) and \( u = x^j \) yields the following two inequalities:
\[
F(x^{j+1}) \leq \frac{1}{4\eta m}(\|x^{j+1} - x^j\|_M^2 + 2\|x^j\|_M^2) \\
- \frac{1}{2\eta m}(1 + \frac{1}{2\eta m}(\sigma_f^M - 2c(p)/\eta))\|x^{j+1}\|_M^2.
\]

(B.3)

\[
F(x^{j+1}) - F(x^j) \\
\leq - \frac{1}{4\eta m}(1 + \frac{1}{2\eta m}(\sigma_f^M - 2c(p)/\eta))\|x^{j+1} - x^j\|_M^2.
\]

(B.4)

Define \( \tau = \frac{1}{2\eta m}(\sigma_f^M - 2c(p)/\eta) \), multiply \( (1 + 2\tau) \) to (B.3), then add it to (B.5) yields
\[
2(1 + \tau)F(x^{j+1}) - F(x^j) \\
\leq \frac{1}{2m\eta}(1 + 2\tau)(\|x^j\|_M^2 - (1 + \tau)\|x^{j+1}\|_M^2).
\]

Multiplying both sides by \( (1 + \tau)^j \) gives
\[
2(1 + \tau)^j F(x^{j+1}) - (1 + \tau)^j F(x^j) \\
\leq \frac{1}{2m\eta}(1 + 2\tau)(1 + \tau)^j\|x^j\|_M^2 - (1 + \tau)^j\|x^{j+1}\|_M^2).
\]

Summing over \( j = 0, 1, \ldots, k - 1 \), we have
\[
(1 + \tau)^k F(x^k) + \sum_{j=0}^{k-1} (1 + \tau)^j F(x^j) - F(x^0) \\
\leq \frac{1}{2m\eta}(1 + 2\tau)(\|x^0\|_M^2 - (1 + \tau)^k\|x^k\|_M^2).
\]

Since \( F(x^j) \geq 0 \), we have
\[
F(x^k)(1 + \tau)^k \leq F(x^0) + \frac{1}{2m\eta}(1 + 2\tau)\|x^0\|_M^2.
\]

By the strong convexity of \( F \), we have \( F(x^0) + \frac{\sigma_f^M}{2\tau}\|x^0\|_M^2 \), therefore
\[
F(x^k)(1 + \tau)^k \leq F(x^0)(2 + \frac{1}{2\tau}). \tag{B.6}
\]

Finally, recall that \( a > 0 \) can be chosen arbitrarily, so we can take
\[
a = \frac{4c(p)}{\eta \sigma_f^M},
\]
and
\[
\eta \leq \min\{\frac{1 - 2c(p)\alpha}{2L_f^M}, \frac{1}{2\sqrt{mL_f}}\} \\
= \min\{\frac{1 - 8c^2(p)/\eta \sigma_f^M}{2L_f^M}, \frac{1}{2\sqrt{mL_f}}\}, \tag{B.7}
\]
\[
\tau = \frac{1}{2\eta}(\sigma_f^M - 2c(p)/\eta) = \frac{1}{4}\eta \sigma_f^M.
\]

In order for the choice of \( \eta \) in (B.7) to be possible, we need
\[
2L_f^M\eta^2 - \eta + 8c^2(p)/\sigma_f^M \leq 0 \tag{B.8}
\]

to have one solution at least, which requires
\[
64\kappa_f^M c^2(p) \leq 1,
\]
under which \( \eta = \frac{1}{2L_f^M} \) satisfy (B.8). As a result, \( m \geq 4 \) makes (B.7) into
\[
\eta \leq \frac{1}{2\sqrt{mL_f}},
\]
and the desired convergence result follows from (B.6). \( \Box \)

C. Proof of Lemma 2

Proof. From Lemma 1, we know that
\[
c(p) = 14\kappa(M)\frac{\tau^p}{1 - \tau^p},
\]
where
\[
\tau = \exp(-\frac{1}{2e\sqrt{\kappa(M)} + 1}).
\]

Therefore, in order for \( 64\kappa_f^M c^2(p) \leq 1 \), we need
\[
\kappa_f^M \kappa^2(M)(\frac{\tau^p}{1 - \tau^p})^2 \leq \frac{1}{64 \times 14^2} = c_1,
\]
which is equivalent to
\[ \tau^p \leq \frac{c_1}{\sqrt[2]{\kappa^p} + \sqrt{c_1}}. \]

Thus, it suffices to require that
\[ \left(\exp\left(-\frac{1}{2e\sqrt{\kappa}M}\right) + 1\right)^p \leq \frac{c}{\sqrt[2]{\kappa^p} M} + \sqrt{c_1}, \]
which gives
\[ p \geq (2e\sqrt{\kappa}M + 1) \ln \frac{\sqrt[2]{\kappa^p} M + \sqrt{c_1}}{c_1}. \]

\[ \square \]

**D. Proof of Theorem 2**

The proof of Theorem 2 is similar to that of Theorem 4.3 of (Allen-Zhu, 2018), so we provide a proof sketch here and omit the details.

1. In (Allen-Zhu, 2018), the proof of Theorem 4.3 is based on Lemma 3.3, here the proof of Theorem 2 is based on Lemma 8, which is an analog of Lemma of 3.3 in our settings.

2. Based on Lemma 8, the proof of Theorem 2 follows in nearly the same way as Theorem 4.3 of (Allen-Zhu, 2018), the only difference is that one needs to replace \( a \) by \( a_j = 2\exp(p) / a \eta \).

3. By setting
\[ a = \frac{4e\exp(p)}{\eta M^2}, \]
and
\[ 64\kappa^p e^2 (p) \leq 1 \]
as in the proof of Theorem 1, the \( \tau \) in Theorem 4.3 of (Allen-Zhu, 2018) becomes \( 1/\sqrt[2]{m} \eta \sigma \kappa^p \), and the convergence result of Theorem 2 follows.

**E. Proof of Theorems 3 and 4**

**Proof of Theorem 3.** From Remark 5, we know that the gradient complexity of SVRG can be expressed as
\[ C_1(m, \varepsilon) = \mathcal{O}\left(\frac{n + m}{\ln(1 + \frac{\sqrt{n}}{\kappa \sigma})} \ln \frac{1}{\varepsilon}\right). \]

Taking the largest possible step size \( \eta = \frac{1}{\sqrt[2]{M} \kappa} \) as in Theorem 1, we have
\[ C_1(m, \varepsilon) = \mathcal{O}\left(\frac{n + m}{\ln(1 + \frac{\sqrt{n}}{\kappa \sigma})} \ln \frac{1}{\varepsilon}\right). \]

Let us first find the optimal \( m = m^* \) for SVRG, let
\[ g(m) = \frac{n + m}{\ln(1 + \frac{\sqrt{n}}{\kappa \sigma})}, \]
then
\[ g'(m) = \frac{\ln(1 + \frac{\sqrt{n}}{\kappa \sigma}) - \frac{\sqrt{n}}{\kappa \sigma} m + m}{\ln^2(1 + \frac{\sqrt{n}}{\kappa \sigma})}. \]

Taking derivative to the numerator gives
\[ \left[ \ln(1 + \frac{\sqrt{n}}{\kappa \sigma}) - \frac{\sqrt{n}}{\kappa \sigma} m + m \right]' = (n + m) \frac{\frac{1}{2} m \kappa f^2}{\left(1 + \frac{\sqrt{n}}{\kappa \sigma}\right)^2} > 0, \]

Therefore, \( m^* \) is given by \( g'(m) = 0 \). Let \( z = \frac{\sqrt{n}}{\kappa \sigma} > 0 \), then
\[ g'(m) = \frac{\ln(1 + z) - \frac{z}{1 + z} \frac{n + m}{2m}}{\ln^2(1 + z)}. \]

Since \( \ln(1 + z) > \frac{z}{1 + z} \) for \( z > 0 \), we know that \( g'(n) > 0 \), therefore, \( m^* < n \).

Let \( m = n^s \) where \( 0 < s < 1 \), we would like to have \( g'(n^s) < 0 \), i.e.,
\[ \ln(1 + z) - \frac{z}{1 + z} \frac{n + m}{2m} \]
since \( \kappa f > n^{\frac{1}{2}} \), we have \( z = \frac{\sqrt{n}}{\kappa \sigma} < \frac{1}{8} \), on the other hand, we have
\[ \left[ \frac{n + m}{\ln^2(1 + z)} \right]' = \frac{1}{\ln^2(1 + z)} > 0. \]

Therefore, it suffices to have
\[ n^{1-s} > 18 \ln \frac{9}{8} - 1 := c_0 > 1. \]

As a result, we have \( m^* \in (\frac{n}{c_0}, n) \), and
\[ C_1(m^*, \varepsilon) = \mathcal{O}\left(\frac{n + m^*}{\ln(1 + \frac{\sqrt{n}}{\kappa \sigma} \ln \frac{1}{\varepsilon})}\right) \]
\[ = \mathcal{O}\left(\frac{n}{\sqrt{n}} \ln \frac{1}{\varepsilon}\right) = \mathcal{O}\left(\kappa f \sqrt{n} \ln \frac{1}{\varepsilon}\right), \]

where in the second equality we have used \( \kappa f > n^{\frac{1}{2}} \).

For our iPreSVRG in Algorithm 1, we have
\[ C_1^i(m, \varepsilon) = \mathcal{O}\left(\frac{n + (1 + pd)m}{\ln(1 + \frac{\sqrt{n}}{\kappa \sigma})} \ln \frac{1}{\varepsilon}\right), \]
thanks to Lemma 2, \( p \) can be chosen as
\[
p = \mathcal{O}(\sqrt{\kappa(M)} \ln (\sqrt{\kappa_f^M} \kappa(M)))
\]
furthermore, we can take \( \eta = \frac{1}{\sqrt{mL_f}} \) due to Theorem 1.
Under these settings, we have
\[
C_1'(m, \varepsilon) = \mathcal{O}(\frac{n + (1+pd)m}{\ln(1 + \frac{1}{\sqrt{m\kappa_f^M}})} \ln \frac{1}{\varepsilon}).
\]
Let us take \( m = m' = \left\lceil \frac{n}{1+pd} \right\rceil \).
If \( n > 1 + pd \), or equivalently \( \kappa_f < n^2d^{-2} \), then
\[
C_1'(m', \varepsilon) = \mathcal{O}(\frac{n}{\ln(1 + \frac{1}{\sqrt{n\kappa_f^M}})} \ln \frac{1}{\varepsilon}).
\]
Since \( p = \mathcal{O}\left(\sqrt{\kappa(M)} \ln (\sqrt{\kappa_f^M} \kappa(M))\right) \), we know that
when \( (\kappa_f^M)^2 \sqrt{\kappa(M)} d < n \), or equivalently \( \kappa_f < n^2d^{-2} \),
we have
\[
\ln(1 + \frac{1}{\sqrt{n\kappa_f^M}}) = \mathcal{O}(\ln n),
\]
therefore
\[
C_1'(m', \varepsilon) = \mathcal{O}(n \ln \frac{1}{\varepsilon}),
\]
and
\[
\min_{m \geq 1} \frac{C_1'(m, \varepsilon)}{C_1'(m', \varepsilon)} \leq \frac{C_1'(m', \varepsilon)}{C_1'(m^*, \varepsilon)} = \mathcal{O}(\frac{\sqrt{n}}{\kappa_f}).
\]
If \( n \leq 1 + pd \), or equivalently \( \kappa_f > n^2d^{-2} \), then \( m = 1 \)
and
\[
C_1'(m, \varepsilon) = \mathcal{O}(\frac{\sqrt{\kappa(M)d}}{\ln(1 + \frac{1}{\sqrt{\kappa_f^M}})} \ln \frac{1}{\varepsilon}),
\]
therefore
\[
\min_{m \geq 1} \frac{C_1'(m, \varepsilon)}{C_1'(1, \varepsilon)} \leq \frac{C_1'(1, \varepsilon)}{C_1'(m^*, \varepsilon)} = \mathcal{O}(\frac{\sqrt{\kappa(M)d}}{\kappa_f \sqrt{n} \ln(1 + \frac{1}{\sqrt{\kappa_f^M}})}).
\]
Since \( \kappa(M) \approx \kappa_f \gg \kappa_f^M \), this ratio becomes \( \mathcal{O}(\frac{d}{\sqrt{\kappa_f^M}}) \).

Proof of Theorem 4. The proof of Theorem 4 is similar and is omitted.