A. Proof of Lemma 1

In this section, we prove the results on the error generated when solving the subproblem (3.2) inexactly 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 $\gamma = \eta\frac{\lambda_{\min}(M)}{\lambda^2_{\max}(M)}$, and is repeat it p times, where $p \geq 1$. Then, $w_{t+1} = w^p_{t+1}$ is an approximate solution to (3.2) that satisfies*

$$\mathbf{0} \in \partial\psi(w_{t+1}) + \frac{1}{\eta}M(w_{t+1} - w_t) + \tilde{\nabla}_t + M\varepsilon^p_{t+1}, \tag{A.1}$$

$$\|\varepsilon^p_{t+1}\|_M \leq \frac{c(p)}{\eta}\|w_{t+1} - w_t\|_M, \tag{A.2}$$

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

$$\underset{y\in\mathbb{R}^d}{\text{minimize}}\, h_1(y) + h_2(y), \tag{A.3}$$

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

$$\begin{aligned} w^0_{t+1} &= w_t, \\ w^{i+1}_{t+1} &= \mathbf{prox}_{\gamma h_1}\big(w^i_{t+1} - \gamma\nabla h_2(w^i_{t+1})\big), \\ w_{t+1} &= w^p_{t+1}, \end{aligned}$$

where $i = 0, 1, ..., p-1$. From the definition of $\mathbf{prox}_{\gamma h_1}$, we have

$$\mathbf{0} \in \partial h_1(w^p_{t+1}) + \nabla h_2(w^{p-1}_{t+1}) + \frac{1}{\gamma}(w^p_{t+1} - w^{p-1}_{t+1}).$$

Compare this with (A.1) gives

$$M\varepsilon^p_{t+1} = \frac{1}{\gamma}(w^p_{t+1} - w^{p-1}_{t+1}) + \nabla h_2(w^{p-1}_{t+1}) - \nabla h_2(w^p_{t+1}).$$

To bound the right hand side, let $w^\star_{t+1}$ 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 α-strongly convex and β-Lipschitz differentiable. Consequently, Prop. 26.16(ii) of (Bauschke et al., 2017) gives

$$\|w^i_{t+1} - w^\star_{t+1}\| \leq \tau^i\|w^0_{t+1} - w^\star_{t+1}\|, \quad \forall i = 0, 1, ..., p,$$

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

Let $a_i = \|w^i_{t+1} - w^\star_{t+1}\|$. Then, $a_i \leq \tau^i a_0$. We can derive

$$\begin{aligned} \|M\varepsilon^p_{t+1}\| &\leq (\frac{1}{\gamma} + \beta)\|w^p_{t+1} - w^{p-1}_{t+1}\| \\ &\leq (\frac{1}{\gamma} + \beta)(a_p + a_{p-1}) \leq (\frac{1}{\gamma} + \beta)(\tau^p + \tau^{p-1})a_0. \end{aligned}$$

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

$$\|M\varepsilon^p_{t+1}\| \leq b(p)\|w_{t+1} - w_t\|, \tag{A.4}$$

where

$$b(p) = (\frac{1}{\gamma} + \frac{\lambda_{\max}(M)}{\eta})\frac{\tau^p + \tau^{p-1}}{1 - \tau^p}. \tag{A.5}$$

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 ε^p_{t+1} and $w_{t+1} - w_t$ be decomposed by

$$\begin{aligned} \varepsilon^p_{t+1} &= \sum_{i=1}^d \alpha_i v_i, \\ w_{t+1} - w_t &= \sum_{i=1}^d \beta_i v_i. \end{aligned}$$

then

$$\begin{aligned} \|\varepsilon^p_{t+1}\|_M &= \sqrt{\sum_{i=1}^d \lambda_i\alpha_i^2} \leq \sqrt{\frac{1}{\lambda_{\min}(M)}\sum_{i=1}^d \lambda_i^2\alpha_i^2} \\ &= \sqrt{\frac{1}{\lambda_{\min}(M)}}\|M\varepsilon^p_{t+1}\|, \\ \|w_{t+1} - w_t\| &= \sqrt{\sum_{i=1}^d \beta_i^2} \leq \sqrt{\frac{1}{\lambda_{\min}(M)}\sum_{i=1}^d \lambda_i\beta_i^2} \\ &= \sqrt{\frac{1}{\lambda_{\min}(M)}}\|w_{t+1} - w_t\|_M. \end{aligned}$$

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

$$\|\varepsilon^p_{t+1}\|_M \leq c(p)\|w_{t+1} - w_t\|_M, \tag{A.6}$$

where

$$c(p) = \frac{1}{\lambda_{\min}(M)}b(p) = \frac{\frac{1}{\gamma} + \frac{\lambda_{\max}(M)}{\eta}}{\lambda_{\min}(M)}\frac{\tau^p + \tau^{p-1}}{1 - \tau^p}.$$

$\square$

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

$$\mathbf{0} \in \partial\psi(w_{t+1}) + \frac{1}{\eta}M(w_{t+1}-w_t) + \tilde{\nabla}_t + M\varepsilon^p_{t+1}, \tag{A.7}$$

$$\|\varepsilon^p_{t+1}\|_M \leq \frac{c(p)}{\eta}\|w_{t+1}-w_t\|_M, \tag{A.8}$$

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

$$\mathbf{0} \in \partial\psi(w_{t+1}^{(r-1,p_0)}) + \frac{1}{\gamma}(w_{t+1}^{(r-1,p_0)} - u_{t+1}^{(r-1,p_0)}) + \nabla h_2(u_{t+1}^{(r-1,p_0)}),$$

compare this with (A.7), we have

$$\begin{aligned} M\varepsilon^p_{t+1} =& \frac{1}{\gamma}(w_{t+1}^{(r-1,p_0)} - u_{t+1}^{(r-1,p_0)}) + \nabla h_2(u_{t+1}^{(r-1,p_0)}) \\ &- \frac{1}{\eta}M(w_{t+1}-w_t) - \tilde{\nabla}_t \\ =& \frac{1}{\gamma}(w_{t+1}^{(r-1,p_0)} - u_{t+1}^{(r-1,p_0)}) \\ &+ \frac{1}{\eta}M(u_{t+1}^{(r-1,p_0)} - w_{t+1}) \end{aligned}$$

where

$$u_{t+1}^{(r-1,p_0)} = w_{t+1}^{(r-1,p_0-1)} + \frac{\theta_{p_0-2}-1}{\theta_{p_0-1}}(w_{t+1}^{(r-1,p_0-1)} - w_{t+1}^{(r-1,p_0-2)}).$$

As a result,

$$\begin{aligned} \|M\varepsilon^p_{t+1}\| \leq& \|\frac{1}{\gamma}(w_{t+1}^{(r-1,p_0)} - u_{t+1}^{(r-1,p_0)})\| \qquad \text{(A.9)} \\ &+ \|\frac{1}{\eta}M(u_{t+1}^{(r-1,p_0)} - w_{t+1})\| \\ \leq& \|\frac{1}{\gamma}(w_{t+1}^{(r-1,p_0)} - w_{t+1}^{(r-1,p_0-1)})\| \\ &+ \frac{1}{\gamma}\|\frac{\theta_{p_0-2}-1}{\theta_{p_0-1}}(w_{t+1}^{(r-1,p_0-1)} - w_{t+1}^{(r-1,p_0-2)})\| \\ &+ \|\frac{1}{\eta}M(w_{t+1}^{(r-1,p_0-1)} - w_{t+1})\| \\ &+ \|\frac{1}{\eta}\frac{\theta_{p_0-2}-1}{\theta_{p_0-1}}M(w_{t+1}^{(r-1,p_0-1)} - w_{t+1}^{(r-1,p_0-2)})\|, \qquad \text{(A.10)} \end{aligned}$$

Let the solution of (3.2) be $w^\star_{t+1}$. 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

$$\Psi(w_{t+1}^{(i,j)}) - \Psi(w^\star_{t+1}) \leq \frac{2\lambda_{\max}(M)\|w_{t+1}^{(i,0)} - w^\star_{t+1}\|^2}{\eta j^2}.$$

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

$$\Psi(w_{t+1}^{(i,j)}) - \Psi(w^\star_{t+1}) \geq \frac{\lambda_{\min}(M)}{2\eta}\|w_{t+1}^{(i,j)} - w^\star_{t+1}\|^2.$$

Therefore,

$$\|w_{t+1}^{(i,j)} - w^\star_{t+1}\| \leq \sqrt{\frac{4\kappa(M)}{j^2}}\|w_{t+1}^{(i,0)} - w^\star_{t+1}\|. \tag{A.11}$$

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

$$\begin{aligned} &\|\frac{1}{\gamma}(w_{t+1}^{(r-1,p_0)} - w_{t+1}^{(r-1,p_0-1)})\| \\ \leq& \frac{1}{\gamma}\|w_{t+1}^{(r-1,p_0)} - w^\star_{t+1}\| \\ &+ \frac{1}{\gamma}\|w_{t+1}^{(r-1,p_0-1)} - w^\star_{t+1}\| \\ \leq& \frac{1}{\gamma}(\frac{4\kappa(M)}{p_0^2})^{\frac{r}{2}}\|w_{t+1}^{(0,0)} - w^\star_{t+1}\| \\ &+ \frac{1}{\gamma}(\frac{4\kappa(M)}{p_0^2})^{\frac{r-1}{2}}(\frac{4\kappa(M)}{(p_0-1)^2})^{\frac{1}{2}}\|w_{t+1}^{(0,0)} - w^\star_{t+1}\|. \end{aligned}$$

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

$$\begin{aligned} &\frac{1}{\gamma}\|\frac{\theta_{p_0-2}-1}{\theta_{p_0-1}}(w_{t+1}^{(r-1,p_0-1)} - w_{t+1}^{(r-1,p_0-2)})\| \\ \leq& \frac{1}{\gamma}(\frac{4\kappa(M)}{p_0^2})^{\frac{r-1}{2}}(\frac{4\kappa(M)}{(p_0-1)^2})^{\frac{1}{2}}\|w_{t+1}^{(0,0)} - w^\star_{t+1}\| \\ &+ \frac{1}{\gamma}(\frac{4\kappa(M)}{p_0^2})^{\frac{r-1}{2}}(\frac{4\kappa(M)}{(p_0-2)^2})^{\frac{1}{2}}\|w_{t+1}^{(0,0)} - w^\star_{t+1}\|, \end{aligned}$$

$$\begin{aligned} &\|\frac{1}{\eta}M(w_{t+1}^{(r-1,p_0-1)} - w_{t+1})\| \\ \leq& \frac{\lambda_{\max}(M)}{\eta}(\frac{4\kappa(M)}{p_0^2})^{\frac{r-1}{2}}(\frac{4\kappa(M)}{(p_0-1)^2})^{\frac{1}{2}}\|w_{t+1}^{(0,0)} - w^\star_{t+1}\|, \\ &+ \frac{\lambda_{\max}(M)}{\eta}(\frac{4\kappa(M)}{p_0^2})^{\frac{r}{2}}\|w_{t+1}^{(0,0)} - w^\star_{t+1}\|, \end{aligned}$$

$$\begin{aligned} &\|\frac{1}{\eta}\frac{\theta_{p_0-2}-1}{\theta_{p_0-1}}M(w_{t+1}^{(r-1,p_0-1)} - w_{t+1}^{(r-1,p_0-2)})\| \\ \leq& \frac{\lambda_{\max}(M)}{\eta}(\frac{4\kappa(M)}{p_0^2})^{\frac{r-1}{2}}(\frac{4\kappa(M)}{(p_0-1)^2})^{\frac{1}{2}}\|w_{t+1}^{(0,0)} - w^\star_{t+1}\| \\ &+ \frac{\lambda_{\max}(M)}{\eta}(\frac{4\kappa(M)}{p_0^2})^{\frac{r-1}{2}}(\frac{4\kappa(M)}{(p_0-2)^2})^{\frac{1}{2}}\|w_{t+1}^{(0,0)} - w^\star_{t+1}\|, \end{aligned}$$

where in the first and third estimate we have used $\frac{\theta_{p_0-2}-1}{\theta_{p_0-1}} \leq$

$\frac{\theta_{p_0-2}}{\theta_{p_0-1}} < 1$. On the other hand, we have

$$\begin{aligned}\|w_{t+1} - w_t\| &= \|w_{t+1}^{(r-1,p_0)} - w_{t+1}^{(0,0)}\| \\ &\geq \|w_{t+1}^{(0,0)} - w_{t+1}^{\star}\| - \|w_{t+1}^{(r-1,p_0)} - w_{t+1}^{\star}\| \\ &\geq (1 - (\frac{4\kappa(M)}{p_0^2})^{\frac{r}{2}})\|w_{t+1}^{(0,0)} - w_{t+1}^{\star}\|.\end{aligned}$$

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 $\tau = (\frac{4\kappa(M)}{p_0^2})^{\frac{1}{2p_0}}$ yields

$$\|M\varepsilon_{t+1}^p\| \leq 2\frac{\lambda_{\max}(M)}{\eta}\frac{b(p)}{1-\tau^p}\|w_{t+1} - w_t\|,$$

where

$$\begin{aligned}b(p) =& \tau^{p-p_0}\big((\frac{4\kappa(M)}{(p_0-1)^2})^{\frac{1}{2}} + (\frac{4\kappa(M)}{(p_0-2)^2})^{\frac{1}{2}}\big) \\ &+ \tau^p + \tau^{p-p_0}(\frac{4\kappa(M)}{(p_0-1)^2})^{\frac{1}{2}}.\end{aligned} \tag{A.12}$$

Similar 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 $\tau = (\frac{4\kappa(M)}{p_0^2})^{\frac{1}{2p_0}}$ is minimized, a simple calculation yields

$$p_0^\star = 2e\sqrt{\kappa(M)}.$$

In order for p_0 to be an integer, we can take

$$p_0 = \lceil 2e\sqrt{\kappa(M)}\rceil,$$

then

$$\begin{aligned}\tau &= (\frac{4\kappa(M)}{p_0^2})^{\frac{1}{2p_0}} \leq (\frac{1}{e^2})^{\frac{1}{2\lceil 2e\sqrt{\kappa(M)}\rceil}} \leq (\frac{1}{e^2})^{\frac{1}{2(2e\sqrt{\kappa(M)}+1)}} \\ &= \exp(-\frac{1}{2e\sqrt{\kappa(M)}+1}).\end{aligned}$$

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

$$\tau^{-p_0}(\frac{4\kappa(M)}{p_0-1})^{\frac{1}{2}} = (\frac{p_0}{p_0-1})^{\frac{1}{p_0}},$$

and

$$p_0 = \lceil 2e\sqrt{\kappa(M)}\rceil \geq \lceil 2e\rceil = 6.$$

On the other hand, a simple calculation shows that $(\frac{p_0}{p_0-1})^{\frac{1}{p_0}}$ is decreasing in p_0, therefore

$$\tau^{-p_0}(\frac{4\kappa(M)}{p_0-1})^{\frac{1}{2}} \leq (\frac{6}{5})^{\frac{1}{6}} < 2,$$

Similarly, one can show that

$$\tau^{-p_0}(\frac{4\kappa(M)}{p_0-2})^{\frac{1}{2}} \leq (\frac{6}{4})^{\frac{1}{6}} < 2.$$

Combining these two inequalities with (B.2) yields

$$b(p) \leq 7\tau^p.$$

□

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*

$$\begin{aligned}&\langle\tilde{\nabla}_t, w_t - u\rangle + \psi(w_{t+1}) - \psi(u) \\ &\leq \langle\tilde{\nabla}_t, w_t - w_{t+1}\rangle + \frac{\|u - w_t\|_M^2}{2\eta} \\ &- \frac{1}{2\eta}\|u - w_{t+1}\|_M^2 - \frac{1}{2\eta}\|w_{t+1} - w_t\|_M^2 \\ &+ \langle M\varepsilon_{t+1}^p, u - w_{t+1}\rangle.\end{aligned}$$

Proof. First, let us rewrite the left hand side as

$$\begin{aligned}&\langle\tilde{\nabla}_t, w_t - u\rangle + \psi(w_{t+1}) - \psi(u) \\ &= \langle\tilde{\nabla}_t, w_t - w_{t+1}\rangle + \langle\tilde{\nabla}_t, w_{t+1} - u\rangle + \psi(w_{t+1}) - \psi(u).\end{aligned}$$

By (4.1) and the definition of subdifferential we have

$$\psi(u) \geq \psi(w_{t+1}) - \langle\tilde{\nabla}_t + \frac{1}{\eta}M(w_{t+1} - w_t) + M\varepsilon_{t+1}^p, u - w_{t+1}\rangle.$$

Combining these two gives

$$\begin{aligned}&\langle\tilde{\nabla}_t, w_t - u\rangle + \psi(w_{t+1}) - \psi(u) \\ \leq&\langle\tilde{\nabla}_t, w_t - w_{t+1}\rangle \\ &+ \langle\frac{1}{\eta}M(w_{t+1} - w_t) + M\varepsilon_{t+1}^p, u - w_{t+1}\rangle \\ =&\langle\tilde{\nabla}_t, w_t - w_{t+1}\rangle + \frac{\|u - w_t\|_M^2}{2\eta} \\ &- \frac{1}{2\eta}\|u - w_{t+1}\|_M^2 - \frac{1}{2\eta}\|w_{t+1} - w_t\|_M^2 \\ &+ \langle M\varepsilon_{t+1}^p, u - w_{t+1}\rangle,\end{aligned}$$

where in the last equality we have applied

$$\langle a-b, c-a\rangle_M = -\frac{1}{2}\|a-b\|_M^2 - \frac{1}{2}\|a-c\|_M^2 + \frac{1}{2}\|b-c\|_M.$$

□

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 $a > 0$, $\eta \leq \frac{1-2c(p)a}{2L_f^M}$, and $u \in \mathbb{R}^{\mathrm{d}}$ we have*

$$
\begin{aligned}
&\mathbb{E}[F(w_{t+1}) - F(u)] \\
\leq &\mathbb{E}[\eta\|\tilde{\nabla}_t - \nabla f(w_t)\|^2_{M^{-1}} + \frac{1-\eta\sigma_f^M}{2\eta}\|u - w_t\|^2_M \\
&- (\frac{1}{2\eta} - \frac{c(p)}{2\eta a})\|u - w_{t+1}\|^2_M].
\end{aligned}
$$

Proof. We have

$$
\begin{aligned}
&\mathbb{E}[F(w_{t+1}) - F(u)] \\
&= \mathbb{E}[f(w_{t+1}) - f(u) + \psi(w_{t+1}) - \psi(u)] \\
\leq &\mathbb{E}[f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t\rangle \\
&+ \frac{L_f^M}{2}\|w_t - w_{t+1}\|^2_M - f(u) + \psi(w_{t+1}) - \psi(u)] \\
&\mathbb{E}[\langle \nabla f(w_t), w_t - u\rangle - \frac{\sigma_f^M}{2}\|u - w_t\|^2_M \\
&+ \langle \nabla f(w_t), w_{t+1} - w_t\rangle + \frac{L_f^M}{2}\|w_t - w_{t+1}\|^2_M \\
&+ \psi(w_{t+1}) - \psi(u)] \\
= &\mathbb{E}[\langle \tilde{\nabla}_t, w_t - u\rangle - \frac{\sigma_f^M}{2}\|u - w_t\|^2_M \qquad \text{(B.1)} \\
&+ \langle \nabla f(w_t), w_{t+1} - w_t\rangle \\
&+ \frac{L_f^M}{2}\|w_t - w_{t+1}\|^2_M + \psi(w_{t+1}) - \psi(u)], \qquad \text{(B.2)}
\end{aligned}
$$

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 $\mathbb{E}[\tilde{\nabla}_t] = \nabla f(w_t)$.

On the other hand, recall that Lemma 4 gives

$$
\begin{aligned}
&\langle \tilde{\nabla}_t, w_t - u\rangle + \psi(w_{t+1}) - \psi(u) \\
&\langle \tilde{\nabla}_t, w_t - w_{t+1}\rangle + \frac{\|u - w_t\|^2_M}{2\eta} \\
&- \frac{1}{2\eta}\|u - w_{t+1}\|^2_M - \frac{1}{2\eta}\|w_{t+1} - w_t\|^2_M \\
&+ \langle M\varepsilon^p_{t+1}, u - w_{t+1}\rangle,
\end{aligned}
$$

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

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

from Lemma 3 and Lemma 1 we know that

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

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

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

Applying this to Lemma 4 yields

$$
\begin{aligned}
&\langle \tilde{\nabla}_t, w_t - u\rangle + \psi(w_{t+1}) - \psi(u) \\
\leq &\langle \tilde{\nabla}_t, w_t - w_{t+1}\rangle + \frac{\|u - w_t\|^2_M}{2\eta} \\
&- \frac{1}{2\eta}\|u - w_{t+1}\|^2_M - \frac{1}{2\eta}\|w_{t+1} - w_t\|^2_M \\
&+ \langle M\varepsilon^p_{t+1}, u - w_{t+1}\rangle \\
&\langle \tilde{\nabla}_t, w_t - w_{t+1}\rangle + \frac{\|u - w_t\|^2_M}{2\eta} \\
&- (\frac{1}{2\eta} - \frac{c(p)}{2a\eta})\|u - w_{t+1}\|^2_M \\
&- (\frac{1}{2\eta} - \frac{c(p)a}{2\eta})\|w_{t+1} - w_t\|^2_M
\end{aligned}
$$

Applying this to (B.2), we arrive at

$$
\begin{aligned}
&\mathbb{E}[F(w_{t+1}) - F(u)] \\
\leq &\mathbb{E}[\langle \tilde{\nabla}_t - \nabla f(w_t), w_t - w_{t+1}\rangle \\
&- \frac{1 - c(p)a - \eta L_f^M}{2\eta}\|w_t - w_{t+1}\|^2_M \\
&+ \frac{1-\eta\sigma_f^M}{2\eta}\|u - w_t\|^2_M - (\frac{1}{2\eta} - \frac{c(p)}{2a\eta})\|u - w_{t+1}\|^2_M] \\
&\mathbb{E}[\frac{\eta}{2(1 - c(p)a - \eta L_f^M)}\|\tilde{\nabla}_t - \nabla f(w_t)\|^2_{M^{-1}} \\
&+ \frac{1-\eta\sigma_f^M}{2\eta}\|u - w_t\|^2_M - (\frac{1}{2\eta} - \frac{c(p)}{2a\eta})\|u - w_{t+1}\|^2_M],
\end{aligned}
$$

where in the second inequality we have applied

$$
\begin{aligned}
&\langle u_1, u_2\rangle = \langle M^{-\frac{1}{2}}u_1, M^{\frac{1}{2}}u_2\rangle \leq \|u_1\|_{M^{-1}}\|u_2\|_M \\
&\leq \frac{1}{2b}\|u_1\|^2_{M_1^{-1}} + \frac{b}{2}\|u_2\|^2_{M^{\frac{1}{2}}} \quad \text{for any } b > 0.
\end{aligned}
$$

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

□

Lemma 6. *Under Assumption 1, we have*

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

Proof. We have

$$
\begin{aligned}
&\mathbb{E}[\|\tilde{\nabla}_t - \nabla f(w_t)\|^2_{M^{-1}}] \\
=&\mathbb{E}[\|\nabla f(w_0) + \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_0) - \nabla f(w_t)\|^2_{M^{-1}}] \\
=&\mathbb{E}[\|\big(\nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_0)\big) - \big(\nabla f(w_t) - \nabla f(w_0)\big)\|^2_{M^{-1}}] \\
\leq&\mathbb{E}[\|\nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_0)\|^2_{M^{-1}}] \\
\leq&(L_f^M)^2\|w_t - w_0\|^2_M,
\end{aligned}
$$

where in the first inequality, we have applied $\mathbb{E}[\|\xi - \mathbb{E}\,\xi\|^2] = \mathbb{E}[\|\xi\|^2 - \|\mathbb{E}\,\xi\|^2$ with $\xi = M^{-\frac{1}{2}}\big(\nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_0)\big)$, 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 $N \sim$**Geom**(p), then*

1. $\mathbb{E}_N\left[C_N - C_{N+1}\right] = \frac{p}{1-p}\mathbb{E}_N\left[C_0 - C_N\right]$*, and*
2. $\mathbb{E}_N\left[C_N\right] = (1-p)\mathbb{E}\left[C_{N+1}\right] + pC_0$.

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

$$
\begin{aligned}
&\mathbb{E}[F(w_{D+1}) - F(u)] \\
\leq&\mathbb{E}[-\frac{1}{4m\eta}\|w_{D+1} - w_0\|^2_M + \frac{\langle w_0 - w_{D+1}, w_0 - u\rangle_M}{m\eta} \\
&-(\frac{\sigma_f^M}{4} - \frac{c(p)}{2a\eta})\|w_{D+1} - u\|^2_M].
\end{aligned}
$$

Proof. By Lemmas 5 and 6, we know that

$$
\begin{aligned}
&\mathbb{E}[F(w_{t+1}) - F(u)] \\
&\mathbb{E}[\eta(L_f^M)^2\|w_0 - w_t\|^2_M + \frac{1-\eta\sigma_f^M}{2\eta}\|u - w_t\|^2_M \\
&-(\frac{1}{2\eta} - \frac{c(p)}{2\eta a})\|u - w_{t+1}\|^2_M].
\end{aligned}
$$

Let $D \sim$ **Geom**$(\frac{1}{m})$ as in Algorithm 1 and take $t = D$, then

$$
\begin{aligned}
&\mathbb{E}[F(w_{D+1}) - F(u)] \\
\leq&\mathbb{E}[\eta(L_f^M)^2\|w_0 - w_D\|^2_M + \frac{1}{2\eta}\|u - w_D\|^2_M \\
&-\frac{1}{2\eta}\|u - w_{D+1}\|^2_M - \frac{\sigma_f^M}{2}\|u - w_D\|^2_M \\
&+\frac{c(p)}{2\eta a}\|u - w_{D+1}\|^2_M] \\
=&\mathbb{E}[\eta(L_f^M)^2\|w_D - w_0\|^2_M + \frac{\|u - w_0\|^2_M - \|u - w_D\|^2_M}{2(m-1)\eta} \\
&-\frac{\sigma_f^M}{2}\|u - w_D\|^2_M + \frac{c(p)}{2a\eta}\|u - w_{D+1}\|^2_M] \\
=&\mathbb{E}[\frac{m-1}{m}\eta(L_f^M)^2\|w_{D+1} - w_0\|^2_M \\
&+\frac{\|u - w_0\|^2_M - \|u - w_{D+1}\|^2_M}{2m\eta}] \\
&-\frac{\sigma_f^M}{2m}\|u - w_0\|^2_M - \frac{\sigma_f^M(m-1)}{2m}\|u - w_{D+1}\|^2_M \\
&+\frac{c(p)}{2a\eta}\|u - w_{D+1}\|^2_M] \\
\leq&\mathbb{E}[\eta(L_f^M)^2\|w_{D+1} - w_0\|^2_M + \frac{\|u - w_0\|^2_M - \|u - w_{D+1}\|^2_M}{2m\eta} \\
&-\frac{\sigma_f^M}{4}\|u - w_{D+1}\|^2_M + \frac{c(p)}{2a\eta}\|u - w_{D+1}\|^2_M] \\
\leq&\mathbb{E}[-\frac{1}{4m\eta}\|w_0 - w_{D+1}\|^2_M \\
&+\frac{\|u - w_0\|^2_M - \|u - w_{D+1}\|^2_M + \|w_0 - w_{D+1}\|^2_M}{2m\eta} \\
&-\frac{\sigma_f^M}{4}\|w_{D+1} - u\|^2_M + \frac{c(p)}{2a\eta}\|u - w_{D+1}\|^2_M] \\
=&\mathbb{E}[-\frac{1}{4m\eta}\|w_{D+1} - w_0\|^2_M + \frac{\langle w_0 - w_{D+1}, w_0 - u\rangle_M}{m\eta} \\
&-(\frac{\sigma_f^M}{4} - \frac{c(p)}{2a\eta})\|w_{D+1} - u\|^2_M],
\end{aligned}
$$

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 $C_N = \|w_d - w_0\|^2_M$, item 2 with $C_N = \|u - w_0\|^2_M - \|u - w_N\|^2_M$, 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 $\eta \leq \frac{1}{2\sqrt{m}L_f^M}$.

□

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 assume $x^\star = \arg\min_{x\in\mathbb{R}^d} F(x) = \mathbf{0}$ and $F(x^*) = 0$.

According to Lemma 8, for any $u \in \mathbb{R}^d$, and $\eta \le \min\{\frac{1-2c(p)a}{2L_f^M}, \frac{1}{2\sqrt{m}L_f^M}\}$ we have

$$\begin{aligned}&\mathbb{E}[F(x^{j+1}) - F(u)]\\ \le&\mathbb{E}[-\frac{1}{4m\eta}\|x^{j+1}-x^j\|_M^2\\ &+\frac{\langle x^j - x^{j+1}, x^j - u\rangle_M}{m\eta} - (\frac{\sigma_f^M}{4} - \frac{c(p)}{2a\eta})\|x^{j+1}-u\|_M^2],\end{aligned}$$

or equivalently,

$$\begin{aligned}&\mathbb{E}[F(x^{j+1}) - F(u)]\\ \le&\mathbb{E}[\frac{1}{4m\eta}\|x^{j+1}-x^j\|_M^2 + \frac{1}{2m\eta}\|x^j - u\|_M^2\\ &-\frac{1}{2m\eta}\|x^{j+1}-u\|_M^2 - (\frac{\sigma_f^M}{4} - \frac{c(p)}{2a\eta})\|x^{j+1}-u\|_M^2].\end{aligned}$$

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

Setting $u = x^* = 0$ and $u = x^j$ yields the following two inequalities:

$$\begin{aligned}F(x^{j+1}) \le& \frac{1}{4m\eta}(\|x^{j+1}-x^j\|_M^2 + 2\|x^j\|_M^2)\\ &-\frac{1}{2m\eta}\big(1+\frac{1}{2}m\eta(\sigma_f^M - \frac{2c(p)}{a\eta})\big)\|x^{j+1}\|_M^2,\end{aligned} \tag{B.3}$$

$$F(x^{j+1}) - F(x^j) \tag{B.4}$$

$$\le -\frac{1}{4m\eta}\big(1+m\eta(\sigma_f^M - \frac{2c(p)}{a\eta})\big)\|x^{j+1}-x^j\|_M^2. \tag{B.5}$$

Define $\tau = \frac{1}{2}m\eta(\sigma_f^M - \frac{2c(p)}{a\eta})$, multiply $(1+2\tau)$ to (B.3), then add it to (B.5) yields

$$\begin{aligned}&2(1+\tau)F(x^{j+1}) - F(x^j)\\ \le&\frac{1}{2m\eta}(1+2\tau)\big(\|x^j\|_M^2 - (1+\tau)\|x^{j+1}\|_M^2\big).\end{aligned}$$

Multiplying both sides by $(1+\tau)^j$ gives

$$\begin{aligned}&2(1+\tau)^{j+1}F(x^{j+1}) - (1+\tau)^j F(x^j)\\ \le&\frac{1}{2m\eta}(1+2\tau)\big((1+\tau)^j\|x^j\|_M^2 - (1+\tau)^{j+1}\|x^{j+1}\|_M^2\big).\end{aligned}$$

Summing over $j = 0, 1, ..., k-1$, we have

$$\begin{aligned}&(1+\tau)^k F(x^k) + \sum_{j=0}^{k-1}(1+\tau)^j F(x^j) - F(x^0)\\ \le&\frac{1}{2m\eta}(1+2\tau)(\|x^0\|_M^2 - (1+\tau)^k\|x^k\|_M^2).\end{aligned}$$

Since $F(x^j) \ge 0$, we have

$$F(x^k)(1+\tau)^k \le F(x^0) + \frac{1}{2m\eta}(1+2\tau)\|x^0\|^2.$$

By the strong convexity of F, we have $F(x^0) \ge \frac{\sigma_f^M}{2}\|x^0\|_M^2$, therefore

$$F(x^k)(1+\tau)^k \le F(x^0)(2+\frac{1}{2\tau}). \tag{B.6}$$

Finally, recall that $a > 0$ can be chosen arbitrarily, so we can take

$$a = \frac{4c(p)}{\eta\sigma_f^M},$$

and

$$\begin{aligned}\eta \le& \min\{\frac{1-2c(p)a}{2L_f^M}, \frac{1}{2\sqrt{m}L_f^M}\}\\ =&\min\{\frac{1-\frac{8c^2(p)}{\eta\sigma_f^M}}{2L_f^M}, \frac{1}{2\sqrt{m}L_f^M}\},\end{aligned} \tag{B.7}$$

$$\tau = \frac{1}{2}m\eta(\sigma_f^M - \frac{2c(p)}{a\eta}) = \frac{1}{4}m\eta\sigma_f^M.$$

In order for the choice of η in (B.7) to be possible, we need

$$2L_f^M\eta^2 - \eta + 8\frac{c^2(p)}{\sigma_f^M} \le 0 \tag{B.8}$$

to have one solution at least, which requires

$$64\kappa_f^M c^2(p) \le 1,$$

under which $\eta = \frac{1}{4L_f^M}$ satisfy (B.8). As a result, $m \ge 4$ makes (B.7) into

$$\eta \le \frac{1}{2\sqrt{m}L_f^M},$$

and the desired convergence result follows from (B.6). $\square$

C. Proof of Lemma 2

Proof. From Lemma 1, we know that

$$c(p) = 14\kappa(M)\frac{\tau^p}{1-\tau^p},$$

where

$$\tau \le \exp(-\frac{1}{2e\sqrt{\kappa(M)}+1}).$$

Therefore, in order for $64\kappa_f^M c^2(p) \le 1$, we need

$$\kappa_f^M\kappa^2(M)(\frac{\tau^p}{1-\tau^p})^2 \le \frac{1}{64\times 14^2} = c_1,$$

which is equivalent to

$$\tau^p \leq \frac{c_1}{\sqrt{\kappa_f^M \kappa(M)} + \sqrt{c_1}}.$$

Thus, it suffices to require that

$$[\exp(-\frac{1}{2e\sqrt{\kappa(M)}+1})]^p \leq \frac{c}{\sqrt{\kappa_f^M \kappa(M)} + \sqrt{c_1}},$$

which gives

$$p \geq (2e\sqrt{\kappa(M)}+1)\ln\frac{\sqrt{\kappa_f^M \kappa(M)} + \sqrt{c_1}}{c_1}.$$

□

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 σ by $\sigma_f^M - \frac{2c(p)}{a\eta}$.
3. By setting
$$a = \frac{4c(p)}{\eta\sigma_f^M},$$
and
$$64\kappa_f^M c^2(p) \leq 1$$
as in the proof of Theorem 1, the τ in Theorem 4.3 of (Allen-Zhu, 2018) becomes $\frac{1}{2}m\eta\sigma_f^M$, 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

$$C_1(m,\varepsilon) = \mathcal{O}(\frac{n+m}{\ln(1+\frac{1}{4}m\eta\sigma_f)}\ln\frac{1}{\varepsilon}).$$

Taking the largest possible step size $\eta = \frac{1}{2\sqrt{m}L_f}$ as in Theorem 1, we have

$$C_1(m,\varepsilon) = \mathcal{O}(\frac{n+m}{\ln(1+\frac{\sqrt{m}}{8\kappa_f})}\ln\frac{1}{\varepsilon}).$$

Let us first find the optimal $m = m^\star$ for SVRG, let

$$g(m) = \frac{n+m}{\ln(1+\frac{\sqrt{m}}{8\kappa_f})},$$

then

$$g'(m) = \frac{\ln(1+\frac{\sqrt{m}}{8\kappa_f}) - \frac{\frac{\sqrt{m}}{8\kappa_f}}{1+\frac{\sqrt{m}}{8\kappa_f}}\frac{n+m}{2m}}{\ln^2(1+z)}.$$

Taking derivative to the numerator gives

$$\begin{aligned}&[\ln(1+\frac{\sqrt{m}}{8\kappa_f}) - \frac{\frac{\sqrt{m}}{8\kappa_f}}{1+\frac{\sqrt{m}}{8\kappa_f}}\frac{n+m}{2m}]' \\ &= (n+m)\frac{\frac{1}{32\kappa_f}m^{-\frac{3}{2}} + 2\frac{m^{-1}}{(16\kappa_f)^2}}{(1+\frac{\sqrt{m}}{8\kappa_f})^2} > 0,\end{aligned}$$

Therefore, $m^\star$ is given by $g'(m) = 0$. Let $z = \frac{\sqrt{m}}{8\kappa_f} > 0$, 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, $m^\star < n$.

Let $m = n^s$ where $0 < s < 1$, we would like to have $g'(n^s) < 0$, i,e.,

$$\frac{\ln(1+z)}{\frac{z}{1+z}} < \frac{1+n^{1-s}}{2}.$$

so that $m^\star \in (n^s, n)$.

Since $\kappa_f > n^{\frac{1}{2}}$, we have $z = \frac{\sqrt{m}}{8\kappa_f} < \frac{1}{8}$, on the other hand, we have

$$[\frac{\ln(1+z)}{\frac{z}{1+z}} < \frac{1+n^{1-s}}{2}]'_z > 0.$$

Therefore, it suffices to have

$$n^{1-s} > 18\ln\frac{9}{8} - 1 := c_0 > 1.$$

As a result, we have $m^\star \in (\frac{n}{c_0}, n)$, and

$$\begin{aligned}C_1(m^\star,\varepsilon) &= \mathcal{O}(\frac{n+m^\star}{\ln(1+\frac{\sqrt{m^\star}}{8\kappa_f})}\ln\frac{1}{\varepsilon}) \\ &= \mathcal{O}(\frac{n}{\frac{\sqrt{n}}{8\kappa_f}}\ln\frac{1}{\varepsilon}) = \mathcal{O}(\kappa_f\sqrt{n}\ln\frac{1}{\varepsilon}),\end{aligned}$$

where in the second equality we have used $\kappa_f > n^{\frac{1}{2}}$.

For our iPreSVRG in Algorithm 1, we have

$$C_1'(m,\varepsilon) = \mathcal{O}(\frac{n+(1+pd)m}{\ln(1+\frac{1}{4}m\eta\sigma^M)}\ln\frac{1}{\varepsilon}),$$

thanks to Lemma 2, p can be chosen as

$$p = \mathcal{O}(\sqrt{\kappa(M)} \ln \big(\sqrt{\kappa_f^M} \kappa(M)\big),$$

furthermore, we can take $\eta = \frac{1}{2\sqrt{m}L_f}$ due to Theorem 1.

Under these settings, we have

$$C_1'(m, \varepsilon) = \mathcal{O}(\frac{n + (1+pd)m}{\ln(1 + \frac{1}{8}\frac{\sqrt{m}}{\kappa_f^M})} \ln \frac{1}{\varepsilon}).$$

Let us take $m = m' = \lceil \frac{n}{1+pd} \rceil$.

If $n > 1 + pd$, or equivalently $\kappa_f < n^2 d^{-2}$, then

$$C_1'(m', \varepsilon) = \mathcal{O}(\frac{n}{\ln(1 + \frac{1}{8}\frac{\sqrt{n}}{\sqrt{pd}\kappa_f^M})} \ln \frac{1}{\varepsilon}).$$

Since $p = \mathcal{O}\Big(\sqrt{\kappa(M)} \ln \big(\sqrt{\kappa_f^M} \kappa(M)\big)\Big)$, we know that when $(\kappa_f^M)^2 \sqrt{\kappa(M)} d < n$, or equivalently $\kappa_f < n^2 d^{-2}$, we have

$$\ln(1 + \frac{1}{8}\frac{\sqrt{n}}{\sqrt{pd}\kappa_f^M}) = \mathcal{O}(\ln n),$$

therefore

$$C_1'(m', \varepsilon) = \mathcal{O}(n \ln \frac{1}{\varepsilon}),$$

and

$$\frac{\min_{m \geq 1} C_1'(m, \varepsilon)}{\min_{m \geq 1} C_1(m, \varepsilon)} \leq \frac{C_1'(m', \varepsilon)}{C_1(m^\star, \varepsilon)} = \mathcal{O}(\frac{\sqrt{n}}{\kappa_f}).$$

If $n \leq 1 + pd$, or equivalently $\kappa_f > n^2 d^{-2}$, then $m = 1$ and

$$C_1'(m, \varepsilon) = \mathcal{O}(\frac{\sqrt{\kappa(M)} d}{\ln(1 + \frac{1}{8}\frac{1}{\kappa_f^M})} \ln \frac{1}{\varepsilon}),$$

therefore

$$\frac{\min_{m \geq 1} C_1'(m, \varepsilon)}{\min_{m \geq 1} C_1(m, \varepsilon)} \leq \frac{C_1'(1, \varepsilon)}{C_1(m^\star, \varepsilon)} = \mathcal{O}(\frac{\sqrt{\kappa(M)} d}{\kappa_f \sqrt{n} \ln(1 + \frac{1}{8}\frac{1}{\kappa_f^M})}).$$

Since $\kappa(M) \approx \kappa_f \gg \kappa_f^M$, this ratio becomes $\mathcal{O}(\frac{d}{\sqrt{n \kappa_f}})$ □

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