Supplementary Materials

A. Proof of Theorem 1

We first recall the following lemma.

Lemma 1 (Lemma 1, (Gong et al., 2013)). *Under Assumption 1.{3}. For any $\eta > 0$ and any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ such that $\mathbf{x} = \mathrm{prox}_{\eta g}(\mathbf{y} - \eta \nabla f(\mathbf{y}))$, one has that*

$$F(\mathbf{x}) \le F(\mathbf{y}) - (\tfrac{1}{2\eta} - \tfrac{L}{2})\|\mathbf{x} - \mathbf{y}\|^2.$$

Applying Lemma 1 with $\mathbf{x} = \mathbf{x}_k, \mathbf{y} = \mathbf{y}_k$, we obtain that

$$F(\mathbf{x}_k) \le F(\mathbf{y}_k) - (\tfrac{1}{2\eta} - \tfrac{L}{2})\|\mathbf{x}_k - \mathbf{y}_k\|^2. \tag{12}$$

Since $\eta < \frac{1}{L}$, it follows that $F(\mathbf{x}_k) \le F(\mathbf{y}_k)$. Moreover, the update rule of APGnc guarantees that $F(\mathbf{y}_{k+1}) \le F(\mathbf{x}_k)$. In summary, for all k the following inequality holds:

$$F(\mathbf{y}_{k+1}) \le F(\mathbf{x}_k) \le F(\mathbf{y}_k) \le F(\mathbf{x}_{k-1}). \tag{13}$$

Combing further with the fact that $F(\mathbf{x}_k), F(\mathbf{y}_k) \ge \inf F > -\infty$ for all k, we conclude that $\{F(\mathbf{x}_k)\}, \{F(\mathbf{y}_k)\}$ converge to the same limit F^*, i.e.,

$$\lim_{k\to\infty} F(\mathbf{x}_k) = \lim_{k\to\infty} F(\mathbf{y}_k) = F^*. \tag{14}$$

On the other hand, by induction we conclude from eq. (13) that for all k

$$F(\mathbf{y}_k) \le F(\mathbf{x}_0), \quad F(\mathbf{x}_k) \le F(\mathbf{x}_0).$$

Combining with Assumption 1.1 that F has bounded sublevel set, we conclude that $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$ are bounded and thus have bounded limit points. Now combining eq. (12) and eq. (13) yields

$$\begin{aligned}(\tfrac{1}{2\eta} - \tfrac{L}{2})\|\mathbf{y}_k - \mathbf{x}_k\|^2 &\le F(\mathbf{y}_k) - F(\mathbf{x}_k) \\ &\le F(\mathbf{y}_k) - F(\mathbf{y}_{k+1}),\end{aligned} \tag{15}$$

which, after telescoping over k and letting $k \to \infty$, becomes

$$\sum_{k=1}^{\infty}(\tfrac{1}{2\eta} - \tfrac{L}{2})\|\mathbf{y}_k - \mathbf{x}_k\|^2 \le F(\mathbf{y}_1) - \inf F < \infty. \tag{16}$$

This further implies that $\|\mathbf{y}_k - \mathbf{x}_k\| \to 0$, and hence $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$ share the same set of limit points Ω. Note that Ω is closed (it is a set of limit points) and bounded, we conclude that Ω is compact in $\mathbb{R}^d$.

By optimality condition of the proximal gradient step of APGnc, we obtain that

$$\begin{aligned}&-\nabla f(\mathbf{y}_k) - \tfrac{1}{\eta}(\mathbf{x}_k - \mathbf{y}_k) \in \partial g(\mathbf{x}_k) \\ \Leftrightarrow\ &\underbrace{\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{y}_k) - \tfrac{1}{\eta}(\mathbf{x}_k - \mathbf{y}_k)}_{\mathbf{u}_k} \in \partial F(\mathbf{x}_k),\end{aligned} \tag{17}$$

which further implies that

$$\begin{aligned}\|\mathbf{u}_k\| &= \|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{y}_k) - \tfrac{1}{\eta}(\mathbf{x}_k - \mathbf{y}_k)\| \\ &\le (L + \tfrac{1}{\eta})\|\mathbf{y}_k - \mathbf{x}_k\| \to 0.\end{aligned} \tag{18}$$

Consider any limit point $\mathbf{z}' \in \Omega$, and w.l.o.g we write $\mathbf{x}_k \to \mathbf{z}',\ \mathbf{y}_k \to \mathbf{z}'$ by restricting to a subsequence. By the definition of the proximal map, the proximal gradient step of APGnc implies that

$$\begin{aligned}&\langle \nabla f(\mathbf{y}_k), \mathbf{x}_k - \mathbf{y}_k\rangle + \tfrac{1}{2\eta}\|\mathbf{y}_k - \mathbf{x}_k\|^2 + g(\mathbf{x}_k) \\ &\le \langle \nabla f(\mathbf{y}_k), \mathbf{z}' - \mathbf{y}_k\rangle + \tfrac{1}{2\eta}\|\mathbf{z}' - \mathbf{y}_k\|^2 + g(\mathbf{z}').\end{aligned} \tag{19}$$

Taking $\limsup$ on both sides and note that $\mathbf{x}_k - \mathbf{y}_k \to 0,\ \mathbf{y}_k \to \mathbf{z}'$, we obtain that $\limsup_{k\to\infty} g(\mathbf{x}_k) \le g(\mathbf{z}')$. Since g is lower semicontinuous and $\mathbf{x}_k \to \mathbf{z}'$, it follows that $\limsup_{k\to\infty} g(\mathbf{x}_k) \ge g(\mathbf{z}')$. Combining both inequalities, we conclude that $\lim_{k\to\infty} g(\mathbf{x}_k) = g(\mathbf{z}')$. Note that the continuity of f yields $\lim_{k\to\infty} f(\mathbf{x}_k) = f(\mathbf{z}')$, we then conclude that $\lim_{k\to\infty} F(\mathbf{x}_k) = F(\mathbf{z}')$. Since $\lim_{k\to\infty} F(\mathbf{x}_k) = F^*$ by eq. (14), we conclude that

$$F(\mathbf{z}') \equiv F^*, \quad \forall \mathbf{z}' \in \Omega. \tag{20}$$

Hence, F remains constant on the compact set Ω. To this end, we have established $\mathbf{x}_k \to \mathbf{z}',\ F(\mathbf{x}_k) \to F(\mathbf{z}')$ and that $\partial F(\mathbf{x}_k) \ni \mathbf{u}_k \to 0$. Recall the definition of limiting sub-differential, we conclude that $0 \in \partial F(\mathbf{z}')$ for all $\mathbf{z}' \in \Omega$.

B. Proof of Theorem 2

Throughout the proof we assume that $F(\mathbf{x}_k) \neq F^*$ for all k because otherwise the algorithm terminates and the conclusions hold trivially. We also denote k_0 as a sufficiently large positive integer.

Combining eq. (12) and eq. (13) yields that

$$F(\mathbf{x}_{k+1}) \le F(\mathbf{x}_k) - (\tfrac{1}{2\eta} - \tfrac{L}{2})\|\mathbf{y}_{k+1} - \mathbf{x}_{k+1}\|^2. \tag{21}$$

Moreover, eq. (17) and eq. (18) imply that

$$\mathrm{dist}_{\partial F(\mathbf{x}_k)}(\mathbf{0}) \le (L + \tfrac{1}{\eta})\|\mathbf{y}_k - \mathbf{x}_k\|. \tag{22}$$

We have shown in Appendix A that $F(\mathbf{x}_k) \downarrow F^*$, and it is also clear that $\mathrm{dist}_\Omega(\mathbf{x}_k) \to 0$. Thus, for any $\epsilon, \delta > 0$ and all $k \ge k_0$, we have

$$\mathbf{x}_k \in \{\mathbf{x} \mid \mathrm{dist}_\Omega(\mathbf{x}) \le \epsilon, F^* < F(\mathbf{x}) < F^* + \delta\}.$$

Since Ω is compact and F is constant on it, the uniformized KL property implies that for all $k \ge k_0$

$$\varphi'(F(\mathbf{x}_k) - F^*)\mathrm{dist}_{\partial F(\mathbf{x}_k)}(\mathbf{0}) \ge 1. \tag{23}$$

Recall that $r_k := F(\mathbf{x}_k) - F^*$. Then eq. (23) is equivalent to

$$\begin{aligned}
1 &\le \left(\varphi'(r_k)\,\mathrm{dist}_{\partial F(\mathbf{x}_k)}(\mathbf{0})\right)^2 \\
&\overset{(i)}{\le} (\varphi'(r_k))^2 \left(\tfrac{1}{\eta} + L\right)^2 \|\mathbf{y}_k - \mathbf{x}_k\|^2 \\
&\overset{(ii)}{\le} (\varphi'(r_k))^2 \frac{\left(\frac{1}{\eta}+L\right)^2}{\frac{1}{2\eta} - \frac{L}{2}} [F(\mathbf{x}_{k-1}) - F(\mathbf{x}_k)] \\
&\le d_1 (\varphi'(r_k))^2 (r_{k-1} - r_k),
\end{aligned}$$

where (i) is due to eq. (22), (ii) follows from eq. (21), and $d_1 = \left(\frac{1}{\eta} + L\right)^2 / \left(\frac{1}{2\eta} - \frac{L}{2}\right)$. Since $\varphi(t) = \frac{c}{\theta}t^\theta$, we have that $\varphi'(t) = ct^{\theta-1}$. Thus the above inequality becomes

$$1 \le d_1 c^2 r_k^{2\theta-2}(r_{k-1} - r_k). \tag{24}$$

It has been shown in (Frankel et al., 2015; Li & Lin, 2015) that sequence $\{r_k\}$ satisfying the above inductive property converges to zero at different rates according to θ as stated in the theorem.

C. Proof of Theorem 3

g **non-convex,** $\epsilon_k = 0$: In this setting, we first prove the following inexact version of Lemma 1.

Lemma 2. *Under Assumption 1.3. For any $\eta > 0$ and any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ such that $\mathbf{x} = \mathrm{prox}_{\eta g}(\mathbf{y} - \eta(\nabla f(\mathbf{y}) + \mathbf{e}))$, one has that*

$$F(\mathbf{x}) \le F(\mathbf{y}) + (\tfrac{L}{2} - \tfrac{1}{2\eta})\|\mathbf{x} - \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|\|\mathbf{e}\|.$$

Proof. By Assumption 1.3 we have that

$$f(\mathbf{x}) \le f(\mathbf{y}) + \langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{y})\rangle + \tfrac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2.$$

Also, by the definition of proximal map, the proximal gradient step implies that

$$g(\mathbf{x}) + \tfrac{1}{2\eta}\|\mathbf{x}-\mathbf{y}+\eta(\nabla f(\mathbf{y})+\mathbf{e})\|^2 \le g(\mathbf{y}) + \tfrac{1}{2\eta}\|\eta(\nabla f(\mathbf{y})+\mathbf{e})\|^2,$$

which, after simplifications becomes that

$$g(\mathbf{x}) \le g(\mathbf{y}) - \tfrac{1}{2\eta}\|\mathbf{x}-\mathbf{y}\|^2 - \langle \mathbf{x}-\mathbf{y}, (\nabla f(\mathbf{y})+\mathbf{e})\rangle.$$

Combine the above two inequalities further gives that

$$F(\mathbf{x}) \le F(\mathbf{y}) + (\tfrac{L}{2} - \tfrac{1}{2\eta})\|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|\|\mathbf{e}\|.$$

□

Using Lemma 2 with $\mathbf{x}=\mathbf{x}_k, \mathbf{y}=\mathbf{y}_k, \mathbf{e}=\mathbf{e}_k$ and notice the fact that $\|\mathbf{e}_k\| \le \gamma\|\mathbf{x}_k-\mathbf{y}_k\|$, we obtain that

$$F(\mathbf{x}_k) \le F(\mathbf{y}_k) + (\gamma + \tfrac{L}{2} - \tfrac{1}{2\eta})\|\mathbf{x}_k-\mathbf{y}_k\|^2. \tag{25}$$

Moreover, the optimality condition of the proximal gradient step with gradient error gives that By optimality condition of the proximal gradient step of APGnc, we obtain that

$$\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{y}_k) - \mathbf{e}_k - \tfrac{1}{\eta}(\mathbf{x}_k-\mathbf{y}_k) \in \partial F(\mathbf{x}_k),$$

which further implies that

$$\mathrm{dist}_{\partial F(\mathbf{x}_k)}(\mathbf{0}) \le (\gamma + L + \tfrac{1}{\eta})\|\mathbf{y}_k-\mathbf{x}_k\|. \tag{26}$$

Notice that eq. (25) and eq. (26) are parallel to the key inequalities eq. (21) and eq. (22) in the analysis of exact APGnc. Thus, by choosing $\eta < \frac{1}{2\gamma+L}$ and redefining $d_1 = (\frac{1}{\eta}+L+\gamma)^2/(\frac{1}{2\eta}-\frac{L}{2}-\gamma)$, all the statements in Theorem 1 remain true and the convergence rates in Theorem 2 remain the same order with a worse constant.

g **convex**: We first present the following lemma.

Lemma 3. *For any $\mathbf{x}, \mathbf{v} \in \mathbb{R}^d$, let $\mathbf{u}' \in \partial_\epsilon g(\mathbf{x})$ such that $\nabla f(\mathbf{x}) + \mathbf{u}'$ has minimal norm. Denote $\xi := \mathrm{dist}_{\partial g(\mathbf{x})}(\mathbf{u}')$, then we have*

$$\mathrm{dist}_{\partial F(\mathbf{x})}(\mathbf{0}) \le \mathrm{dist}_{\nabla f(\mathbf{x})+\partial_\epsilon g(\mathbf{x})}(\mathbf{0}) + \xi. \tag{27}$$

Proof. We observe the following

$$\begin{aligned}
\mathrm{dist}_{\partial F(\mathbf{x})}(\mathbf{0}) &= \min_{\mathbf{u}\in\partial g(\mathbf{x})} \|\nabla f(\mathbf{x}) + \mathbf{u}\| \\
&= \min_{\mathbf{u}\in\partial g(\mathbf{x})} \|\nabla f(\mathbf{x}) + \mathbf{u}' + \mathbf{u} - \mathbf{u}'\|, \ \forall \mathbf{u}' \in \partial_\epsilon g(\mathbf{x}) \\
&\le \|\nabla f(\mathbf{x}) + \mathbf{u}'\| + \min_{\mathbf{u}\in\partial g(\mathbf{x})} \|\mathbf{u}-\mathbf{u}'\|, \ \forall \mathbf{u}' \in \partial_\epsilon g(\mathbf{x}) \\
&\le \min_{\mathbf{u}'\in\partial_\epsilon} g(\mathbf{x})\|\nabla f(\mathbf{x}) + \mathbf{u}'\| + \xi \\
&= \mathrm{dist}_{\nabla f(\mathbf{x})+\partial_\epsilon g(\mathbf{x})}(\mathbf{0}) + \xi.
\end{aligned} \tag{28}$$

□

Recall that we have two inexactness, i.e., $\mathbf{x}_k = \mathrm{prox}_{\eta g}^{\epsilon_k}(\mathbf{y}_k - \eta(\nabla f(\mathbf{y}_k)+\mathbf{e}_k))$. Following a proof similar to that of Lemma 2 and notice that $\epsilon_k \le \delta\|\mathbf{x}_k-\mathbf{y}_k\|^2$, we can obtain that

$$\begin{aligned}
F(\mathbf{x}_k) &\le F(\mathbf{y}_k) + (\gamma + \tfrac{L}{2} - \tfrac{1}{2\eta})\|\mathbf{x}_k-\mathbf{y}_k\|^2 + \epsilon_k \\
&\le F(\mathbf{y}_k) + (\gamma' + \tfrac{L}{2} - \tfrac{1}{2\eta})\|\mathbf{x}_k-\mathbf{y}_k\|^2
\end{aligned} \tag{29}$$

for some $\gamma' > \gamma > 0$. Since g is convex, by Lemma 2 in (Schmidt et al., 2011) one can exhibit $\mathbf{v}_k$ with $\|\mathbf{v}_k\| \le \sqrt{2\eta\epsilon_k}$ such that

$$\tfrac{1}{\eta}[\mathbf{y}_k - \mathbf{x}_k - \eta(\nabla f(\mathbf{y}_k) + \mathbf{e}_k) - \mathbf{v}_k] \in \partial_{\epsilon_k} g(\mathbf{x}_k).$$

This implies that

$$\mathrm{dist}_{\nabla f(\mathbf{x}_k) + \partial_{\epsilon_k} g(\mathbf{x}_k)}(\mathbf{0}) \le (\gamma + \tfrac{1}{\eta} + L)\|\mathbf{x}_k - \mathbf{y}_k\| + \sqrt{\tfrac{2\epsilon_k}{\eta}}.$$

Apply Lemma 3 and notice that $\epsilon_k \le \delta\|\mathbf{x}_k - \mathbf{y}_k\|^2, \xi_k \le \lambda\|\mathbf{x}_k - \mathbf{y}_k\|$, we obtain that

$$\mathrm{dist}_{\partial F(\mathbf{x}_k)}(\mathbf{0}) \le (\gamma' + \frac{1}{\eta} + L)\|\mathbf{x}_k - \mathbf{y}_k\| \tag{30}$$

for some $\gamma' > \gamma > 0$. Now eq. (29) and eq. (30) are parallel to the key inequalities eq. (21) and eq. (22) in the analysis of exact APGnc. Thus, by choosing $\eta < \frac{1}{2\gamma' + L}$ and redefining $d_1 = (\frac{1}{\eta} + L + \gamma')^2/(\frac{1}{2\eta} - \frac{L}{2} - \gamma')$, all the statements in Theorem 1 remain true and the convergence rates in Theorem 2 remain the same order with a worse constant.

D. Proof of Theorem 4

We first define the following quantities for the convenience of the proof.

$$c_t = c_{t+1}(1 + \tfrac{1}{m}) + \tfrac{\eta L^2}{2}, \;\; c_m = 0, \tag{31}$$
$$R_k^t := \mathbb{E}\left[F(\mathbf{x}_k^t) + c_t\|\mathbf{x}_k^t - \mathbf{x}_k^0\|^2\right], \tag{32}$$
$$\bar{\mathbf{x}}_k^{t+1} = \mathrm{prox}_{\eta g}(\mathbf{x}_k^t - \eta\nabla f(\mathbf{x}_k^t)). \tag{33}$$

Note that $\bar{\mathbf{x}}_k^{t+1}$ is a reference sequence introduced for the convenience of analysis, and is not being computed in the implementation of the algorithm. Then it has been shown in the proof of Theorem 5 of (Reddi et al., 2016b) that

$$R_k^{t+1} \le R_k^t + \left(L - \tfrac{1}{2\eta}\right)\mathbb{E}\left[\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2\right]. \tag{34}$$

Telescoping eq. (34) over t from $t = 1$ to $t = m-1$, we obtain

$$\mathbb{E}[F(\mathbf{x}_k^m)] \le \mathbb{E}\left[F(\bar{\mathbf{x}}_k^1) + c_1\|\bar{\mathbf{x}}_k^1 - \mathbf{x}_k^0\|^2 + \sum_{t=1}^{m-1}\left(L - \tfrac{1}{2\eta}\right)\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2\right]. \tag{35}$$

Following from eq. (31), a simple induction shows that $c_t \le \eta L^2 m$. Setting $\eta < \frac{1}{2L}$ and recalling that $F(\mathbf{y}_k) \le F(\mathbf{x}_{k-1}^m)$., eq. (35) further implies that

$$\mathbb{E}[F(\mathbf{y}_{k+1})] \le \mathbb{E}[F(\mathbf{x}_k^m)] \le \mathbb{E}[F(\bar{\mathbf{x}}_k^1)] + \eta L^2 m\mathbb{E}[\|\bar{\mathbf{x}}_k^1 - \mathbf{x}_k^0\|^2]. \tag{36}$$

Now telescoping eq. (34) again over t from $t = 0$ to $t = m-1$ and applying eq. (36), we obtain

$$\mathbb{E}[F(\mathbf{x}_k^m)] \le \mathbb{E}[F(\mathbf{y}_k)] + \sum_{t=0}^{m-1}(L - \tfrac{1}{2\eta})\mathbb{E}\left[\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2\right]. \tag{37}$$

Combining all the above facts, we conclude that for $\eta < \frac{1}{2L}$

$$\mathbb{E}[F(\mathbf{y}_k)] \le \mathbb{E}[F(\mathbf{y}_{k-1})] \le \ldots \le F(\mathbf{y}_0). \tag{38}$$

Since $\mathbb{E}[F(\cdot)]$ is bounded below, $\mathbb{E}[F(\mathbf{y}_k)]$ decreases to a finite limit, say, F^*. Define $r_k = \mathbb{E}\left[F(\mathbf{y}_k) - F^*\right]$, and assume $r_k > 0$ for all k (since otherwise $r_k = 0$ and the algorithm terminates). Applying the KŁ property with $\theta = 1/2$, we obtain

$$\tfrac{1}{c}(F(\mathbf{x}) - F^*)^{\frac{1}{2}} \le \mathrm{dist}_{\partial F(\mathbf{x})}(\mathbf{0}). \tag{39}$$

Setting $\mathbf{x} = \bar{\mathbf{x}}_k^1$, we further obtain

$$\tfrac{1}{c^2}(F(\bar{\mathbf{x}}_k^1) - F^*) \le \mathrm{dist}^2_{\partial F(\bar{\mathbf{x}}_k^1)}(\mathbf{0}) \le \left(L + \tfrac{1}{\eta}\right)^2\|\bar{\mathbf{x}}_k^1 - \mathbf{y}_k\|^2, \tag{40}$$

where the last inequality is due to eq. (33). Taking expectation over both sides and using eq. (36), we obtain

$$\tfrac{1}{c^2}\mathbb{E}[F(\mathbf{x}_k^m) - F^*] - \tfrac{\eta L^2 m}{c^2}\mathbb{E}\left[\|\bar{\mathbf{x}}_k^1 - \mathbf{x}_k^0\|^2\right] \le \left(L + \tfrac{1}{\eta}\right)^2 \mathbb{E}\left[\|\bar{\mathbf{x}}_k^1 - \mathbf{y}_k\|^2\right]. \tag{41}$$

Noting that $\mathbf{x}_k^0 = \mathbf{y}_k$ and $\mathbb{E}F(\mathbf{y}_{k+1}) \le \mathbb{E}F(\mathbf{x}_k^m)$, we then rearrange the above inequality and obtain

$$\tfrac{1}{c^2}\mathbb{E}[F(\mathbf{y}_{k+1}) - F^*] \le \tfrac{1}{c^2}\mathbb{E}[F(\mathbf{x}_k^m) - F^*] \le \left[\left(L + \tfrac{1}{\eta}\right)^2 + \tfrac{\eta L^2 m}{c^2}\right]\mathbb{E}\left[\|\bar{\mathbf{x}}_k^1 - \mathbf{y}_k\|^2\right] \tag{42}$$

$$\le \frac{\left(L+\frac{1}{\eta}\right)^2 + \frac{\eta L^2 m}{c^2}}{\frac{1}{2\eta} - L}\left(\mathbb{E}[F(\mathbf{y}_k)] - E[F(\mathbf{y}_{k+1})]\right), \tag{43}$$

which can be further rewritten as

$$r_{k+1} \le d\left(r_k - r_{k+1}\right), \tag{44}$$

where $d = \frac{c^2\left(L+\frac{1}{\eta}\right)^2 + \eta L^2 m}{\frac{1}{2\eta} - L}$. Then a simple induction yields that

$$r_{k+1} \le \left(\tfrac{d}{d+1}\right)^{k+1}\left(F(\mathbf{y}_0) - F^*\right). \tag{45}$$

E. Proof of Theorem 5

We first introduce some auxiliary lemmas.

Lemma 4. *Consider the convex function g and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ such that $\mathbf{y} = \mathrm{prox}_{\eta g}^{\epsilon}(\mathbf{x})$ for some $\epsilon > 0$. Then, there exists $\|\mathbf{i}\| \le \sqrt{2\eta\epsilon}$ that satisfies the following inequality for all $\mathbf{z} \in \mathbb{R}^d$.*

$$g(\mathbf{y}) + \tfrac{1}{2\eta}\|\mathbf{y} - \mathbf{x}\|^2 \le g(\mathbf{z}) + \tfrac{1}{2\eta}\|\mathbf{z} - \mathbf{x}\|^2 - \tfrac{1}{2\eta}\|\mathbf{y} - \mathbf{z}\|^2 + \langle \mathbf{y} - \mathbf{z}, \tfrac{1}{\eta}\mathbf{i}\rangle + \epsilon. \tag{46}$$

Proof. By Lemma 2 in (Schmidt et al., 2011), there exists $\|\mathbf{i}\| \le \sqrt{2\eta\epsilon}$ such that

$$\tfrac{1}{\eta}\left(\mathbf{x} - \mathbf{y} - \mathbf{i}\right) \in \partial_\epsilon g(\mathbf{y}). \tag{47}$$

Then, the definition of ϵ-subdifferential implies that

$$g(\mathbf{z}) - g(\mathbf{y}) \ge \langle \mathbf{z} - \mathbf{y}, \partial_\epsilon g(\mathbf{y})\rangle - \epsilon = \langle \mathbf{z} - \mathbf{y}, \tfrac{1}{\eta}\left(\mathbf{x} - \mathbf{y} - \mathbf{i}\right)\rangle - \epsilon, \ \forall\, \mathbf{z} \in \mathbb{R}^d. \tag{48}$$

The desired result follows by rearranging the above inequality. □

Lemma 5. *Consider the convex function g and $\mathbf{x}, \mathbf{y}, \mathbf{d} \in \mathbb{R}^d$ such that $\mathbf{y} = \mathrm{prox}_{\eta g}^{\epsilon}(\mathbf{x} - \eta\mathbf{d})$ for some $\epsilon > 0$. Then, there exists $\|\mathbf{i}\| \le \sqrt{2\eta\epsilon}$ that satisfies the following inequality for all $\mathbf{z} \in \mathbb{R}^d$.*

$$g(\mathbf{y}) = \langle \mathbf{y} - \mathbf{z}, \mathbf{d} - \tfrac{1}{\eta}\mathbf{i}\rangle \le g(\mathbf{z}) + \tfrac{1}{2\eta}\left[\|\mathbf{z} - \mathbf{x}\|^2 - \|\mathbf{y} - \mathbf{z}\|^2 - \|\mathbf{y} - \mathbf{x}\|^2\right] + \epsilon. \tag{49}$$

Proof. By Lemma 4, we obtain the following inequality for all $\mathbf{z} \in \mathbb{R}^d$.

$$\begin{aligned} g(\mathbf{y}) &+ \langle \mathbf{y} - \mathbf{x}, \mathbf{d}\rangle + \tfrac{1}{2\eta}\|\mathbf{y} - \mathbf{x}\|^2 + \frac{\eta}{2}\|\mathbf{d}\|^2 \\ &\le g(\mathbf{z}) + \tfrac{1}{2\eta}\|\mathbf{z} - \mathbf{x} + \eta\mathbf{d}\|^2 - \tfrac{1}{2\eta}\|\mathbf{y} - \mathbf{z}\|^2 + \langle \mathbf{y} - \mathbf{z}, \tfrac{1}{\eta}\mathbf{i}\rangle + \epsilon \\ &= g(\mathbf{z}) + \langle \mathbf{z} - \mathbf{x}, \mathbf{d}\rangle\tfrac{1}{2\eta}\|\mathbf{z} - \mathbf{x}\|^2 + \tfrac{\eta}{2}\|\mathbf{d}\|^2 - \tfrac{1}{2\eta}\|\mathbf{y} - \mathbf{z}\|^2 + \langle \mathbf{y} - \mathbf{z}, \tfrac{1}{\eta}\mathbf{i}\rangle + \epsilon. \end{aligned} \tag{50}$$

The desired result follows by rearranging the above inequality. □

Lemma 6. *Consider the convex function g and $\mathbf{x}, \mathbf{y}, \mathbf{d} \in \mathbb{R}^d$ such that $\mathbf{y} = \mathrm{prox}_{\eta g}^{\epsilon}(\mathbf{x} - \eta\mathbf{d})$ for some $\epsilon > 0$. Then, there exists $\|\mathbf{i}\| \le \sqrt{2\eta\epsilon}$ that satisfies the following inequality for all $\mathbf{z} \in \mathbb{R}^d$.*

$$F(\mathbf{y}) + \langle \mathbf{y} - \mathbf{z}, \mathbf{d} - \tfrac{1}{\eta}\mathbf{i} - \nabla f(\mathbf{x})\rangle \le F(\mathbf{z}) + \left(\tfrac{L}{2} - \tfrac{1}{2\eta}\right)\|\mathbf{y} - \mathbf{x}\|^2 + \left(\tfrac{L}{2} + \tfrac{1}{2\eta}\right)\|\mathbf{z} - \mathbf{x}\|^2 - \tfrac{1}{2\eta}\|\mathbf{y} - \mathbf{z}\|^2 + \epsilon. \tag{51}$$

Proof. By Lipschitz continuity of ∇f, we obtain

$$f(\mathbf{y}) \le f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \tfrac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2, \tag{52}$$

$$f(\mathbf{x}) \le f(\mathbf{z}) + \langle \nabla f(\mathbf{x}), \mathbf{x}-\mathbf{z}\rangle + \tfrac{L}{2}\|\mathbf{x}-\mathbf{z}\|^2. \tag{53}$$

Adding the above inequalities together yields

$$f(\mathbf{y}) \le f(\mathbf{z}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{z}\rangle + \tfrac{L}{2}\left[\|\mathbf{y}-\mathbf{x}\|^2 + \|\mathbf{z}-\mathbf{x}\|^2\right]. \tag{54}$$

Combining with Lemma 5, we then obtain the desired result. □

Recall the reference sequence $\bar{\mathbf{x}}_k^{t+1} = \mathrm{prox}_{\eta g}(\mathbf{x}_k^t - \eta\nabla f(\mathbf{x}_k^t))$. Applying Lemma 6 with $\epsilon = 0$, $\mathbf{y} = \bar{\mathbf{x}}_k^{t+1}$, $\mathbf{z} = \mathbf{x}_k^t$, and $\mathbf{d} = \nabla f(\mathbf{x}_k^t)$ and taking expectation on both sides, we obtain

$$\mathbb{E}[F(\bar{\mathbf{x}}_k^{t+1})] \le \mathbb{E}\left[F(\mathbf{x}_k^t) + \left(\tfrac{L}{2} - \tfrac{1}{2\eta}\right)\|\bar{\mathbf{x}}_k^{t+1}) - \mathbf{x}_k^t\|^2 - \tfrac{1}{2\eta}\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2\right]. \tag{55}$$

Similarly, applying Lemma 6 with $\epsilon = \epsilon_k^t$, $\mathbf{y} = \mathbf{x}_k^{t+1}$, $\mathbf{z} = \bar{\mathbf{x}}_k^{t+1}$, $\mathbf{d} = \mathbf{v}_k^t$ and taking expectation on both sides, we obtain

$$\begin{aligned}\mathbb{E}[F(\mathbf{x}_k^{t+1})] \le \mathbb{E}\Big[&F(\bar{\mathbf{x}}_k^{t+1}) + \langle \mathbf{x}_k^{t+1} - \bar{\mathbf{x}}_k^{t+1}, \nabla f\left(\mathbf{x}_k^t\right) - \mathbf{v}_k^t + \tfrac{1}{\eta}\mathbf{i}_k\rangle \\ &+ \left(\tfrac{L}{2} - \tfrac{1}{2\eta}\right)\|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\|^2 + \left(\tfrac{L}{2} + \tfrac{1}{2\eta}\right)\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2 - \tfrac{1}{2\eta}\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^{t+1}\|^2 + \epsilon_k^t\Big].\end{aligned} \tag{56}$$

Adding eq. (55) and eq. (56) together yields

$$\mathbb{E}[F(\mathbf{x}_k^{t+1})] \le \mathbb{E}\left[F(\mathbf{x}_k^t) + \left(L - \tfrac{1}{2\eta}\right)\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2 + \left(\tfrac{L}{2} - \tfrac{1}{2\eta}\right)\|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\|^2 - \tfrac{1}{2\eta}\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^{t+1}\|^2 + T\right] \tag{57}$$

where $T = \langle \mathbf{x}_k^{t+1} - \bar{\mathbf{x}}_k^{t+1}, \nabla f\left(\mathbf{x}_k^t\right) - \mathbf{v}_k^t + \frac{\mathbf{i}_k}{\eta}\rangle + \epsilon_k^t$. Now we bound $\mathbb{E}[T]$ as follows.

$$\mathbb{E}[T] \le \tfrac{1}{2\eta}\mathbb{E}\left[\|\mathbf{x}_k^{t+1} - \mathbf{x}_k^{t+1}\|^2\right] + \tfrac{\eta}{2}\mathbb{E}\left[\|\nabla f\left(\mathbf{x}_k^t\right) - \mathbf{v}_k^t + \tfrac{\mathbf{i}_k}{\eta}\|^2\right] + \epsilon_k^t \tag{58}$$

$$\le \tfrac{1}{2\eta}\mathbb{E}\left[\|\mathbf{x}_k^{t+1} - \mathbf{x}_k^{t+1}\|^2\right] + \eta\mathbb{E}\left[\|\nabla f\left(\mathbf{x}_k^t\right) - \mathbf{v}_k^t\|^2\right] + \eta\mathbb{E}\left[\|\tfrac{\mathbf{i}_k}{\eta}\|^2\right] + \epsilon_k^t \tag{59}$$

$$\le \tfrac{1}{2\eta}\mathbb{E}\left[\|\mathbf{x}_k^{t+1} - \mathbf{x}_k^{t+1}\|^2\right] + \eta\mathbb{E}\left[\|\nabla f\left(\mathbf{x}_k^t\right) - \mathbf{v}_k^t\|^2\right] + 3\epsilon_k^t. \tag{60}$$

By Lemma 3 of (Reddi et al., 2016b), it holds that $\mathbb{E}\left[\|\nabla f\left(\mathbf{x}_k^t\right) - \mathbf{v}_k^t\|^2\right] \le L^2\mathbb{E}\left[\|\mathbf{x}_k^t - \mathbf{x}_k^0\|^2\right]$. Combining with the above inequality, we further obtain that

$$\mathbb{E}[T] \le \tfrac{1}{2\eta}\mathbb{E}\left[\|\mathbf{x}_k^{t+1} - \mathbf{x}_k^{t+1}\|^2\right] + \eta L^2\mathbb{E}\left[\|\mathbf{x}_k^t - \mathbf{x}_k^0\|^2\right] + 3\epsilon_k^t. \tag{61}$$

Substituting the above result into eq. (57), we obtain

$$\mathbb{E}[F(\mathbf{x}_k^{t+1})] \le \mathbb{E}\left[F(\mathbf{x}_k^t) + \left(L - \tfrac{1}{2\eta}\right)\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2 + \left(\tfrac{L}{2} - \tfrac{1}{2\eta}\right)\|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\|^2 + \eta L^2\|\mathbf{x}_k^t - \mathbf{x}_k^0\|^2 + 3\epsilon_k^t\right]. \tag{62}$$

Recalling that $R_k^t := \mathbb{E}\left[F(\mathbf{x}_k^t) + c_t\|\mathbf{x}_k^t - \mathbf{x}_k^0\|^2\right]$, where $c_t = \eta L^2\frac{(1+\beta)^{m-t}-1}{\beta}$ with $\beta > 0$. Then, we can upper bound R_k^{t+1} as

$$R_k^{t+1} = \mathbb{E}\left[F(\mathbf{x}_k^{t+1}) + c_{t+1}\|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t + \mathbf{x}_k^t - \mathbf{x}_k^0\|^2\right] \tag{63}$$

$$= \mathbb{E}\left[F(\mathbf{x}_k^{t+1}) + c_{t+1}\left(\|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\|^2 + \|\mathbf{x}_k^t - \mathbf{x}_k^0\|^2 + 2\langle \mathbf{x}_k^{t+1} - \mathbf{x}_k^t, \mathbf{x}_k^t - \mathbf{x}_k^0\rangle\right)\right] \tag{64}$$

$$\le \mathbb{E}\left[F(\mathbf{x}_k^{t+1}) + c_{t+1}\left(1 + \tfrac{1}{\beta}\right)\|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\|^2 + c_{t+1}\left(1+\beta\right)\|\mathbf{x}_k^t - \mathbf{x}_k^0\|^2\right] \tag{65}$$

$$\le \mathbb{E}\Big[F(\mathbf{x}_k^t) + \left(L - \tfrac{1}{2\eta}\right)\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2 + \left[c_{t+1}\left(1 + \tfrac{1}{\beta}\right) + \tfrac{L}{2} - \tfrac{1}{2\eta}\right]\|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\|^2 \tag{66}$$

$$+ \left[c_{t+1}\left(1+\beta\right) + \eta L^2\right]\|\mathbf{x}_k^t - \mathbf{x}_k^0\|^2 + 3\epsilon_k^t\Big]. \tag{67}$$

Setting $\beta = 1/m$ in c_t and observe that

$$c_t = \eta L^2 \tfrac{(1+\beta)^{m-t}-1}{\beta} = \eta L^2 m \left((1+\beta)^{m-t} - 1\right) \le \eta L^2 m\,(e-1) \le 2\eta L^2 m, \tag{68}$$

which further implies that

$$c_{t+1}\left(1+\tfrac{1}{\beta}\right) + \tfrac{L}{2} \le 2\eta L^2 m(1+m) \le 4\eta L^2 m^2 + \tfrac{L}{2} = 4\rho L m^2 + \tfrac{L}{2} \le \tfrac{1}{2\eta}. \tag{69}$$

Also note that $c_t = c_{t+1}(1+\beta) + \eta L^2$. Collecting all these facts, R_k^{t+1} can be further upper bounded by

$$R_k^{t+1} \le R_k^t + \mathbb{E}\left[\left(L - \tfrac{1}{2\eta}\right)\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2 + 3\epsilon_k^t\right]. \tag{70}$$

Telescoping eq. (70) from $t = 1$ to $t = m-1$, we obtain

$$\mathbb{E}[F(\mathbf{x}_k^m)] \le \mathbb{E}\left[F(\bar{\mathbf{x}}_k^1) + c_1\|\bar{\mathbf{x}}_k^1 - \mathbf{x}_k^0\|^2 + \sum_{t=1}^{m-1}\left(L - \tfrac{1}{2\eta}\right)\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2 + \sum_{t=1}^{m-1} 3\epsilon_k^t\right]. \tag{71}$$

Again, telescoping eq. (70) from $t = 0$ to $t = m-1$ we obtain

$$\mathbb{E}[F(\mathbf{y}_{k+1})] \le \mathbb{E}[F(\mathbf{x}_k^m)] \le \mathbb{E}[F(\mathbf{y}_k)] + \sum_{t=0}^{m-1}(L - \tfrac{1}{2\eta})\mathbb{E}\left[\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2\right] + 3\sum_{t=0}^{m-1}\mathbb{E}\left[\epsilon_k^t\right]. \tag{72}$$

Assume $\sum\limits_{t=0}^{m-1}\mathbb{E}\left[\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2\right] > 0$, because otherwise the algorithm is terminated. Assume that there exists $\alpha > 0$ such that $3\sum\limits_{t=0}^{m-1}\mathbb{E}\left[\epsilon_k^t\right] \le \alpha \sum\limits_{t=0}^{m-1}\mathbb{E}\left[\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2\right]$ and $\frac{1}{2\eta} - L - \alpha > 0$. Then eq. (72) further implies that

$$\mathbb{E}[F(\mathbf{y}_{k+1})] \le \mathbb{E}[F(\mathbf{x}_k^m)] \le \mathbb{E}[F(\mathbf{y}_k)] + \sum_{t=0}^{m-1}(L - \tfrac{1}{2\eta} + \alpha)\mathbb{E}\left[\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2\right]. \tag{73}$$

That is, we have $\mathbb{E}[F(\mathbf{y}_k)] \le \mathbb{E}[F(\mathbf{y}_{k-1})] \le \ldots \le F(\mathbf{y}_0)$, and hence $\mathbb{E}[F(\mathbf{y}_k)] \downarrow F^*$. We can further upper bound eq. (71) as

$$\begin{aligned}
\mathbb{E}[F(\mathbf{x}_k^m)] \le&\, \mathbb{E}\left[F(\bar{\mathbf{x}}_k^1) + c_1\|\bar{\mathbf{x}}_k^1 - \mathbf{x}_k^0\|^2 + \sum_{t=1}^{m-1}\left(L - \tfrac{1}{2\eta}\right)\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2 + \sum_{t=1}^{m-1} 3\epsilon_k^t\right] \\
\le&\, E\left[F(\bar{\mathbf{x}}_k^1) + c_1\|\bar{\mathbf{x}}_k^1 - \mathbf{x}_k^0\|^2 - \left(L - \tfrac{1}{2\eta}\right)\|\bar{\mathbf{x}}_k^1 - \mathbf{x}_k^0\|^2 + \sum_{t=0}^{m-1}\left(L - \tfrac{1}{2\eta}\right)\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2 + \sum_{t=0}^{m-1} 3\epsilon_k^t\right] \\
\le&\, \mathbb{E}\left[F(\bar{\mathbf{x}}_k^1) + \left(c_1 + \tfrac{1}{2\eta}\right)\|\bar{\mathbf{x}}_k^1 - \mathbf{x}_k^0\|^2 + \sum_{t=0}^{m-1}\left(L - \tfrac{1}{2\eta} + \alpha\right)\|\bar{\mathbf{x}}_k^{t+1} - \mathbf{x}_k^t\|^2\right] \\
\le&\, \mathbb{E}\left[F(\bar{\mathbf{x}}_k^1)\right] + \mathbb{E}\left[\left(2\eta L^2 m + \tfrac{1}{2\eta}\right)\|\bar{\mathbf{x}}_k^1 - \mathbf{x}_k^0\|^2\right].
\end{aligned} \tag{74}$$

Define $r_k = \mathbb{E}\left[F(\mathbf{y}_k) - F^*\right]$, and suppose $r_k > 0$ for all k (otherwise the algorithm terminates in finite steps). Applying the KŁ condition with $\theta = 1/2$, we obtain

$$\tfrac{1}{c}(F(\mathbf{x}) - F^*)^{\frac{1}{2}} \le \mathrm{dist}_{\partial F(\mathbf{x})}(\mathbf{0}). \tag{75}$$

Setting $\mathbf{x} = \bar{\mathbf{x}}_k^1$, we obtain

$$\frac{1}{c^2}(F(\bar{\mathbf{x}}_k^1) - F^*) \le \mathrm{dist}^2_{\partial F(\bar{\mathbf{x}}_k^1)}(\mathbf{0}) \le \left(L + \tfrac{1}{\eta}\right)^2\|\bar{\mathbf{x}}_k^1 - \mathbf{y}_k\|^2. \tag{76}$$

Taking expectation on both sides and using the result from eq. (74), we obtain

$$\frac{1}{c^2}\mathbb{E}[F(\mathbf{x}_k^m) - F^*] - \frac{2\eta L^2 m + \frac{1}{2\eta}}{c^2}\mathbb{E}\left[\|\bar{\mathbf{x}}_k{}^1 - \mathbf{x}_k^0\|^2\right] \le \left(L + \tfrac{1}{\eta}\right)^2 \mathbb{E}\left[\|\bar{\mathbf{x}}_k^1 - \mathbf{y}_k\|^2\right]. \tag{77}$$

Note that $\mathbf{x}_k^0 = \mathbf{y}_k$. Then rearranging the above inequality yields

$$\tfrac{1}{c^2}\mathbb{E}[F(\mathbf{y}_{k+1}) - F^*] \le \frac{1}{c^2}E[F(\mathbf{x}_k^m) - F^*] \le \left[\left(L + \tfrac{1}{\eta}\right)^2 + \frac{2\eta L^2 m + \frac{1}{2\eta}}{c^2}\right]\mathbb{E}\left[\|\bar{\mathbf{x}}_k^1 - \mathbf{y}_k\|^2\right] \tag{78}$$

$$\le \frac{\left(L+\frac{1}{\eta}\right)^2 + \frac{2\eta L^2 m + \frac{1}{2\eta}}{c^2}}{\frac{1}{2\eta} - L - \alpha}\left(\mathbb{E}[F(\mathbf{y}_k)] - \mathbb{E}[F(\mathbf{y}_{k+1})]\right), \tag{79}$$

which can be rewritten as $r_{k+1} \le d\,(r_k - r_{k+1})$ with $d = \frac{c^2\left(L+\frac{1}{\eta}\right)^2 + 2\eta L^2 m + \frac{1}{2\eta}}{\frac{1}{2\eta} - L - \alpha}$. Then, induction yields that

$$r_{k+1} \le \frac{d}{d+1} r_k \le \left(\frac{d}{d+1}\right)^{k+1}\left(F(\mathbf{y}_0) - F^*\right). \tag{80}$$