A. Proof in Section 3

A.1. Proof of Lemma 1

Proof. We prove the first part. The second part was proved in (Nesterov, 2015). Recall that

$$f(x_1) - f(x_2) \leq \langle \nabla f(x_2), x_1 - x_2 \rangle + \frac{L}{1+v} \|x_1 - x_2\|^{1+v}.$$  \hfill (12)

Define $\phi(y) = f(x + y) - f(x) - \langle \nabla f(x), y \rangle$. By $(L, v)$-Hölder continuity of $\nabla f(x)$, one has $\phi(y) \leq \frac{L}{1+v} \|y\|^{1+v}$ due to (12). Denote $\psi(y) = \frac{L}{1+v} \|y\|^{1+v}$.

Given the definition of $\phi(y)$ and $\psi(y)$, we could derive their convex conjugates, denoted by $\phi^*(u)$ and $\psi^*(u)$. For $\phi^*(u)$, one has

$$\phi^*(u) = \sup_y \langle y, u \rangle - \phi(y)$$

$$= \sup_y \langle y, u \rangle - [f(x + y) - f(x) - \langle \nabla f(x), y \rangle]$$

$$= \sup_y \langle y, u + \nabla f(x) \rangle - f(x) + f(x)$$

$$\eqref{eq:phi_conjugate} \sup_z \langle z - x, u + \nabla f(x) \rangle - f(z) + f(x)$$

$$= \sup_z \langle z, u + \nabla f(x) \rangle - f(z) + f(x) - \langle x, u + \nabla f(x) \rangle$$

$$\eqref{eq:subdifferential} f^*(u + \nabla f(x)) + f(x) - \langle x, \nabla f(x) \rangle - \langle x, u \rangle$$

$$\eqref{eq:subdifferential} f^*(u + \nabla f(x)) - f^*(\nabla f(x)) - \langle x, u \rangle,$$

where $\eqref{eq:phi_conjugate}$ is due to letting $z = x + y$. $\eqref{eq:subdifferential}$ is due to the definition of convex conjugate and $\eqref{eq:subdifferential}$ is due to Fenchel-Young inequality (in this case, the equality holds), i.e., $f(x) + f^*(\nabla f(x)) = \langle x, \nabla f(x) \rangle$.

For $\psi^*(u)$, one has

$$\psi^*(u) = \sup_y \langle y, u \rangle - \frac{L}{1+v} \|y\|^{1+v}$$

$$\eqref{eq:psi_conjugate} \langle y^*(u), u \rangle - \frac{L}{1+v} \|y^*(u)\|^{1+v}$$

$$\eqref{eq:optimization} \|y^*(u)\|_2 \|u\|_2 - \frac{L}{1+v} \|y^*(u)\|^{1+v}_2$$

$$\eqref{eq:optimization} \left(\frac{1}{L}\right)^\frac{1}{2} \|u\|_2^{1+\frac{1}{2}} - \frac{L}{1+v} \left(\frac{1}{L}\right)^\frac{1}{2} \|u\|_2^{1+\frac{1}{2}}$$

$$= \left(1 - \frac{1}{1+v}\right) \left(\frac{1}{L}\right)^\frac{1}{2} \|u\|_2^{1+\frac{1}{2}},$$

where $\eqref{eq:phi_conjugate}$ is due to letting $y^*(u) \in \arg \max_y \langle y, u \rangle = \frac{L}{1+v} \|y\|^{1+v}$. $\eqref{eq:optimization}$ is due to $u = L\|y^*(u)\|^{2-1} \cdot y^*(u)$ and thus $\langle u, y^*(u) \rangle = \|u\|_2 \cdot \|y^*(u)\|_2$. $\eqref{eq:optimization}$ is due to $\|u\| \cdot \|y^*(u)\| = \langle u, y^*(u) \rangle = L\|y^*(u)\|_2^{2-1} \langle y^*(u), y^*(u) \rangle = L\|y^*(u)\|^{2+1}_2$.

Due to Lemma 19 of (Shalev-Shwartz & Singer, 2010), if $\phi(y) \leq \psi(y)$, then one has $\phi^*(u) \geq \psi^*(u)$ and thus for all $u$ and $x$,

$$f^*(u + \nabla f(x)) - f^*(\nabla f(x)) - \langle x, u \rangle \geq \left(1 - \frac{1}{1+v}\right) \left(\frac{1}{L}\right)^\frac{1}{2} \|u\|_2^{1+\frac{1}{2}}.$$ \hfill (13)

Let $u'$ be any point in the relative interior of the domain of $f^*$. Then we need to prove that if $x \in \partial f^*(u')$, then $u' = \nabla f(x)$.
By Fenchel-Young inequality, one has $\langle x, u' \rangle = f(x) + f^*(u')$ and $\langle x, \nabla f(x) \rangle = f(x) + f^*(\nabla f(x))$. By (13),

\[
0 = f(x) - f(x) \\
= \langle x, \nabla f(x) \rangle - f^*(\nabla f(x)) - (\langle x, u' \rangle - f^*(u')) \\
= f^*(u') - f^*(\nabla f(x)) - \langle x, u' - \nabla f(x) \rangle \\
\geq \left(1 - \frac{1}{1 + v}\right) \left(\frac{1}{L}\right)^{\frac{1}{2}} \|u' - \nabla f(x)\|_2^{1 + \frac{1}{2}},
\]

which implies that $u' = \nabla f(x)$. Thus,

\[
f^*(u + u') - f^*(u') - \langle \partial f^*(u'), u \rangle \geq \frac{1}{2} \frac{2v}{1 + v} \left(\frac{1}{L}\right)^{\frac{1}{2}} \|u\|_2^{1 + \frac{1}{2}}
\]

implies $f^*$ is $(\rho, p)$-uniformly convex with $\rho = \frac{2v}{1 + v} \frac{1}{L^p}$ and $p = 1 + \frac{1}{v}$.

\[\square\]

A.2. Proof of Lemma 2

Proof. First consider

\[
\|\nabla_x f_0(x, y) - \nabla_x f_0(x', y')\|^2 \\
= \|\nabla g(x) - \nabla g(x') - \nabla \ell(x)^\top y + \nabla \ell(x')^\top y'\|^2 \\
\leq 2\|\nabla g(x) - \nabla g(x')\|^2 + 2\|\nabla \ell(x)^\top y - \nabla \ell(x')^\top y + \nabla \ell(x')^\top y - \nabla \ell(x')^\top y'\|^2 \\
\leq 2L_g^2 \|x - x'\|^2 + 4D_g^2 \|\nabla \ell(x) - \nabla \ell(x')\|^2 + 4G_f^2 \|y - y'\|^2 \\
\leq 2L_g^2 \|x - x'\|^2 + 4L_g^2 D_g^2 \|x - x'\|^2 + 4G_f^2 \|y - y'\|^2 \tag{14}
\]

Then consider

\[
\|\nabla_y f_0(x, y) - \nabla_y f_0(x', y')\|^2 \\
= \|\ell(x) - \ell(x')\|^2 \leq G_f^2 \|x - x'\|^2. \tag{15}
\]

Combining the above two inequalities (14) and (15), one has

\[
\|\nabla_x f(x, y) - \nabla_x f(x', y')\|^2 + \|\nabla_y f(x, y) - \nabla_y f(x', y')\|^2 \\
\leq (2L_g^2 + 4L_g^2 D_g^2 + G_f^2) \|x - x'\|^2 + 4G_f^2 \|y - y'\|^2 \\
\leq L^2 (\|x - x'\|^2 + \|y - y'\|^2),
\]

where $L = \sqrt{\max(2L_g^2 + 4L_g^2 D_g^2 + G_f^2, 4G_f^2)}$.

\[\square\]

A.3. Proof of Proposition 1

Proof. This analysis is borrowed from the proof of Theorem 2 in (Xu et al., 2019). For completeness, we include it here. Let $w = (x, y)$, $\nabla_x f_0^{(t)} = \nabla_x f_0(x_t, y_t)$, $\nabla_y f_0^{(t)} = \nabla_y f_0(x_t, y_t)$, $\nabla f_0^{(t)} = (\nabla_x f_0^{(t)}, \nabla_y f_0^{(t)})$ and $\nabla f_0^{(t)} = (\nabla_x f_0^{(t)}, \nabla_y f_0^{(t)})$. By the udpate of $x_{t+1} = \Pi_X[x_t - \eta \nabla_x f_0^{(t)}]$, we know

\[
x_{t+1} = \arg\min_{x \in \mathbb{R}^d} \left\{ I_X(x) + \langle \nabla_x f_0^{(t)}, x - x_t \rangle + \frac{1}{2\eta} \|x - x_t\|^2 \right\}.
\]

and

\[
\langle \nabla_x f_0^{(t)}, x_{t+1} - x_t \rangle + \frac{1}{2\eta} \|x_{t+1} - x_t\|^2 \leq 0. \tag{16}
\]
Stochastic Optimization for Non-convex Inf-Projection Problems

Similarly, by the update of \( y_{t+1} = P_{\eta h}[y_t - \eta \tilde{\nabla}_y f_0^{(t)}] \), we know

\[
y_{t+1} = \arg \min_{y \in \text{dom}(h)} \left\{ h(y) + \langle \tilde{\nabla}_y f_0^{(t)}, y - y_t \rangle + \frac{1}{2\eta} \| y - y_t \|^2 \right\},
\]

and

\[
h(y_{t+1}) + \langle \tilde{\nabla}_y f_0^{(t)}, y_{t+1} - y_t \rangle + \frac{1}{2\eta} \| y_{t+1} - y_t \|^2 \leq h(y_t). \tag{17}
\]

Using the inequalities (16) and (17), and the fact that \( w = (x, y) \), we get

\[
h(y_{t+1}) + \langle \tilde{\nabla} f_0^{(t)}, w_{t+1} - w_t \rangle + \frac{1}{2\eta} \| w_{t+1} - w_t \|^2 \leq h(y_t). \tag{18}
\]

We know from Lemma 2 that \( f_0(w) \) is \( L \)-smooth, thus

\[
f_0(w_{t+1}) \leq f_0(w_t) + \langle \nabla f_0^{(t)}, w_{t+1} - w_t \rangle + \frac{L}{2} \| w_{t+1} - w_t \|^2. \tag{19}
\]

Combining the inequalities (18) and (19) and using the fact that \( f(w) = f_0(w) + h(y) \) we have

\[
\frac{1 - \eta L}{2\eta} \| w_{t+1} - w_t \|^2 \leq f(w_t) - f(w_{t+1}) + \langle \nabla f_0^{(t)} - \tilde{\nabla} f_0^{(t)}, w_{t+1} - w_t \rangle. \tag{20}
\]

Applying Young’s inequality \( \langle a, b \rangle \leq \frac{1}{2L} \| a \|^2 + \frac{L}{2} \| b \|^2 \) to the last inequality of (20), we then have

\[
\frac{1 - 2\eta L}{2\eta} \| w_{t+1} - w_t \|^2 \leq f(w_t) - f(w_{t+1}) + \frac{1}{2L} \| \nabla f_0^{(t)} - \tilde{\nabla} f_0^{(t)} \|^2. \tag{21}
\]

Summing (21) across \( t = 0, \ldots, T - 1 \), we have

\[
\frac{1 - 2\eta L}{2\eta} \sum_{t=0}^{T-1} \| w_{t+1} - w_t \|^2 \leq f(w_0) - f(w_T) + \frac{1}{2L} \sum_{t=0}^{T-1} \| \nabla f_0^{(t)} - \tilde{\nabla} f_0^{(t)} \|^2 \leq M + \frac{1}{2L} \sum_{t=0}^{T-1} \| \nabla f_0^{(t)} - \tilde{\nabla} f_0^{(t)} \|^2, \tag{22}
\]

where the last inequality uses the Assumption 1 (v).

Next, by Exercise 8.8 and Theorem 10.1 of (Rockafellar & Wets, 1998), we know from the updates of \( x_{t+1} \) and \( y_{t+1} \) that

\[
-\tilde{\nabla}_x f_0^{(t)} - \frac{1}{\eta} (x_{t+1} - x_t) \in \partial I_X(x_{t+1}),
\]

\[
-\tilde{\nabla}_y f_0^{(t)} - \frac{1}{\eta} (y_{t+1} - y_t) \in \partial h(y_{t+1}),
\]

and thus

\[
\nabla f_0^{(t+1)} - \tilde{\nabla} f_0^{(t)} - \frac{w_{t+1} - w_t}{\eta} \in \nabla f_0^{(t+1)} + (\partial h(y_{t+1}), \partial I_X(x_{t+1})) = \partial f(w_{t+1}). \tag{23}
\]

Multiplying \( \frac{2}{\eta} \) on both sides of (20) we get

\[
\frac{2}{\eta} \langle \nabla f_0^{(t)} - \nabla f_0^{(t+1)}, w_{t+1} - w_t \rangle + \frac{1 - \eta L}{\eta^2} \| w_{t+1} - w_t \|^2 \leq 2(f(w_t) - f(w_{t+1})) + \frac{2}{\eta} \langle \nabla f_0^{(t)} - \tilde{\nabla} f_0^{(t+1)}, w_{t+1} - w_t \rangle. \tag{24}
\]
By the fact that \( \frac{2}{\eta} \langle \widetilde{\nabla} f_0^{(t)} - \nabla f_0^{(t+1)}, w_{t+1} - w_t \rangle = \| \widetilde{\nabla} f_0^{(t)} - \nabla f_0^{(t+1)} \|^2 - \frac{1}{2} \| w_{t+1} - w_t \|^2 \), then
\[
\| \widetilde{\nabla} f_0^{(t)} - \nabla f_0^{(t+1)} \|^2 \leq \frac{w_{t+1} - w_t}{\eta} \|
\]
\[
\leq \| \nabla f_0^{(t)} - \nabla f_0^{(t+1)} \|^2 + \frac{L}{\eta} ||w_{t+1} - w_t||^2 + \frac{2(f(w_t) - f(w_{t+1}))}{\eta} + \frac{2L}{\eta} ||w_{t+1} - w_t||^2
\]
\[
\leq 2\| \nabla f_0^{(t)} - \nabla f_0^{(t+1)} \|^2 + 2\| \nabla f_0^{(t)} - \nabla f_0^{(t+1)} \|^2 + \frac{3L}{\eta} ||w_{t+1} - w_t||^2 + \frac{2(f(w_t) - f(w_{t+1}))}{\eta}
\]
where the second inequality is due to Cauchy-Schwarz inequality and the smoothness of \( f(w) \); the third inequality is due to Young’s inequality; and the last inequality is due to the smoothness of \( f(w) \). Summing above inequality across \( t = 0, \ldots, T - 1 \), we have
\[
\sum_{t=0}^{T-1} \| \nabla f_0^{(t)} - \nabla f_0^{(t+1)} \|^2 + \frac{w_{t+1} - w_t}{\eta} \|
\]
\[
\leq 2 \sum_{t=0}^{T-1} \| \nabla f_0^{(t)} - \nabla f_0^{(t+1)} \|^2 + (2L^2 + \frac{3L}{\eta}) \sum_{t=0}^{T-1} ||w_{t+1} - w_t||^2 + \frac{2(f(w_0) - f(w_T))}{\eta}
\]
\[
\leq 2 \sum_{t=0}^{T-1} \| \nabla f_0^{(t)} - \nabla f_0^{(t+1)} \|^2 + \frac{2}{\eta^2} \sum_{t=0}^{T-1} ||w_{t+1} - w_t||^2 + \frac{2M}{\eta},
\]
where the last inequality uses the Assumption 1 (v). Combining above inequality with (22) and (23) we obtain
\[
E_R[\text{dist}(0, \hat{\omega} f(w_R)^2)]
\]
\[
\leq \frac{1}{T} \sum_{t=0}^{T-1} E[\| \nabla f_0^{(t)} - \nabla f_0^{(t+1)} \|^2] - \frac{(x_{t+1} - x_t, \eta y_{t+1} - \eta y_t)}{\eta^2}
\]
\[
\leq 2 \sum_{t=0}^{T-1} E[\| \nabla f_0^{(t)} - \nabla f_0^{(t+1)} \|^2] + \frac{2M}{\eta^2} + \frac{2}{\eta^2(1 - 2\eta L)} \left( 2M + \frac{1}{L} \sum_{t=0}^{T-1} E[\| \nabla f_0^{(t)} - \nabla f_0^{(t+1)} \|^2] \right)
\]
\[
= \frac{2c(1 - 2c) + 2}{c(1 - 2c)} \sum_{t=0}^{T-1} E[\| \nabla f_0^{(t)} - \nabla f_0^{(t+1)} \|^2] + \frac{6 - 4c}{1 - 2c} \frac{M}{\eta^2}
\]
\[
\leq \frac{2c(1 - 2c) + 2}{c(1 - 2c)} \sum_{t=0}^{T-1} b(t) + 4 \frac{6 - 4c}{1 - 2c} \frac{M}{\eta^2}
\]
\[
\leq \frac{2c(1 - 2c) + 2}{c(1 - 2c)} \frac{\sigma_0^2 b(T) + 1}{b(T)} + 4 \frac{6 - 4c}{1 - 2c} \frac{M}{\eta^2}
\]
where \( 0 < c < \frac{1}{2} \), the last second inequalit is due to the bounded variance of stochastic gradient and the last inequality uses the fact that \( \sum_{t=1}^T \frac{1}{T} \leq \log(T) + 1. \)

\[ \square \]

A.4. Proof of Lemma 3

Proof. First, we derive \( \nabla F(\hat{x}) \) for any \( \hat{x} \) as follows
\[
\nabla F(\hat{x}) \equiv \nabla g(\hat{x}) - \nabla \ell(\hat{x})^T y^*(\hat{x})
\]
\[
= \nabla g(\hat{x}) - \nabla \ell(\hat{x})^T \hat{y} + \nabla \ell(\hat{x})^T (\hat{y} - y^*(\hat{x}))
\]
\[
= \nabla x f(\hat{x}, \hat{y}) + \nabla \ell(\hat{x})^T (\hat{y} - y^*(\hat{x}))
\]
where $\nabla \ell(x)$ is the Jacobian matrix of $\ell$ at $x$, and $y^*(x) = \arg \min_{y \in \text{dom}(h)} h(y) - \langle y, \ell(x) \rangle = \arg \min_{y \in \text{dom}(h)} f(x, y)$. Here $y^*(x)$ is unique given $x$, since uniform convexity ensures the unique solution ($\nabla h^*$ is Hölder continuous so that $h$ is uniformly convex). Equality (1) above is due to Theorem 10.58 of (Rockafellar & Wets, 2009) and unique $y^*(x)$.

Then by triangle inequality and Cauchy-Schwarz inequality, one has

$$\|\nabla F(\tilde{x})\|_2 \leq \|\nabla f(\tilde{x}, \tilde{y})\|_2 + \|\nabla \ell(\tilde{y})^\top (\tilde{y} - y^*(\tilde{x}))\|_2$$

$$\leq \|\nabla f(\tilde{x}, \tilde{y})\|_2 + \|\nabla \ell(\tilde{y})\|_2 \cdot \|y - y^*(\tilde{x})\|_2$$

$$\leq \|\nabla f(\tilde{x}, \tilde{y})\|_2 + G_\ell \left( \frac{1}{\varrho} \|\partial y f(\tilde{x}, \tilde{y})\|_2 \right)^{\frac{1}{v}}$$

$$= \|\nabla f(\tilde{x}, \tilde{y})\|_2 + G_\ell \left( \frac{1 + v}{2v} L_h^{\frac{1}{2}} \|\partial y f(\tilde{x}, \tilde{y})\|_2 \right)^{\nu}$$

$$= \|\nabla f(\tilde{x}, \tilde{y})\|_2 + G_\ell \left( \frac{1 + v}{2v} \right)^{\nu} L_h \|\partial y f(\tilde{x}, \tilde{y})\|_2^2,$$

where the last inequality is due to $(\varrho, p)$-uniformly convex of $f(\tilde{x}, \tilde{y})$ in $y$ given $\tilde{x}$, i.e., (4). The first equality is due to Lemma 1 that $\varrho = \frac{2v}{1 + v} \cdot \frac{1}{L_h^{\frac{1}{2}}} + p = 1 + \frac{v}{1 + v}$.

**B. Proof in Section 4**

**B.1. Proof of Theorem 3**

**Proof.** Recall the notations $f^k_x(x) = g(x) - y^\top \tilde{\ell}(x)$, $f^k_y(y) = h(y) - y^\top (x_{k+1})$, where $\tilde{\ell}(x) = \ell(x_k) + \nabla \ell(x_k)(x - x_k)$ for the case $\text{dom}(h) \subseteq \mathbb{R}^n$ and $\tilde{\ell}(x) = \ell(x)$ for the case $\text{dom}(h) \subseteq \mathbb{R}^m$.

Define

$$H^k_x(x) = f^k_x(x) + R^k_x(x), \quad H^k_y(y) = f^k_y(y) + R^k_y(y)$$

and

$$v_k = \arg \min_{x \in X} H^k_x(x), \quad u_k = \arg \min_{y \in \text{dom}(h)} H^k_y(y).$$

Both of which are well-defined and unique due to the strong convexity of $H^k$.

Recall that a stochastic gradient of $f^k_x(x)$ can be computed by $\partial g(x; \xi_g) - \nabla \ell(x_k; \xi_\ell) \top y_k$ for $\text{dom}(h) \subseteq \mathbb{R}^n$ or $\partial g(x; \xi_g) - \nabla \ell(x_k; \xi_\ell) \top y_k$ for $\text{dom}(h) \subseteq \mathbb{R}^m$. A stochastic gradient of $f^k_y(y)$ can be computed by $\partial h(y; \xi_h) - \ell(x_{k+1}; \xi_\ell)$, where $\xi_g, \xi_\ell, \xi_h, \xi_\ell'$ denote independent random variables. Then for $f^k_x(x)$ we have

$$\mathbb{E}[\|\nabla f^k_x(x)\|^2] = \mathbb{E}[\|\partial g(x; \xi_g) - \nabla \ell(x_k; \xi_\ell) \top y_k\|^2]$$

$$\leq \mathbb{E}[\|\partial g(x; \xi_g)\|^2] + 2\mathbb{E}[\|\nabla \ell(x_k; \xi_\ell) \top y_k\|^2]$$

$$\leq 2\sigma^2 + 2\mathbb{E}[\|\nabla \ell(x_k; \xi_\ell) \top y_k\|^2]$$

or

$$\mathbb{E}[\|\nabla f^k_x(x)\|^2] = \mathbb{E}[\|\partial g(x; \xi_g) - \nabla \ell(x; \xi_\ell) \top y_k\|^2]$$

$$\leq \mathbb{E}[\|\partial g(x; \xi_g)\|^2] + 2\mathbb{E}[\|\nabla \ell(x; \xi_\ell) \top y_k\|^2]$$

$$\leq 2\sigma^2 + 2\mathbb{E}[\|\nabla \ell(x; \xi_\ell) \top y_k\|^2]$$

$$\leq 2\sigma^2 + 2D^2 \sigma^2$$
where the second inequality uses Assumption 1 (ii); the third inequality uses the assumption of 
\( \max(\|y_k\|^2, E[\|\ell(x_{k+1}; \xi)\|^2]) \leq D^2 \) for all \( k \in \{1, \ldots, K\} \); the last inequality is due to Assumption 1 (iii). 
For \( f_y^k(y) \) we have
\[
\mathbb{E}\|\nabla f_y^k(y)\|^2 = \mathbb{E}\|\partial h(y; \xi_k) - \ell(x_{k+1}; \xi'_k)\|^2 \\
\leq 2\mathbb{E}\|\partial h(y; \xi_k)\|^2 + 2\mathbb{E}\|\ell(x_{k+1}; \xi'_k)\|^2 \\
\leq 2(\sigma^2 + D^2)
\]
where the second inequality uses Assumption 1 (iv) and the assumption of \( \max(\|y_k\|^2, E[\|\ell(x_{k+1}; \xi)\|^2]) \leq D^2 \) for all \( k \in \{1, \ldots, K\} \). We define a constant \( G \), which will be used in our analysis:
\[
G := 17 \max\{2\sigma^2 + 2D^2\sigma^2, 2\sigma^2 + 2D^2\},
\]
which is in fact the role of \( 17\sigma^2 \) in the result of Proposition 2.

Next we could proceed to prove Theorem 3.

Here we focus on the analysis using the convergence result in Proposition 2 corresponding to the non-smooth \( f(z) \). Similar analysis can be done for using the result corresponding to smooth \( f \). Applying Proposition 2 to both \( H_x^k \) and \( H_y^k \) and adding their convergence bound together, we have
\[
\mathbb{E}\left[H_x^k(x_{k+1}) + H_y^k(y_{k+1}) - H_x^k(x_k) - H_y^k(y_k)\right] \\
\leq \frac{G^2}{\gamma(T_k^x + 1)} + \frac{\gamma}{4T_k^y(T_k^x + 1)} \mathbb{E}\|x_k - \nu_k\|^2 + \frac{G^2}{\mu(T_k^y + 1)} + \frac{\mu}{4T_k^y(T_k^x + 1)} \mathbb{E}\|y_k - u_k\|^2
\]
(25)

The following inequalities hold due to the strong convexity of these two functions \( H_x^k(v_k) \leq H_x^k(x_k) - \frac{\gamma}{2}\mathbb{E}\|x_k - \nu_k\|^2 \) and \( H_y^k(u_k) \leq H_y^k(y_k) - \frac{\mu}{2}\mathbb{E}\|y_k - u_k\|^2 \). Plug the above two inequalities to (25),
\[
\mathbb{E}\left[H_x^k(x_{k+1}) + H_y^k(y_{k+1}) - H_x^k(x_k) - H_y^k(y_k)\right] \\
\leq \frac{G^2}{\gamma(T_k^x + 1)} + \frac{\gamma}{4T_k^y(T_k^x + 1)} \mathbb{E}\|x_k - \nu_k\|^2 + \frac{\mu}{4T_k^y(T_k^x + 1)} \mathbb{E}\|y_k - u_k\|^2
\]
Recall the definition of \( H_x^k(x) = f_x^k(x) + R_x^k(x) = f_x^k(x) + \frac{\gamma}{2}\|x - x_k\|^2 \) and \( H_y^k(y) = f_y^k(y) + R_y^k(y) = f_y^k(y) + \frac{\mu}{2}\|y - y_k\|^2 \). Since \( T_k^x \geq 1 \) and \( T_k^y \geq 1 \), we have
\[
\mathbb{E}\left[f_x^k(x_{k+1}) + f_y^k(y_{k+1}) - f_x^k(x_k) - f_y^k(y_k)\right] \\
\leq \frac{G^2}{\gamma(T_k^x + 1)} + \frac{\gamma}{4\mathbb{E}\|x_k - \nu_k\|^2} + \frac{\mu}{4\mathbb{E}\|y_k - u_k\|^2}
\]
(26)
As a result, we have
\[
\frac{1}{4}\left(\mathbb{E}\|x_k - \nu_k\|^2 + \gamma\|x_{k+1} - x_k\|^2 + \mu\|y_k - u_k\|^2 + \delta\|y_{k+1} - y_k\|^2\right) \\
\leq \mathbb{E}\left[f_x^k(x_k) + f_y^k(y_k) - f_x^k(x_{k+1}) - f_y^k(y_{k+1})\right] + \frac{G^2}{\gamma(T_k^x + 1)} + \frac{G^2}{\mu(T_k^y + 1)}
\]
(27)
Let \( T_k^x \geq \frac{k}{\gamma + 1} \) and \( T_k^y \geq \frac{k}{\mu + 1} \), we have
\[
\frac{1}{4}\left(\mathbb{E}\|x_k - \nu_k\|^2 + \gamma\|x_{k+1} - x_k\|^2 + \mu\|y_k - u_k\|^2 + \delta\|y_{k+1} - y_k\|^2\right) \\
\leq \mathbb{E}\left[f_x^k(x_k) + f_y^k(y_k) - f_x^k(x_{k+1}) - f_y^k(y_{k+1})\right] + \frac{2G^2}{k}
\]
Stochastic Optimization for Non-convex Inf-Projection Problems

Let us consider the first term in the R.H.S of above inequality. For DC functions with \( \text{dom}(h) \subseteq \mathbb{R}^n \), recall \( f^k_x(x) = g(x) - y_k^T(\ell(x_k) + \nabla \ell(x_k)(x - x_k)) \), \( f^k_y(y) = h(y) - y^T \ell(x_{k+1}) \). We have

\[
\begin{align*}
    f^k_x(x_k) + f^k_y(y_k) - f^k_x(x_{k+1}) - f^k_y(y_{k+1}) \\
    = g(x_k) - y_k^T \ell(x_k) + h(y_k) - y_k^T \ell(x_{k+1}) - g(x_{k+1}) + y_k^T \ell(x_k) + y_k^T \ell(x_{k+1}) \nonumber \\
    = h(y_k) - y_k^T \ell(x_k) - (g(x_{k+1}) + h(y_{k+1}) + y_k^T \ell(x_{k+1}) \nonumber \\
    \leq f(x_k, y_k) - f(x_{k+1}, y_{k+1})
\end{align*}
\]

where we use \( y_k \in \mathbb{R}^m \) and the convexity of \( \ell(.) \), i.e., \( \ell(x_k) + \nabla \ell(x_k)(x_{k+1} - x_k) \leq \ell(x_{k+1}) \).

For Bi-convex functions, recall \( f^k_x(x) = g(x) - y_k^T \ell(x), f^k_y(y) = h(y) - y^T \ell(x_{k+1}) \). We have

\[
\begin{align*}
    f^k_x(x_k) + f^k_y(y_k) - f^k_x(x_{k+1}) - f^k_y(y_{k+1}) \\
    = g(x_k) - y_k^T \ell(x_k) + h(y_k) - y_k^T \ell(x_{k+1}) - g(x_{k+1}) + y_k^T \ell(x_k) + y_k^T \ell(x_{k+1}) \nonumber \\
    = g(x_k) + h(y_k) - y_k^T \ell(x_k) - (g(x_{k+1}) + h(y_{k+1}) + y_k^T \ell(x_{k+1}) \nonumber \\
    = f(x_k, y_k) - f(x_{k+1}, y_{k+1})
\end{align*}
\]

Hence, we have

\[
\begin{align*}
    \frac{1}{4}(E[\|v_k - v_k\|^2 + \|x_{k+1} - x_k\|^2 + \|y_k - u_k\|^2 + \|y_{k+1} - y_k\|^2]) \\
    \leq E[f(x_k, y_k) - f(x_{k+1}, y_{k+1})] + \frac{2G^2}{k}
\end{align*}
\] (28)

Next, we can bound the sequence of \( x_k \) and \( y_k \) separately. Let us focus on the sequence of \( x_k \) and the analysis for the sequence of \( y_k \) is similar.

\[
\begin{align*}
    \frac{1}{4}E[\|v_k - v_k\|^2 + \|x_{k+1} - x_k\|^2] \leq E[f(x_k, y_k) - f(x_{k+1}, y_{k+1})] + \frac{2G^2}{k}
\end{align*}
\] (29)

Next dividing \( \gamma \) and then multiplying \( \omega_k \) and on both sides and taking summation over \( k = 1, \ldots, K \) where \( \alpha \geq 1 \), one has

\[
\begin{align*}
    \frac{1}{4}E\left[\sum_{k=1}^{K} \omega_k(\|v_k - v_k\|^2 + \|x_{k+1} - x_k\|^2)\right] \\
    \leq \sum_{k=1}^{K} \frac{\omega_k}{k}E[f(x_k, y_k) - f(x_{k+1}, y_{k+1})] + \frac{2G^2}{4k} \sum_{k=1}^{K} \frac{\omega_k}{k}.
\end{align*}
\] (30)

For the LHS of (30), we have

\[
E[\|v_\tau - v_\tau\|^2 + \|x_{\tau+1} - x_\tau\|^2] = \frac{\sum_{k=1}^{K} \omega_k E[\|v_k - v_k\|^2 + \|x_{k+1} - x_k\|^2]}{\sum_{k=1}^{K} \omega_k},
\]

where \( \tau \) is sampled by \( P(\tau = k) = \frac{k^\alpha}{\sum_{s=1}^{K} s^\alpha} \).

For the RHS of (30), let us consider the first term. According to the setting \( \omega_k = k^\alpha \) with \( \alpha \geq 1 \) and following the similar
analysis of Theorem 2 in (Chen et al., 2018), we have
\[
\sum_{k=1}^{L} \omega_k \big(f(x_k, y_k) - f(x_{k+1}, y_{k+1})\big)
\]
\[
= \sum_{k=1}^{K} (\omega_{k-1} f(x_k, y_k) - \omega_k f(x_{k+1}, y_{k+1})) + \sum_{k=1}^{K} (\omega_k - \omega_{k-1}) f(x_k, y_k)
\]
\[
= \omega_0 f(x_1, y_1) - \omega_K f(x_{K+1}, y_{K+1}) + \sum_{k=1}^{K} (\omega_k - \omega_{k-1}) f(x_k, y_k)
\]
\[
\leq \sum_{k=1}^{K} (\omega_k - \omega_{k-1}) M = M \omega_K = MK^\alpha,
\]
where the third equality is due to \(\omega_0 = 0\) and the inequality is due to Assumption 1 (v). Then for the second term of RHS of (30),
\[
\sum_{k=1}^{K} \frac{\omega_k}{k} \leq \sum_{k=1}^{K} k^{\alpha-1} \leq K K^{\alpha-1} = K^\alpha.
\]
Plugging the above three terms back into (30) and dividing both sides by \(\sum_{k=1}^{K} \omega_k\), we have
\[
\mathbb{E}[\|x_r - v_r\|^2 + \|x_{r+1} - x_r\|^2] \leq \frac{4(M + 2G^2)(\alpha + 1)}{\gamma K}, \quad (31)
\]
due to \(\sum_{k=1}^{K} k^{\alpha} \geq \int_0^K s^{\alpha} ds = \frac{K^{\alpha+1}}{\alpha+1}\).

Similarly, by setting \(\omega_k = k^\alpha\) with \(\alpha \geq 1\) and following the similar analysis of Theorem 2 in (Chen et al., 2018), we have
\[
\mathbb{E}[\|y_r - u_r\|^2 + \|y_{r+1} - y_r\|^2] \leq \frac{4(M + 2G^2)(\alpha + 1)}{\mu K}, \quad (32)
\]
In addition, we have
\[
\mathbb{E}[\|x_{r+1} - v_r\|^2] \leq 2\mathbb{E}[\|x_r - v_r\|^2 + \|x_{r+1} - x_r\|^2] \leq \frac{8(M + 2G^2)(\alpha + 1)}{\gamma K},
\]
\[
\mathbb{E}[\|y_{r+1} - u_r\|^2] \leq 2\mathbb{E}[\|y_r - u_r\|^2 + \|y_{r+1} - u_r\|^2] \leq \frac{8(M + 2G^2)(\alpha + 1)}{\mu K}.
\]

\[\square\]

B.2. Proof of Proposition 2

Proof. This proof is similar to the proof of Proposition 2 in (Xu et al., 2018a). For completeness, we include it here.

Smooth Case. When \(f(z)\) is \(L\)-smooth and \(R(z)\) is \(\gamma\)-strongly convex, we then first have the following lemma from (Zhao & Zhang, 2015).

Lemma 5. Under the same assumptions in Proposition 2, we have
\[
\mathbb{E}[H(z_{t+1}) - H(z)] \leq \frac{\|z_t - z\|^2}{2\eta_t} - \frac{\|z - z_{t+1}\|^2}{2\eta_t} - \frac{\gamma}{2} \|z - z_{t+1}\|^2 + \eta_t \sigma^2.
\]
The proof of this lemma is similar to the analysis to proof of Lemma 1 in (Zhao & Zhang, 2015). Its proof can be found in the analysis of Lemma 7 in (Xu et al., 2018a).

Let us set $w_t = t$, then by Lemma 5 we have

$$
\sum_{t=1}^{T} w_{t+1}(H(z_{t+1}) - H(z)) \\
\leq \sum_{t=1}^{T} \left( \frac{w_{t+1}}{2\eta_t} \|z - z_t\|^2 - \frac{w_{t+1}}{2\eta_t} \|z - z_{t+1}\|^2 - \frac{\gamma w_{t+1}}{2} \|z - z_{t+1}\|^2 \right) + \sum_{t=1}^{T} \eta_t w_{t+1} \sigma^2 \\
\leq \sum_{t=1}^{T} \left( \frac{w_t}{2\eta_{t-1}} - \frac{\gamma w_t}{2} \right) \|z - z_t\|^2 + \frac{w_t/\eta_t + \gamma w_t}{2} \|z - z_t\|^2 + \sum_{t=1}^{T} \eta_t w_{t+1} \sigma^2 \\
\leq \frac{2\gamma}{3} \|z - z_1\|^2 + \sum_{t=1}^{T} \frac{3\sigma^2}{\gamma},
$$

where the last inequality is due to the settings of $\eta_t$ and $w_t$ such that $\frac{w_{t+1}}{\eta_t} - \frac{w_t}{\eta_{t-1}} - \gamma w_t = \frac{2(t+1)^2}{3} - \frac{2t^2}{3} - \gamma t = \frac{2(t+1)^2}{3} \leq 0$.

Then by the convexity of $H = f + R$ and the update of $\hat{z}_T$, we know

$$H(\hat{z}_T) - H(z) \leq \frac{4\gamma \|z - z_1\|^2}{3T(T+3)} + \frac{6\sigma^2}{(T+3)\gamma}.$$

We complete the proof of smooth case by letting $z = z_*$ in above inequality.

**Non-smooth Case.** We then consider the case of $f(z)$ is non-smooth. Recall that the update of $z_{t+1}$ is

$$z_{t+1} = \arg \min_{z \in \Omega} \partial f(z_t; \xi_t) \top z + R(z) + \frac{1}{2\eta_t} \|z - z_t\|^2.$$

By the optimality condition of $z_{t+1}$ and the strong convexity of above objective function, we know for any $z \in \Omega$,

$$\partial f(z_t; \xi_t) \top z + R(z) + \frac{1}{2\eta_t} \|z - z_t\|^2 \geq \partial f(z_t; \xi_t) \top z_{t+1} + R(z_{t+1}) + \frac{1}{2\eta_t} \|z_{t+1} - z_t\|^2 + \frac{1}{2\eta_t + \gamma} \|z - z_{t+1}\|^2,$$

which implies

$$\partial f(z_t; \xi_t) \top (z_t - z) + R(z_{t+1}) - R(z) \leq (z_t - z_{t+1}) \top \partial f(z_t; \xi_t) - \frac{1}{2\eta_t} \|z_{t+1} - z_t\|^2 + \frac{1}{2\eta_t} \|z - z_t\|^2 - \frac{1}{2\eta_t + \gamma} \|z - z_{t+1}\|^2 \leq \frac{\eta_t \|\partial f(z_t; \xi_t)\|^2}{2} + \frac{1}{2\eta_t} \|z - z_t\|^2 - \frac{1}{2\eta_t + \gamma} \|z - z_{t+1}\|^2.$$

Taking expectation on both sides of above inequality and using the convexity of $f(z)$, then we get

$$E[f(z_t) - f(z) + R(z_{t+1}) - R(z)] \leq \frac{\eta_t \sigma^2}{2} + E \left[ \frac{1}{2\eta_t} \|z - z_t\|^2 - \frac{1}{2\eta_t + \gamma} \|z - z_{t+1}\|^2 \right].$$

Multiplying both sides of above inequality by $w_t = t$ and taking summation over $t = 1, \ldots, T$, then

$$E \left[ \sum_{t=1}^{T} w_t(f(z_t) - f(z) + R(z_{t+1}) - R(z)) \right] \leq \sum_{t=1}^{T} 2\sigma^2 w_t \eta_t + \sum_{t=1}^{T} \frac{w_t/\eta_t + w_t \gamma}{2} \|z - z_{t+1}\|^2.$$
Stochastic Optimization for Non-convex Inf-Projection Problems

We rewrite above inequality, then
\[
E \left[ \sum_{t=1}^{T} w_t(f(z_t) - f(z) + R(z_t) - R(z)) \right] \\
\leq E \left[ \sum_{t=1}^{T} w_t(R(z_t) - R(z_{t+1})) \right] + \sum_{t=1}^{T} 2\sigma^2 w_t \eta_t + E \left[ \sum_{t=1}^{T} \left( \frac{w_t - \frac{w_{t-1}}{\eta_{t-1}} + w_{t-1}\gamma}{2} \right) \|z - z_t\|^2 \right] \\
\leq E \left[ \sum_{t=1}^{T} w_t(R(z_t) - R(z_{t+1})) \right] + \frac{8\sigma^2 T}{\gamma} + \frac{\gamma z - z_1\|^2}{8},
\]

where the last inequality is due to \( w_t = t, \eta_t = 1/(\gamma t), \) and \( \frac{w_t}{2\eta_t} - \frac{w_{t-1}/\eta_{t-1} + w_{t-1}\gamma}{2} \leq 0, \forall t \geq 2. \) The let us consider the first term, we have
\[
E \left[ \sum_{t=1}^{T} w_t(f(z_t) - f(z) + R(z_t) - R(z)) \right] \\
\leq w_0 R(z_1) - w_T R(z_{T+1}) + E \left[ \sum_{t=1}^{T} (w_t - w_{t-1}) R(z_t) \right] + \frac{8\sigma^2 T}{\gamma} + \frac{\gamma z - z_1\|^2}{8} \\
= E \left[ \sum_{t=1}^{T} (w_t - w_{t-1}) R(z_t) \right] + \frac{8\sigma^2 T}{\gamma} + \frac{\gamma z - z_1\|^2}{8}. \quad (33)
\]

Next, we want to show for any \( z_t \) we have
\[
E[\|z_t - z_1\|^2] \leq \frac{\sigma^2}{\gamma^2}. \quad (34)
\]

We prove it by induction. It is easy to show that the inequality (34) holds for \( t = 1. \) We then assume the inequality (34) holds for \( t. \) By the update of \( z_{t+1} = \arg\min_{z \in \Omega} \partial f(z; \xi_t)^T z + \frac{1}{2}\|z - z_1\|^2 + \frac{1}{\eta_t} \|z - z_t\|^2 = \arg\min_{z \in \Omega} \frac{1}{2}\|z - z_{t+1}\|^2, \)
where \( z_{t+1} = \frac{\gamma z + \frac{1}{\eta_t} z - \partial f(z; \xi_t)}{\gamma + \frac{1}{\eta_t}}. \) Then
\[
E[\|z_{t+1} - z_1\|^2] \leq E[\|z_{t+1} - z_1\|^2] = \frac{1}{(\gamma + \frac{1}{\eta_t})^2} E \left[ \| \frac{z_{t+1} - z_1}{\eta_t} - \partial f(z; \xi_t) \|^2 \right] \\
\leq \frac{1}{(\gamma + \frac{1}{\eta_t})^2} \left( \frac{1 + \gamma \eta_t}{\eta_t^2} E[\|z_t - z_1\|^2] + \left( 1 + \frac{1}{\gamma \eta_t} \right) E[\|\partial f(z_t; \xi_t)\|^2] \right) \\
\leq \frac{1}{(\gamma + \frac{1}{\eta_t})^2} \left( \frac{1 + \gamma \eta_t}{\eta_t^2} \sigma^2 + \left( 1 + \frac{1}{\gamma \eta_t} \right) \sigma^2 \right) = \frac{\sigma^2}{\gamma^2}.
\]

Then by induction we know the inequality (34) holds for all \( t \geq 1. \) Combining inequalities (33) and (34) we get
\[
E \left[ \sum_{t=1}^{T} w_t(f(z_t) - f(z) + R(z_t) - R(z)) \right] \\
\leq E \left[ \sum_{t=1}^{T} (w_t - w_{t-1}) \sigma^2 \right] + \frac{8\sigma^2 T}{\gamma} + \frac{\gamma z - z_1\|^2}{8} \\
= \frac{17\sigma^2 T}{2\gamma} + \frac{\gamma z - z_1\|^2}{8}.
\]

Then by the convexity of \( H = f + R \) and the update of \( \hat{z}_T, \) we know
\[
E \left[ H(\hat{z}_T) - H(z) \right] \leq \frac{17\sigma^2}{\gamma (T+1)} + \frac{\gamma z - z_1\|^2}{4T(T+1)}.
\]

We complete the proof of non-smooth case by letting \( z = z_\ast \) in above inequality.
B.3. Proof of Lemma 4

Part I. We consider $g$ is $L_g$-smooth and $\ell$ is $G_\ell$-Lipschitz continuous. Due to the first order optimality of $f^k_\ell(x)$ at $v_k$ (and smoothness of $f^k_\ell(x)$),

$$0 = \nabla g(v_k) - \nabla \ell(x_k)y_k + \gamma(v_k - x_k)$$

$$= \nabla g(x_k) - \nabla \ell(x_k)y^*(x_k) + \gamma(v_k - x_k)$$

$$+ \nabla g(v_k) - \nabla g(x_k) + \nabla \ell(x_k)(y^*(x_k) - y_k)$$

$$= \nabla F(x_k) + \gamma(v_k - x_k)$$

$$+ \nabla g(v_k) - \nabla g(x_k) + \nabla \ell(x_k)(y^*(x_k) - y_k),$$

where $y^*(x_k) = \arg\min_{y \in \text{dom}(h)} h(y) - \ell(x_k)^Ty$. Let $\tilde{f}^k_\ell(y) = h(y) - \ell(x_k)^Ty$. The second equality is due to Theorem 10.13 of (Rockafellar & Wets, 2009) and the uniqueness of $y^*(x_k)$ ($h$ is uniformly convex).

To bound $\|\nabla F(x_k)\|$ we have,

$$\|\nabla F(x_k)\| \leq \gamma \|v_k - x_k\| + \|\nabla g(v_k) - \nabla g(x_k)\| + \|\nabla \ell(x_k)\| \cdot \|y^*(x_k) - u_k + u_k - y_k\|$$

$$\leq \gamma \|v_k - x_k\| + L_g \|v_k - x_k\| + G_\ell \|y^*(x_k) - u_k\| + G_\ell \|u_k - y_k\|.$$

To handle $\Box$, we could use the $(\varrho, p)$-uniform convexity of $\tilde{f}^k_\ell$ (since $\nabla h^*$ is assumed to be $(L_{h^*}, v)$-Hölder continuous) as follows

$$\|y^*(x_k) - u_k\|^{p-1} \leq \frac{1}{\varrho} \|\partial \tilde{f}^k_\ell(y^*(x_k)) - \partial \tilde{f}^k_\ell(u_k)\|$$

$$= \frac{1}{\varrho} \|\partial h(u_k) - \ell(x_k)\|$$

$$\leq \frac{1}{\varrho} \left( \|\partial h(u_k) - \ell(x_{k+1})\| + \|\ell(x_{k+1}) - \ell(x_k)\| \right)$$

$$\leq \frac{1}{\varrho} \|\mu(u_k - y_k)\| + \frac{G_\ell}{\varrho} \|x_{k+1} - x_k\|,$$

where the first inequality is due to (4). The first equality is due to the first order optimality of $\tilde{f}^k_\ell(y)$ at $y^*(x_k)$, i.e., $0 \in \partial \tilde{f}^k_\ell(y^*(x_k))$. The third inequality is due to the first order optimality of $\tilde{f}^k_\ell(y) + R^k_\ell(y)$ at $u_k$, i.e., $0 \in \partial h(u_k) - \ell(x_{k+1}) + \mu(u_k - y_k)$.

Since $\varrho = \frac{2\varphi}{1-v}$ and $v = \frac{1}{p-1}$ (Lemma 1), one has $\|y^*(x_k) - u_k\| \leq \mu^v \left( \frac{1+v}{2v} \right)^{\varphi} L_{h^*} \|u_k - y_k\|^{v} + G_\ell^v \left( \frac{1+v}{2v} \right)^{\varphi} L_{h^*} \|u_k - y_k\|^{v}$.

Therefore, $\|\nabla F(x_k)\| \leq \gamma \|v_k - x_k\| + L_g \|v_k - x_k\|$

$$+ G_\ell \mu^v \left( \frac{1+v}{2v} \right)^{\varphi} L_{h^*} \|u_k - y_k\|^{v} + G_\ell^v \left( \frac{1+v}{2v} \right)^{\varphi} L_{h^*} \|x_{k+1} - x_k\|^{v} + G_\ell \|u_k - y_k\|.$$

Part II. We consider $g$ is non-smooth and $\ell$ is $G_\ell$-Lipschitz continuous and $L_\ell$-smooth and $\max_{y \in \text{dom}(h)} \|y\| \leq D$. Due to the first order optimality of $f^k_\ell$ at $v_k$,

$$0 \in \partial g(v_k) - \nabla \ell(x_k)y_k + \gamma(v_k - x_k)$$

$$= \partial g(v_k) - \nabla \ell(v_k)y^*(v_k) + \gamma(v_k - x_k)$$

$$+ \nabla g(v_k) - \nabla g(x_k) + \nabla \ell(x_k)(y^*(x_k) - y_k)$$

$$= \partial F(v_k) + \gamma(v_k - x_k)$$

$$+ \nabla g(v_k) - \nabla g(x_k) + \nabla \ell(x_k)(y^*(x_k) - y_k),$$

The second equality is due to Theorem 10.13 of (Rockafellar & Wets, 2009) and the uniqueness of $y^*(v_k)$ ($h$ is uniformly convex).
Therefore, by $G_\ell$-Lipschitz continuity of $\ell$, $L_\ell$-smoothness of $\ell$ and $\max_{y \in \text{dom}(h)} \|y\| \leq D_y$, we have
\[
\text{dist}(0, \partial F(v_k)) \leq \gamma \|v_k - x_k\| + G_\ell \|y^*(v_k) - y_k\| + D_y L_\ell \|v_k - x_k\|
\]
\[
\leq \gamma \|v_k - x_k\| + G_\ell \|u_k - y_k\| + G_\ell \|y^*(v_k) - u_k\| + D_y L_\ell \|v_k - x_k\|.
\]

To deal with $\triangledown$, one could employ $(\varrho, p)$-uniform convexity of $\hat{f}^k_y = h(y) - \ell(v_k)^\top y$,
\[
\|y^*(v_k) - u_k\|^{p-1} \leq \frac{1}{\varrho} \|\partial h(u_k) - \ell(v_k)\|
\]
\[
\leq \frac{1}{\varrho} \|\partial h(u_k) - \ell(x_{k+1})\| + \|\ell(x_{k+1}) - \ell(v_k)\|
\]
\[
\leq \frac{\mu}{\varrho} \|u_k - y_k\| + \frac{G_\ell}{\varrho} \|x_{k+1} - v_k\|,
\]
where the first inequality is due to $(\varrho, p)$-uniform convexity of $h$ and the first order optimality of $\hat{f}^k_y$ at $y^*(v_k)$. The last inequality is due to the first order optimality of $f^k_y + R^k_y$ at $u_k$, i.e., $0 \in \partial h(u_k) - \ell(x_{k+1}) + \mu(u_k - y_k)$, and $G_\ell$-Lipschitz continuity of $\ell$.

Since $\nabla h^*$ is $(L^*_h, v)$-Hölder continuous by assumption, by Lemma 1, $\varrho = \frac{2v}{1 + \varphi} \left( \frac{1}{L_h^*} \right)^{\frac{1}{v}}$. Then one has
\[
\|y^*(v_k) - u_k\| \leq \left( \frac{1 + v}{2v} \right)^v L^*_h \left( \mu \|u_k - y_k\| + G_\ell \|x_{k+1} - v_k\| \right)^v.
\]

Therefore, one has
\[
\text{dist}(0, \partial F(v_k)) \leq \gamma \|v_k - x_k\| + G_\ell \|u_k - y_k\|
\]
\[
+ G_\ell \left( \frac{1 + v}{2v} \right)^v L^*_h \left( \mu \|u_k - y_k\| + G_\ell \|x_{k+1} - v_k\| \right)^v + DL_\ell \|v_k - x_k\|.
\]