A. Proof of Convergence of Continuous-time Dynamics

In this section, we provide a detailed Lyapunov function based analysis for the convergence of continuous-time dynamics (3.9).

Proof of Theorem 4.1. Applying Itô Formula (Itô, 1944) to the Lyapunov function in (4.1) yields

$$\mathrm{d}\mathcal{E}_t = \frac{\partial \mathcal{E}_t}{\partial t}\mathrm{d}t + \left\langle \frac{\partial \mathcal{E}_t}{\partial \boldsymbol{X}_t}, \mathrm{d}\boldsymbol{X}_t \right\rangle + \left\langle \frac{\partial \mathcal{E}_t}{\partial \boldsymbol{Y}_t}, \mathrm{d}\boldsymbol{Y}_t \right\rangle + \frac{\delta \dot{\beta}_t^2}{2\mu^2}\operatorname{tr}\left(\boldsymbol{\sigma}_t^\top \frac{\partial^2 \mathcal{E}_t}{\partial \boldsymbol{Y}_t^2}\boldsymbol{\sigma}_t\right)\mathrm{d}t, \tag{A.1}$$

where we use $\boldsymbol{\sigma}_t$ to make the expression $\boldsymbol{\sigma}(\boldsymbol{X}_t, t)$ compact when no confusion arises. Note that we have

$$\frac{\partial \mathcal{E}_t}{\partial t} = \dot{\beta}_t e^{\beta_t}(f(\boldsymbol{X}_t) - f(\mathbf{x}^*) + \mu D_h(\mathbf{x}^*, \nabla h^*(\boldsymbol{Y}_t))),$$
$$\frac{\partial \mathcal{E}_t}{\partial \boldsymbol{X}_t} = e^{\beta_t}\nabla f(\boldsymbol{X}_t), \quad \frac{\partial \mathcal{E}_t}{\partial \boldsymbol{Y}_t} = \mu e^{\beta_t}(\nabla h^*(\boldsymbol{Y}_t) - \mathbf{x}^*).$$

Thus, combining the above partial derivatives with the SDE in (3.9), we obtain

$$\left\langle \frac{\partial \mathcal{E}_t}{\partial \boldsymbol{X}_t}, \mathrm{d}\boldsymbol{X}_t \right\rangle = \dot{\beta}_t e^{\beta_t}\Big\langle \nabla f(\boldsymbol{X}_t), \nabla h^*(\boldsymbol{Y}_t) - \boldsymbol{X}_t\rangle,$$

and

$$\left\langle \frac{\partial \mathcal{E}_t}{\partial \boldsymbol{Y}_t}, \mathrm{d}\boldsymbol{Y}_t \right\rangle = -\dot{\beta}_t e^{\beta_t}\langle \nabla f(\boldsymbol{X}_t)\mathrm{d}t + \mu(\boldsymbol{Y}_t - \nabla h(\boldsymbol{X}_t)) + \sqrt{\delta}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{B}_t, \nabla h^*(\boldsymbol{Y}_t) - \mathbf{x}^*\rangle.$$

Submitting the above equations back into (A.1) yields

$$\begin{aligned}
\mathrm{d}\mathcal{E}_t &= \dot{\beta}_t e^{\beta_t}[f(\boldsymbol{X}_t) - f(\mathbf{x}^*) + \langle \nabla f(\boldsymbol{X}_t), \mathbf{x}^* - \boldsymbol{X}_t\rangle]\mathrm{d}t + \dot{\beta}_t e^{\beta_t}\mu D_h(\mathbf{x}^*, \nabla h^*(\boldsymbol{Y}_t))\mathrm{d}t \\
&\quad - \mu\dot{\beta}e^{\beta_t}\langle \nabla h^*(\boldsymbol{Y}_t) - \mathbf{x}^*, \boldsymbol{Y}_t - \nabla h(\boldsymbol{X}_t)\rangle\mathrm{d}t - \dot{\beta}e^{\beta_t}\langle \nabla h^*(\boldsymbol{Y}_t) - \mathbf{x}^*, \boldsymbol{\sigma}_t\mathrm{d}\boldsymbol{B}_t\rangle + \frac{\delta\dot{\beta}_t^2}{2\mu^2}\operatorname{tr}\left(\boldsymbol{\sigma}_t^\top\frac{\partial^2\mathcal{E}_t}{\partial\boldsymbol{Y}_t^2}\boldsymbol{\sigma}_t\right)\mathrm{d}t \\
&\le \mu\dot{\beta}_t e^{\beta_t}[-D_h(\mathbf{x}^*, \boldsymbol{X}_t) + D_h(\mathbf{x}^*, \nabla h^*(\boldsymbol{Y}_t))]\mathrm{d}t - \mu\dot{\beta}_t e^{\beta_t}\langle \nabla h^*(\boldsymbol{Y}_t) - \mathbf{x}^*, \boldsymbol{Y}_t - \nabla h(\boldsymbol{X}_t)\rangle\mathrm{d}t \\
&\quad - \dot{\beta}e^{\beta_t}\langle \nabla h^*(\boldsymbol{Y}_t) - \mathbf{x}^*, \boldsymbol{\sigma}_t\mathrm{d}\boldsymbol{B}_t\rangle + \frac{\delta\dot{\beta}_t^2}{2\mu^2}\operatorname{tr}\left(\boldsymbol{\sigma}_t^\top\frac{\partial^2\mathcal{E}_t}{\partial\boldsymbol{Y}_t^2}\boldsymbol{\sigma}_t\right)\mathrm{d}t,
\end{aligned}$$

where the inequality is due to the strong convexity of f. By three point identity (Chen & Teboulle, 1993), we have

$$D_h(\mathbf{x}^*, \nabla h^*(\boldsymbol{Y}_t)) + D_h(\nabla h^*(\boldsymbol{Y}_t), \boldsymbol{X}_t) - D_h(\mathbf{x}^*, \boldsymbol{X}_t) = \langle \nabla h(\boldsymbol{X}_t) - \boldsymbol{Y}_t, \mathbf{x}^* - \nabla h^*(\boldsymbol{Y}_t)\rangle.$$

Therefore, we obtain

$$\begin{aligned}
\mathrm{d}\mathcal{E}_t &\le -\mu\dot{\beta}_t e^{\beta_t}D_h(\nabla h^*(\boldsymbol{Y}_t), \boldsymbol{X}_t)\mathrm{d}t - \dot{\beta}e^{\beta_t}\langle \nabla h^*(\boldsymbol{Y}_t) - \mathbf{x}^*, \boldsymbol{\sigma}_t\mathrm{d}\boldsymbol{B}_t\rangle + \frac{\delta\dot{\beta}_t^2}{2\mu^2}\operatorname{tr}\left(\boldsymbol{\sigma}_t^\top\frac{\partial^2\mathcal{E}_t}{\partial\boldsymbol{Y}_t^2}\boldsymbol{\sigma}_t\right)\mathrm{d}t \\
&\le -\dot{\beta}e^{\beta_t}\langle \nabla h^*(\boldsymbol{Y}_t) - \mathbf{x}^*, \boldsymbol{\sigma}_t\mathrm{d}\boldsymbol{B}_t\rangle + \frac{\delta\dot{\beta}_t^2}{2\mu^2}\operatorname{tr}\left(\boldsymbol{\sigma}_t^\top\frac{\partial^2\mathcal{E}_t}{\partial\boldsymbol{Y}_t^2}\boldsymbol{\sigma}_t\right)\mathrm{d}t.
\end{aligned}$$

Integrating $\mathrm{d}\mathcal{E}_t$ from 0 to t, we obtain

$$\mathcal{E}_t \le \mathcal{E}_0 - \int_0^t \dot{\beta}e^{\beta_r}\langle \nabla h^*(\boldsymbol{Y}_r) - \mathbf{x}^*, \boldsymbol{\sigma}_r\mathrm{d}\boldsymbol{B}_r\rangle + \int_0^t \frac{\delta\dot{\beta}_r^2 e^{\beta_r}}{2\mu}\operatorname{tr}\left(\boldsymbol{\sigma}_r^\top\nabla^2 h^*(\boldsymbol{Y}_r)\boldsymbol{\sigma}_r\right)\mathrm{d}r.$$

Taking expectation and by the property of Brownian motion, we have

$$\mathbb{E}[\mathcal{E}_t] \le \mathcal{E}_0 + \mathbb{E}\left[\int_0^t \frac{\delta\dot{\beta}_r^2 e^{\beta_r}}{2\mu}\operatorname{tr}\left(\boldsymbol{\sigma}_r^\top\nabla^2 h^*(\boldsymbol{Y}_r)\boldsymbol{\sigma}_r\right)\mathrm{d}r\right].$$

By the definition of $\mathcal{E}_t$ and the non-negativity of D_{h^*}, we obtain the upper bound on the expected primal function value gap

$$\mathbb{E}[f(\boldsymbol{X}_t) - f(\mathbf{x}^*)] \le e^{-\beta_t}\mathcal{E}_0 + e^{-\beta_t}\mathbb{E}\left[\int_0^t \frac{\delta\dot{\beta}_r^2 e^{\beta_r}}{2\mu}\operatorname{tr}\left(\boldsymbol{\sigma}_r^\top\nabla^2 h^*(\boldsymbol{Y}_r)\boldsymbol{\sigma}_r\right)\mathrm{d}r\right],$$

which completes the proof. □

B. Proof of Convergence for Discrete-time Algorithms

We provide proofs of convergence rates of the proposed Algorithms here. For the shorthand of notation, we use $\Delta(\mathbf{x}) = G(\mathbf{x};\xi) - \nabla f(\mathbf{x})$ to denote the bias between the stochastic gradient and the full gradient, where $\mathbf{x} \in \mathbb{R}^d$ and ξ is a random vector.

B.1. Proof of Theorem 5.1

Proof of Theorem 5.1. For the ease of presentation, we define $\mathcal{D}_k = f(\mathbf{x}_k) - f(\mathbf{x}^*) + \mu D_{h^*}(\mathbf{y}_k, \nabla h(\mathbf{x}^*))$. By the definition of Lyapunov function in (5.4), we have $\mathcal{E}_k = A_k \mathcal{D}_k$. Then

$$\begin{aligned}
\mathbb{E}[\mathcal{D}_{k+1} - \mathcal{D}_k] &= \mathbb{E}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) + \mu D_{h^*}(\mathbf{y}_{k+1}, \nabla h(\mathbf{x}^*)) - \mu D_{h^*}(\mathbf{y}_k, \nabla h(\mathbf{x}^*)] \\
&= \mathbb{E}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) + \mu\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1}), \mathbf{y}_k - \mathbf{y}_{k+1}\rangle - \mu D_{h^*}(\mathbf{y}_k, \mathbf{y}_{k+1})] \\
&= \mathbb{E}\bigg[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) + \frac{A_{k+1} - A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1}), \nabla f(\mathbf{x}_{k+1})\rangle\bigg] \\
&\quad + \mathbb{E}\bigg[\frac{A_{k+1} - A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1}), \mu\big(\mathbf{y}_{k+1} - \nabla h(\mathbf{x}_{k+1})\big) + \Delta(\mathbf{x}_{k+1})\rangle - \mu D_{h^*}(\mathbf{y}_k, \mathbf{y}_{k+1})\bigg] \\
&= \mathbb{E}\bigg[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) + \frac{A_{k+1} - A_k}{A_k}\langle \mathbf{x}^* - \mathbf{x}_{k+1}, \nabla f(\mathbf{x}_{k+1})\rangle + \langle \mathbf{x}_k - \mathbf{x}_{k+1}, \nabla f(\mathbf{x}_{k+1})\rangle\bigg] \\
&\quad + \mathbb{E}\bigg[\frac{A_{k+1} - A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1}), \mu\big(\mathbf{y}_{k+1} - \nabla h(\mathbf{x}_{k+1})\big) + \Delta(\mathbf{x}_{k+1})\rangle - \mu D_{h^*}(\mathbf{y}_k, \mathbf{y}_{k+1})\bigg],
\end{aligned}$$

where the second equality is due to the three point identity, the third equality follows from (5.3b) and the last one follows from (5.3a). By strong convexity of f we have

$$\begin{aligned}
\langle \mathbf{x}^* - \mathbf{x}_{k+1}, \nabla f(\mathbf{x}_{k+1})\rangle &\le f(\mathbf{x}^*) - f(\mathbf{x}_{k+1}) - \mu D_h(\mathbf{x}^*, \mathbf{x}_{k+1}), \\
\langle \mathbf{x}_k - \mathbf{x}_{k+1}, \nabla f(\mathbf{x}_{k+1})\rangle &\le f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) - \mu D_h(\mathbf{x}_k, \mathbf{x}_{k+1}),
\end{aligned}$$

which immediately implies

$$\begin{aligned}
&\mathbb{E}[\mathcal{D}_{k+1} - \mathcal{D}_k] \\
&\le \mathbb{E}\bigg[\frac{A_{k+1} - A_k}{A_k}\big(f(\mathbf{x}^*) - f(\mathbf{x}_{k+1}) - \mu D_h(\mathbf{x}^*, \mathbf{x}_{k+1})\big) - \mu D_h(\mathbf{x}_k, \mathbf{x}_{k+1})\bigg] \\
&\quad + \mathbb{E}\bigg[\frac{A_{k+1} - A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1}), \mu\big(\mathbf{y}_{k+1} - \nabla h(\mathbf{x}_{k+1})\big) + \Delta(\mathbf{x}_{k+1})\rangle - \mu D_{h^*}(\mathbf{y}_k, \mathbf{y}_{k+1})\bigg] \\
&= \mathbb{E}\bigg[\frac{A_{k+1} - A_k}{A_k}\big(-\mathcal{D}_{k+1} + \mu D_{h^*}(\mathbf{y}_{k+1}, \nabla h(\mathbf{x}^*)) - \mu D_h(\mathbf{x}^*, \mathbf{x}_{k+1})\big) - \mu D_h(\mathbf{x}_k, \mathbf{x}_{k+1})\bigg] \\
&\quad + \mathbb{E}\bigg[\frac{A_{k+1} - A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1}), \mu\big(\mathbf{y}_{k+1} - \nabla h(\mathbf{x}_{k+1})\big) + \Delta(\mathbf{x}_{k+1})\rangle - \mu D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k))\bigg] \\
&\le \frac{A_{k+1} - A_k}{A_k}\mathbb{E}[-\mathcal{D}_{k+1} + \mu D_{h^*}(\mathbf{y}_{k+1}, \nabla h(\mathbf{x}^*)) - \mu D_h(\mathbf{x}^*, \mathbf{x}_{k+1})] - \frac{\mu\mu_h}{2}\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\|^2 \\
&\quad + \frac{A_{k+1} - A_k}{A_k}\mathbb{E}\big[\mu\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1}), \mathbf{y}_{k+1} - \nabla h(\mathbf{x}_{k+1})\rangle + \langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1}), \Delta(\mathbf{x}_{k+1})\rangle\big] \\
&= \frac{A_{k+1} - A_k}{A_k}\mathbb{E}\big[-\mathcal{D}_{k+1} - \mu D_h(\nabla h^*(\mathbf{y}_{k+1}), \mathbf{x}_{k+1})\big] \\
&\quad - \frac{\mu\mu_h}{2}\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\|^2 + \frac{A_{k+1} - A_k}{A_k}\mathbb{E}\big[\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1}), \Delta(\mathbf{x}_{k+1})\rangle\big] \\
&\le \frac{A_{k+1} - A_k}{A_k}\mathbb{E}[-\mathcal{D}_{k+1}] + \frac{A_{k+1} - A_k}{A_k}\mathbb{E}[\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1}), \Delta(\mathbf{x}_{k+1})\rangle] - \frac{\mu\mu_h}{2}\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\|^2,
\end{aligned}$$

where the second inequality is due to the strong convexity of h and the last equality follows from three point identity. We now give an upper bound of the last term in the above inequality. Note that $\mathbf{y}_k$ is independent of $\Delta(\mathbf{x}_{k+1})$ according to

(5.3b) and that $\mathbb{E}[\Delta(\mathbf{x}_{k+1})] = \mathbf{0}$, then

$$\begin{aligned}
&\frac{A_{k+1}-A_k}{A_k}\mathbb{E}[\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1}), \Delta(\mathbf{x}_{k+1})\rangle] - \frac{\mu\mu_h}{2}\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\|^2 \\
&= \frac{A_{k+1}-A_k}{A_k}\mathbb{E}[\langle \nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1}), \Delta(\mathbf{x}_{k+1})\rangle] - \frac{\mu\mu_h}{2}\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\|^2 \\
&\le \frac{(A_{k+1}-A_k)^2\sigma^2}{2\mu\mu_h A_k^2},
\end{aligned}$$

where the last inequality use the fact that $bx - ax^2/2 \le b^2/(2a)$ for $a > 0$ and $\mathbb{E}[\|\Delta(\mathbf{x}_{k+1})\|^2|\mathbf{x}_{k+1}] \le \sigma^2$. This immediately yields

$$\mathbb{E}[A_{k+1}\mathcal{D}_{k+1} - A_k\mathcal{D}_k] \le \frac{(A_{k+1}-A_k)^2}{A_k}\frac{\sigma^2}{2\mu\mu_h}.$$

Plugging the definition that $\mathcal{E}_k = A_k\mathcal{D}_k$ into the above inequality, and summing from 0 to $k-1$, we obtain

$$\mathbb{E}[\mathcal{E}_k] \le \mathcal{E}_0 + \frac{\sigma^2}{2\mu\mu_h}\sum_{i=0}^{k-1}\frac{(A_{i+1}-A_i)^2}{A_i}.$$

Choose $A_k = k(k+1)/2$ and $A_0 = 1$, we can see that the step size of Algorithm 1 is $1/k$. Applying Lemma D.3 and the definition of $\mathcal{E}_k$ yields the following convergence rate

$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] \le \frac{\mathbb{E}[\mathcal{E}_k]}{A_k} \le \frac{2\mathcal{E}_0}{k(k+1)} + \frac{2\sigma^2}{(k+1)\mu\mu_h}.$$

□

B.2. Proof of Theorem 5.3

We first lay down the following technical lemma about the gradient bound of f over $\mathcal{X}$.

Lemma B.1. Suppose f is μ-strongly convex and L-smooth. Then we have

$$\|\nabla f(\mathbf{x})\|_* \le \frac{L\sqrt{2M_{h,\mathcal{X}}}}{\sqrt{\mu_h}} + \|\nabla f(\mathbf{x}^*)\|_*,$$

where $M_{h,\mathcal{X}} = \sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} D_h(\mathbf{x},\mathbf{x}')$.

Proof of Theorem 5.3. Recall the Lyapunov function in (5.4) and the definition that $\mathcal{E}_k = A_k\mathcal{D}_k$. Similar to the proof of Theorem 5.1 we have

$$\begin{aligned}
\mathbb{E}[\mathcal{D}_{k+1} - \mathcal{D}_k] &= \mathbb{E}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) + \mu D_{h^*}(\mathbf{y}_{k+1}, \nabla h(\mathbf{x}^*)) - \mu D_{h^*}(\mathbf{y}_k, \nabla h(\mathbf{x}^*)] \\
&= \mathbb{E}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) + \mu\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_k), \mathbf{y}_k - \mathbf{y}_{k+1}\rangle + \mu D_{h^*}(\mathbf{y}_{k+1}, \mathbf{y}_k)] \\
&= \mathbb{E}\left[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) + \frac{A_{k+1}-A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_k), \nabla f(\mathbf{x}_{k+1})\rangle\right] \\
&\quad + \mathbb{E}\left[\frac{A_{k+1}-A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_k), \mu\big(\mathbf{y}_k - \nabla h(\mathbf{x}_{k+1})\big) + \Delta(\mathbf{x}_{k+1})\rangle + \mu D_{h^*}(\mathbf{y}_{k+1}, \mathbf{y}_k)\right] \\
&= \mathbb{E}\left[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) + \frac{A_{k+1}-A_k}{A_k}\langle \mathbf{x}^* - \mathbf{x}_{k+1}, \nabla f(\mathbf{x}_{k+1})\rangle + \langle \mathbf{x}_k - \mathbf{x}_{k+1}, \nabla f(\mathbf{x}_{k+1})\rangle\right] \\
&\quad + \mathbb{E}\left[\frac{A_{k+1}-A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_k), \mu\big(\mathbf{y}_k - \nabla h(\mathbf{x}_{k+1})\big) + \Delta(\mathbf{x}_{k+1})\rangle + \mu D_{h^*}(\mathbf{y}_{k+1}, \mathbf{y}_k)\right],
\end{aligned}$$

where the second equation comes from the three point identity, the third and fourth equations are due to (5.5b) and (5.5a) respectively. Note that h is μ_h-strongly convex and L_h-smooth. Without loss of generality, we assume that $L_h = 1$;

otherwise, we can replace h with $h' = h/L_h$ which is μ_h/L_h-strongly convex and 1-smooth. By strong convexity of f we have

$$\begin{aligned}
\mathbb{E}[\mathcal{D}_{k+1} - \mathcal{D}_k] &\le \mathbb{E}\Big[\frac{A_{k+1}-A_k}{A_k}\big(f(\mathbf{x}^*) - f(\mathbf{x}_{k+1}) - \mu D_h(\mathbf{x}^*, \mathbf{x}_{k+1})\big) - \mu D_h(\mathbf{x}_k, \mathbf{x}_{k+1})\Big] \\
&\quad + \mathbb{E}\Big[\frac{A_{k+1}-A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_k), \mu\big(\mathbf{y}_k - \nabla h(\mathbf{x}_{k+1})\big) + \Delta(\mathbf{x}_{k+1})\rangle + \mu D_{h^*}(\mathbf{y}_{k+1}, \mathbf{y}_k)\Big] \\
&\le \frac{A_{k+1}-A_k}{A_k}\mathbb{E}\big[-\mathcal{D}_{k+1} + \mu\big(D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) - D_h(\mathbf{x}^*, \mathbf{x}_{k+1})\big)\big] \\
&\quad + \mathbb{E}\Big[\frac{A_{k+1}-A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_k), \mu\big(\mathbf{y}_k - \nabla h(\mathbf{x}_{k+1})\big) + \Delta(\mathbf{x}_{k+1})\rangle + \mu D_{h^*}(\mathbf{y}_{k+1}, \mathbf{y}_k)\Big] \\
&= \frac{A_{k+1}-A_k}{A_k}\mathbb{E}\Big[-\mathcal{D}_{k+1} + \mu\big(\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_{k+1}), \nabla h(\mathbf{x}_{k+1}) - \mathbf{y}_{k+1}\rangle - D_h(\nabla h^*(\mathbf{y}_{k+1}), \mathbf{x}_{k+1})\big)\Big] \\
&\quad + \mathbb{E}\Big[\frac{A_{k+1}-A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_k), \mu\big(\mathbf{y}_k - \nabla h(\mathbf{x}_{k+1})\big) + \Delta(\mathbf{x}_{k+1})\rangle + \mu D_{h^*}(\mathbf{y}_{k+1}, \mathbf{y}_k)\Big] \\
&= -\frac{A_{k+1}-A_k}{A_k}\mathcal{D}_{k+1} + \mu\mathbb{E}\Big[D_{h^*}(\mathbf{y}_{k+1}, \mathbf{y}_k) + \frac{A_{k+1}-A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_k), \Delta(\mathbf{x}_{k+1})\rangle\Big] \\
&\quad + \mu\frac{A_{k+1}-A_k}{A_k}\mathbb{E}[\langle \nabla h(\mathbf{x}_{k+1}) - \mathbf{y}_{k+1}, \nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\rangle - D_{h^*}(\nabla h(\mathbf{x}_{k+1}), \mathbf{y}_{k+1})] \\
&\quad + \mu\frac{A_{k+1}-A_k}{A_k}\mathbb{E}[\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_k), \nabla h(\mathbf{x}_{k+1}) - \mathbf{y}_{k+1}\rangle + \langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_k), \mathbf{y}_k - \nabla h(\mathbf{x}_{k+1})\rangle] \\
&\le -\frac{A_{k+1}-A_k}{A_k}\mathcal{D}_{k+1} + \mu\mathbb{E}\Big[\frac{A_{k+1}-A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_k), \mathbf{y}_k - \mathbf{y}_{k+1}\rangle + D_{h^*}(\mathbf{y}_{k+1}, \mathbf{y}_k)\Big] \\
&\quad + \mu\frac{A_{k+1}-A_k}{A_k}\mathbb{E}[\|\nabla h(\mathbf{x}_{k+1}) - \mathbf{y}_{k+1}\|_*\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\| - \frac{1}{2}\|\nabla h(\mathbf{x}_{k+1}) - \mathbf{y}_{k+1}\|_*^2]
\end{aligned}$$

where the second inequality follows from the definition of $\mathcal{D}_{k+1}$ and that $D_h(\mathbf{x}_k, \mathbf{x}_{k+1})$ is non-negative, the first equality is due to the three point identity, and the last inequality is due to the $1/\mu_h$-smoothness of h^*, Cauchy's inequality and the fact that $\mathbf{y}_k$ is independent of $\Delta(\mathbf{x}_{k+1})$ and $\mathbb{E}[\Delta(\mathbf{x}_{k+1})] = \mathbf{0}$. For the last term in the above inequality, we have

$$\begin{aligned}
&\|\nabla h(\mathbf{x}_{k+1}) - \mathbf{y}_{k+1}\|_*\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\| - \frac{1}{2}\|\nabla h(\mathbf{x}_{k+1}) - \mathbf{y}_{k+1}\|_*^2 \\
&= -\frac{1}{2}\big(\|\nabla h(\mathbf{x}_{k+1}) - \mathbf{y}_{k+1}\|_* - \|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\|\big)^2 + \frac{1}{2}\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\|^2 \\
&\le \frac{1}{2}\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\|^2.
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\mathbb{E}[\mathcal{D}_{k+1} - \mathcal{D}_k] &\le -\frac{A_{k+1}-A_k}{A_k}\mathcal{D}_{k+1} + \mu\mathbb{E}\Big[\frac{A_{k+1}-A_k}{A_k}\langle \mathbf{x}^* - \nabla h^*(\mathbf{y}_k), \mathbf{y}_k - \mathbf{y}_{k+1}\rangle + D_{h^*}(\mathbf{y}_{k+1}, \mathbf{y}_k)\Big] \\
&\quad + \mu\frac{A_{k+1}-A_k}{A_k}\mathbb{E}\Big[\frac{1}{2}\|\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\|^2\Big] \\
&\le -\frac{A_{k+1}-A_k}{A_k}\mathcal{D}_{k+1} + \mu\mathbb{E}\Big[\frac{A_{k+1}-A_k}{A_k}\|\mathbf{x}^* - \nabla h^*(\mathbf{y}_k)\|\|\mathbf{y}_{k+1} - \mathbf{y}_k\|_* + \frac{1}{2\mu_h}\|\mathbf{y}_{k+1} - \mathbf{y}_k\|_*^2\Big] \\
&\quad + \mu\frac{A_{k+1}-A_k}{A_k}\mathbb{E}\Big[\frac{1}{2\mu_h^2}\|\mathbf{y}_{k+1} - \mathbf{y}_k\|_*^2\Big] \\
&\le -\frac{A_{k+1}-A_k}{A_k}\mathcal{D}_{k+1} + \mu\mathbb{E}\Big[\frac{(A_{k+1}-A_k)\sqrt{2M_{h,\mathcal{X}}}}{\sqrt{\mu_h}A_k}\|\mathbf{y}_{k+1} - \mathbf{y}_k\|_* + \frac{1}{2\mu_h}\|\mathbf{y}_{k+1} - \mathbf{y}_k\|_*^2\Big] \\
&\quad + \mu\frac{A_{k+1}-A_k}{A_k}\mathbb{E}\Big[\frac{1}{2\mu_h^2}\|\mathbf{y}_{k+1} - \mathbf{y}_k\|_*^2\Big],
\end{aligned} \tag{B.1}$$

where the second inequality follows from Cauchy's inequality and the $1/\mu_h$-smoothness of h^*. Further by (5.5b) we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{y}_{k+1}-\mathbf{y}_k\|_*] &= \frac{A_{k+1}-A_k}{A_k}\mathbb{E}\left[\left\|\frac{1}{\mu}(\nabla f(\mathbf{x}_{k+1})+\Delta(\mathbf{x}_{k+1}))+\mathbf{y}_k-\nabla h(\mathbf{x}_{k+1})\right\|_2\right]\\ &\le \frac{A_{k+1}-A_k}{\mu A_k}\mathbb{E}[\|\nabla f(\mathbf{x}_{k+1})\|_*+\|\Delta(\mathbf{x}_{k+1})\|_*+\mu\|\mathbf{y}_k-\nabla h(\mathbf{x}_{k+1})\|_*]\\ &\le \frac{(A_{k+1}-A_k)}{\mu A_k}\left(\frac{L\sqrt{2M_{h,\mathcal{X}}}}{\sqrt{\mu_h}}+\|\nabla f(\mathbf{x}^*)\|_*+\sigma+\mu\mathbb{E}[\|\mathbf{y}_k-\nabla h(\mathbf{x}_{k+1})\|_*]\right),\end{aligned} \tag{B.2}$$

where the second inequality is due to Lemma B.1. By (5.5a) we have that

$$\begin{aligned}\|\mathbf{y}_k-\nabla h(\mathbf{x}_{k+1})\|_* &= \left\|\nabla h\left(\mathbf{x}_{k+1}+\frac{A_k}{A_{k+1}-A_k}(\mathbf{x}_{k+1}-\mathbf{x}_k)\right)-\nabla h(\mathbf{x}_{k+1})\right\|_*\\ &\le \frac{A_k}{A_{k+1}-A_k}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|_2\\ &= \|\nabla h^*(\mathbf{y}_k)-\mathbf{x}_{k+1}\|_2\\ &\le \sqrt{\frac{2M_{h,\mathcal{X}}}{\mu_h}}.\end{aligned} \tag{B.3}$$

Substituting (B.1) with and (B.2) and (B.3) yields

$$\begin{aligned}\mathbb{E}[\mathcal{D}_{k+1}-\mathcal{D}_k] &\le -\frac{A_{k+1}-A_k}{A_k}\mathcal{D}_{k+1}+\frac{\mu\sqrt{2M_{h,\mathcal{X}}}(A_{k+1}-A_k)}{\sqrt{\mu_h}A_k}\mathbb{E}[\|\mathbf{y}_{k+1}-\mathbf{y}_k\|_*]+\frac{\mu}{2\mu_h}\mathbb{E}[\|\mathbf{y}_{k+1}-\mathbf{y}_k\|_2^2]\\ &\quad+\frac{\mu(A_{k+1}-A_k)}{2\mu_h^2A_k}\mathbb{E}[\|\mathbf{y}_{k+1}-\mathbf{y}_k\|_*^2]\\ &\le -\frac{A_{k+1}-A_k}{A_k}\mathcal{D}_{k+1}+\frac{\sqrt{2M_{h,\mathcal{X}}}(A_{k+1}-A_k)^2}{\sqrt{\mu_h}A_k^2}C_0+\frac{(A_{k+1}-A_k)^2}{2\mu\mu_hA_k^2}C_0^2\\ &\quad+\frac{(A_{k+1}-A_k)^3}{2\mu\mu_h^2A_k^3}C_0^2\\ &\le -\frac{A_{k+1}-A_k}{A_k}\mathcal{D}_{k+1}+\frac{\sqrt{2M_{h,\mathcal{X}}}(A_{k+1}-A_k)^2}{\sqrt{\mu_h}A_k^2}C_0+\frac{(A_{k+1}-A_k)^2}{2\mu\mu_h^2A_k^2}\left(\mu_h+\frac{A_{k+1}-A_k}{A_k}\right)C_0^2,\end{aligned}$$

where $C_0=(L+\mu)\sqrt{2M_{h,\mathcal{X}}/\mu_h}+\sigma+\|\nabla f(\mathbf{x}^*)\|_*$. Multiply A_k to both sides and we get

$$\mathbb{E}[A_{k+1}\mathcal{D}_{k+1}-A_k\mathcal{D}_k]\le\frac{(A_{k+1}-A_k)^2}{2\mu_h^2\mu A_k}\left[\left(\mu_h+\frac{A_{k+1}-A_k}{A_k}\right)C_0^2+2\sqrt{2M_{h,\mathcal{X}}\mu_h^3}\mu C_0\right].$$

Note that we have $\mathcal{E}_k=A_k\mathcal{D}_k$. Setting $A_k=k(k+1)/2$, $A_0=1$ and summing from 0 to $k-1$ yields

$$\mathcal{E}_k\le\mathcal{E}_0+\frac{(\mu_h+2)C_0^2+2\sqrt{2M_{h,\mathcal{X}}\mu_h^3}\mu C_0}{2\mu_h^2\mu}\sum_{j=0}^{k-1}\frac{(A_{j+1}-A_j)^2}{A_j}.$$

Finally, we plug in the definition of $\mathcal{E}_k$ to get

$$\mathbb{E}[f(\mathbf{x}_k)-f(\mathbf{x}^*)]\le\frac{2\mathcal{E}_0}{k(k+1)}+\frac{(\mu_h+2)C_0^2+2\sqrt{2M_{h,\mathcal{X}}\mu_h^3}\mu C_0}{\mu_h^2\mu(k+1)}.$$

□

B.3. Proof of Theorem 5.5

Proof of Theorem 5.5. Since we have $\mathcal{E}_k = A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*) + \mu D_h(\mathbf{x}^*, \nabla h(\mathbf{y}_k)))$ by (5.4). We have

$$\begin{aligned}
\mathbb{E}[\mathcal{E}_{k+1} - \mathcal{E}_k] &= \mathbb{E}[A_{k+1}f(\mathbf{x}_{k+1}) - A_k f(\mathbf{x}_k) - (A_{k+1} - A_k)f(\mathbf{x}^*)] \\
&\quad + \mathbb{E}[\mu A_{k+1} D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) - \mu A_k D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))] \\
&= \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{z}_{k+1})) + A_k(f(\mathbf{z}_{k+1}) - f(\mathbf{x}_k)) + (A_{k+1} - A_k)(f(\mathbf{z}_{k+1}) - f(\mathbf{x}^*))] \\
&\quad + \mathbb{E}[\mu A_{k+1} D_h(, \mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) - \mu A_k D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))] \\
&\le \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{z}_{k+1})) + A_k[\langle \nabla f(\mathbf{z}_{K+1}), \mathbf{z}_{K+1} - \mathbf{x}_k\rangle - \mu D_h(\mathbf{x}_k, \mathbf{z}_{k+1})]] \\
&\quad + \mathbb{E}[(A_{k+1} - A_k)[\langle \nabla f(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}^*\rangle - \mu D_h(\mathbf{x}^*, \mathbf{z}_{k+1})]] \\
&\quad + \mathbb{E}[\mu A_{k+1}[D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))] + \mu(A_{k+1} - A_k) D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))] \\
&= \mathbb{E}[A_{k+1}(f(\mathbf{x}_{k+1}) - f(\mathbf{z}_{k+1})) + (A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle] \\
&\quad + \mathbb{E}[-\mu A_k D_h(\mathbf{x}_k, \mathbf{z}_{k+1}) + \mu(A_{k+1} - A_k)[D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \mathbf{z}_{k+1})]] \\
&\quad + \mathbb{E}[\mu A_{k+1}[D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k))]], \qquad \text{(B.4)}
\end{aligned}$$

where the inequality is due to the strong convexity of f and the last equation follows from (5.6a). Denote $C_k = (A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle$, and then we proceed to bound C_k as follows.

$$\begin{aligned}
C_k &= (A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\rangle + (A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_{k+1} - \mathbf{x}^*\rangle \\
&= (A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\rangle + \mu A_{k+1}\langle \mathbf{y}_k - \mathbf{y}_{k+1}, \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^*\rangle \\
&\quad + \mu(A_{k+1} - A_k)\langle \nabla h(\mathbf{z}_{k+1}) - \mathbf{y}_k, \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^*\rangle - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^*\rangle \\
&= (A_{k+1} - A_k)\langle \nabla f(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\rangle - \mu A_{k+1} D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k)) \\
&\quad + \mu A_{k+1}[D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1})] - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^*\rangle \\
&\quad + \mu(A_{k+1} - A_k)\langle \nabla h(\mathbf{z}_{k+1}) - \mathbf{y}_k, \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^*\rangle \\
&\le (A_{k+1} - A_k)\langle \nabla \widetilde{f}(\mathbf{z}_{k+1}; \xi_{k+1}), \nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1})\rangle \\
&\quad - \mu(A_{k+1} - A_k) D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k)) - \frac{\mu\mu_h A_k}{2}\|\nabla h^*(\mathbf{y}_{k+1}) - \nabla h^*(\mathbf{y}_k)\|_2^2 \\
&\quad + \mu A_{k+1}[D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle \\
&\quad + \mu(A_{k+1} - A_k)\langle \nabla h(\mathbf{z}_{k+1}) - \mathbf{y}_k, \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^*\rangle, \qquad \text{(B.5)}
\end{aligned}$$

where the second equality is due to (5.6b), the third equality follows from the three point identity and the inequality is a result of the strong convexity of h. To make it brief, we define

$$\mathbf{w} = \frac{A_{k+1} - A_k}{A_{k+1}}\nabla h^*(\mathbf{y}_{k+1}) + \frac{A_k}{A_{k+1}}\mathbf{x}_k. \qquad \text{(B.6)}$$

We immediately have that $\mathbf{z}_{k+1} - \mathbf{w} = (A_{k+1} - A_k)/A_{k+1}(\nabla h^*(\mathbf{y}_k) - \nabla h^*(\mathbf{y}_{k+1}))$ by (5.6a). Therefore we can further bound C_k in the following way.

$$\begin{aligned}
C_k &\le A_{k+1}\langle \nabla \widetilde{f}(\mathbf{z}_{k+1}; \xi_{k+1}), \mathbf{z}_{k+1} - \mathbf{w}\rangle - \frac{\mu\mu_h A_{k+1}^2 A_k}{2(A_{k+1} - A_k)^2}\|\mathbf{z}_{k+1} - \mathbf{w}\|_2^2 \\
&\quad + \mu A_{k+1}[D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle \\
&\quad - \mu(A_{k+1} - A_k) D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k)) + \mu(A_{k+1} - A_k)\langle \nabla h(\mathbf{z}_{k+1}) - \mathbf{y}_k, \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^*\rangle \\
&\le A_{k+1}\langle \nabla \widetilde{f}(\mathbf{z}_{k+1}; \xi_{k+1}), \mathbf{z}_{k+1} - \mathbf{w}\rangle - \frac{\mu\mu_h A_{k+1}^2 A_k}{(A_{k+1} - A_k)^2} D_h(\mathbf{z}_{k+1}, \mathbf{w}) \\
&\quad + \mu A_{k+1}[D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle \\
&\quad - \mu(A_{k+1} - A_k) D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k)) + \mu(A_{k+1} - A_k)\langle \nabla h(\mathbf{z}_{k+1}) - \mathbf{y}_k, \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^*\rangle \\
&\le A_{k+1}\langle \nabla \widetilde{f}(\mathbf{z}_{k+1}; \xi_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle - \frac{\mu\mu_h^2 A_{k+1}^2 A_k}{2(A_{k+1} - A_k)^2}\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|_2^2 \\
&\quad + \mu A_{k+1}[D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] - (A_{k+1} - A_k)\langle \Delta(\mathbf{z}_{k+1}), \nabla h^*(\mathbf{y}_k) - \mathbf{x}^*\rangle \\
&\quad - \mu(A_{k+1} - A_k) D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k)) + \mu(A_{k+1} - A_k)\langle \nabla h(\mathbf{z}_{k+1}) - \mathbf{y}_k, \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^*\rangle, \qquad \text{(B.7)}
\end{aligned}$$

where in the second inequality we use the smoothness of h and in the last inequality we use (5.6c) and the strong convexity of h. Take expectation and note that $\Delta(\mathbf{z}_{k+1})$ is independent of $\mathbf{y}_k$, we have

$$\begin{aligned}
\mathbb{E}[C_k] \leq \mathbb{E}\Bigg[& A_{k+1}\langle \nabla f(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle - \frac{\mu\mu_h^2 A_{k+1}^2 A_k}{2(A_{k+1}-A_k)^2}\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|_2^2\Bigg] \\
& + \mathbb{E}[\mu A_{k+1}[D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] + A_{k+1}\langle \Delta(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle] \\
& + \mathbb{E}[-\mu(A_{k+1}-A_k)D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k)) + \mu(A_{k+1}-A_k)\langle \nabla h(\mathbf{z}_{k+1}) - \mathbf{y}_k, \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^*\rangle] \\
\leq \mathbb{E}\Bigg[& A_{k+1}\left(f(\mathbf{z}_{k+1}) - f(\mathbf{x}_{k+1}) + \frac{L}{2}\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|_2^2\right) - \frac{\mu\mu_h^2 A_{k+1}^2 A_k}{2(A_{k+1}-A_k)^2}\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|_2^2\Bigg] \\
& + \mathbb{E}[\mu A_{k+1}[D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_{k+1}))] + A_{k+1}\langle \Delta(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle] \\
& + \mathbb{E}[-\mu(A_{k+1}-A_k)D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k)) + \mu(A_{k+1}-A_k)\langle \nabla h(\mathbf{z}_{k+1}) - \mathbf{y}_k, \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^*\rangle],
\end{aligned}$$

where the second inequality follows from Assumption 3.4. Submitting the above inequality back into B.4 yields

$$\begin{aligned}
& \mathbb{E}[\mathcal{E}_{k+1} - \mathcal{E}_k] \\
& \leq \mathbb{E}\left[\left(\frac{LA_{k+1}}{2} - \frac{\mu\mu_h^2 A_{k+1}^2 A_k}{2(A_{k+1}-A_k)^2}\right)\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|^2 - \mu A_k D_h(\mathbf{x}_k, \mathbf{z}_{k+1})\right] \\
& \quad + \mathbb{E}[\mu(A_{k+1}-A_k)[D_h(\mathbf{x}^*, \nabla h^*(\mathbf{y}_k)) - D_h(\mathbf{x}^*, \mathbf{z}_{k+1})] - A_{k+1}\langle \Delta(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle] \\
& \quad + \mu(A_{k+1}-A_k)\mathbb{E}[-D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k)) + \langle \nabla h(\mathbf{z}_{k+1}) - \mathbf{y}_k, \nabla h^*(\mathbf{y}_{k+1})) - \mathbf{x}^*\rangle] \\
& = \mathbb{E}\left[\left(\frac{LA_{k+1}}{2} - \frac{\mu\mu_h^2 A_{k+1}^2 A_k}{2(A_{k+1}-A_k)^2}\right)\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|_2^2 - A_{k+1}\langle \Delta(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle - \mu A_k D_h(\mathbf{x}_k, \mathbf{z}_{k+1})\right] \\
& \quad + \mu(A_{k+1}-A_k)\mathbb{E}[\langle \nabla h(\mathbf{z}_{k+1}) - \mathbf{y}_k, \mathbf{x}^* - \nabla h^*(\mathbf{y}_k)\rangle - D_h(\nabla h^*(\mathbf{y}_k), \mathbf{z}_{k+1})] \\
& \quad + \mu(A_{k+1}-A_k)\mathbb{E}[-D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k)) + \langle \nabla h(\mathbf{z}_{k+1}) - \mathbf{y}_k, \nabla h^*(\mathbf{y}_{k+1}) - \mathbf{x}^*\rangle] \\
& \leq \mathbb{E}\left[\left(\frac{LA_{k+1}}{2} - \frac{\mu\mu_h^2 A_{k+1}^2 A_k}{2(A_{k+1}-A_k)^2}\right)\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|_2^2 - A_{k+1}\langle \Delta(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle\right] \\
& \quad - \mu(A_{k+1}-A_k)\mathbb{E}[D_h(\nabla h^*(\mathbf{y}_k), \mathbf{z}_{k+1}) + D_h(\nabla h^*(\mathbf{y}_{k+1}), \nabla h^*(\mathbf{y}_k))] \\
& \quad + \mu(A_{k+1}-A_k)\mathbb{E}[\langle \nabla h(\mathbf{z}_{k+1}) - \mathbf{y}_k, \nabla h^*(\mathbf{y}_{k+1}) - \nabla h^*(\mathbf{y}_k)\rangle] \\
& = \mathbb{E}\left[\left(\frac{LA_{k+1}}{2} - \frac{\mu\mu_h^2 A_{k+1}^2 A_k}{2(A_{k+1}-A_k)^2}\right)\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|_2^2 - A_{k+1}\langle \Delta(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle\right] \\
& \quad - \mu(A_{k+1}-A_k)\mathbb{E}[D_h(\nabla h^*(\mathbf{y}_{k+1}), \mathbf{z}_{k+1})],
\end{aligned}$$

where the first equation follows from the three point identity, in the last inequality we drop the negative term $-\mu A_k D_h(\mathbf{x}_k, \mathbf{z}_{k+1})$ and the last equation again follows from the three point identity. We further rearrange the above inequality and obtain

$$\begin{aligned}
\mathbb{E}[\mathcal{E}_{k+1} - \mathcal{E}_k] & \leq \mathbb{E}\left[\left(\frac{LA_{k+1}}{2} - \frac{\mu\mu_h^2 A_{k+1}^2 A_k}{2(A_{k+1}-A_k)^2}\right)\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|_2^2 - A_{k+1}\langle \Delta(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle\right] \\
& \quad - \mu(A_{k+1}-A_k)\mathbb{E}[D_h(\nabla h^*(\mathbf{y}_{k+1}), \mathbf{z}_{k+1})] \\
& \leq \underbrace{\mathbb{E}\left[\frac{A_{k+1}}{2}\left(L - \frac{\mu\mu_h^2 A_{k+1}}{(A_{k+1}-A_k)^2}\right)\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|_2^2\right]}_{S_1} \\
& \quad + \underbrace{\mathbb{E}\left[-\frac{\mu\mu_h^2 A_{k+1}^2 (A_k - 1)}{2(A_{k+1}-A_k)^2}\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|_2^2 - A_{k+1}\langle \Delta(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{x}_{k+1}\rangle\right]}_{S_2},
\end{aligned} \tag{B.8}$$

where we used the fact that $D_h(\nabla h^*(\mathbf{y}_{k+1}), \mathbf{z}_{k+1}) \geq 0$. We choose $A_k = \mu\mu_h^2(k+1)(k+2)/(4L) + 1$ and thus

$$S_1 \leq \mathbb{E}\left[\frac{A_{k+1}}{2}\left(L - L\frac{k+3}{k+2}\right)\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|_2^2\right] \leq 0. \tag{B.9}$$

For S_2, using Cauchy-Schwartz inequality yields

$$\begin{aligned} S_2 &\leq \mathbb{E}\left[A_{k+1}\|\Delta(\mathbf{z}_{k+1})\|_2\|\mathbf{z}_{k+1}-\mathbf{x}_{k+1}\|_2 - \frac{\mu\mu_h^2 A_{k+1}^2(A_k-1)}{2(A_{k+1}-A_k)^2}\|\mathbf{z}_{k+1}-\mathbf{x}_{k+1}\|_2^2\right] \\ &\leq \mathbb{E}\left[\frac{(A_{k+1}-A_k)^2}{2\mu\mu_h^2(A_k-1)}\|\Delta(\mathbf{z}_{k+1})\|_2^2\right] \\ &\leq \frac{(A_{k+1}-A_k)^2\sigma^2}{2\mu\mu_h^2(A_k-1)}, \end{aligned} \tag{B.10}$$

where for the second inequality we use the simple inequality that $bx - ax^2/2 \leq b^2/2a,\ \forall a > 0$ and the last inequality follows from the fact that $\mathbb{E}[\|\Delta(\mathbf{z}_{k+1})\|_2^2] \leq \sigma^2$. Then we submit (B.9) and (B.10) back into (B.8) and obtain

$$\mathbb{E}[\mathcal{E}_{k+1} - \mathcal{E}_k] \leq \frac{(A_{k+1}-A_k)^2\sigma^2}{2\mu\mu_h^2(A_k-1)}.$$

Summing up the above inequality from 0 to $k-1$ yields

$$\mathbb{E}[\mathcal{E}_k] \leq A_0\mathcal{E}_0 + \frac{\sigma^2}{2L}\sum_{i=0}^{k-1}\frac{i+2}{i+1} \leq A_0\mathcal{E}_0 + \frac{\sigma^2 k}{L}.$$

Recall the definition that $\mathcal{E}_k = A_k(f(\mathbf{x}_k) - f(\mathbf{x}^*) + \mu D_h(\mathbf{x}^*, \nabla h(\mathbf{y}_k)))$, we have

$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] \leq \frac{\mathcal{E}_k}{A_k} \leq \frac{4LA_0\mathcal{E}_0 + 4\sigma^2 k}{\mu\mu_h^2(k+1)(k+2)},$$

which completes the proof. □

C. Proof of Technical Lemmas

In this section, we provide the proof of technical lemmas used in the proof of main theorems.

Proof of Lemma B.1. By smoothness of f we have

$$\|\nabla f(\mathbf{x})\|_* = \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|_* + \|\nabla f(\mathbf{x}^*)\|_* \leq L\|\mathbf{x}-\mathbf{x}^*\| + \|\nabla f(\mathbf{x}^*)\|_*.$$

By the strong convexity of h we have

$$D_h(\mathbf{x},\mathbf{x}') = h(\mathbf{x}) - h(\mathbf{x}') - \langle\nabla h(\mathbf{x}'), \mathbf{x}-\mathbf{x}'\rangle \geq \frac{\mu_h}{2}\|\mathbf{x}-\mathbf{x}'\|^2.$$

Combining the two inequalities above yields

$$\|\nabla f(\mathbf{x})\|_* \leq \frac{\sqrt{2}L}{\sqrt{\mu_h}}\sqrt{D_h(\mathbf{x},\mathbf{x}^*)} + \|\nabla f(\mathbf{x}^*)\|_* \leq \frac{L\sqrt{2M_{h,\mathcal{X}}}}{\sqrt{\mu_h}} + \|\nabla f(\mathbf{x}^*)\|_*,$$

where $M_{h,\mathcal{X}} = \sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} D_h(\mathbf{x},\mathbf{x}')$. □

D. Auxiliary Lemmas

The following standard results on duality and Bregman divergence are wildly used in the analysis of mirror descent (Lan, 2012; Ghadimi & Lan, 2012; Krichene et al., 2015; Krichene & Bartlett, 2017). Detailed discussion on the properties of Bregman divergence can be found in Banerjee et al. (2005).

Lemma D.1. If h is μ_h-strongly convex for some constant $\mu_h > 0$, then its conjugate function h^* is $1/\mu_h$-smooth.

Lemma D.2. Suppose h is strongly convex, for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, it holds that

$$\nabla h^*(\nabla h(\mathbf{x}))) = \mathbf{x}, \quad D_h(\mathbf{x},\mathbf{x}') = D_{h^*}(\nabla h(\mathbf{x}'), \nabla h(\mathbf{x})).$$

The following lemma characterizes the upper bound of p-series $\sum_{j=1}^{k} 1/j^p$.

Lemma D.3. (Chlebus, 2009) For $p < 0$, the divergence rate of p-series is given by

$$1 + \frac{k^{1-p} - 1}{1-p} \leq \sum_{j=1}^{k} \frac{1}{j^p} \leq \frac{(k+1)^{1-p} - 1}{1-p}.$$

Lemma D.4 (Three Point Identity (Chen & Teboulle, 1993))**.** Let $D_h(\cdot, \cdot)$ be a Bregman divergence with distance generating function h. For any $\mathbf{a}, \mathbf{b}$ that are interior points of dom h and $\mathbf{c} \in$ dom h, we have

$$D_h(\mathbf{c}, \mathbf{a}) + D_h(\mathbf{a}, \mathbf{b}) - D_h(\mathbf{c}, \mathbf{b}) = \langle \nabla h(\mathbf{b}) - \nabla h(\mathbf{a}), \mathbf{c} - \mathbf{a} \rangle.$$