A. Illustration of Algorithm 2

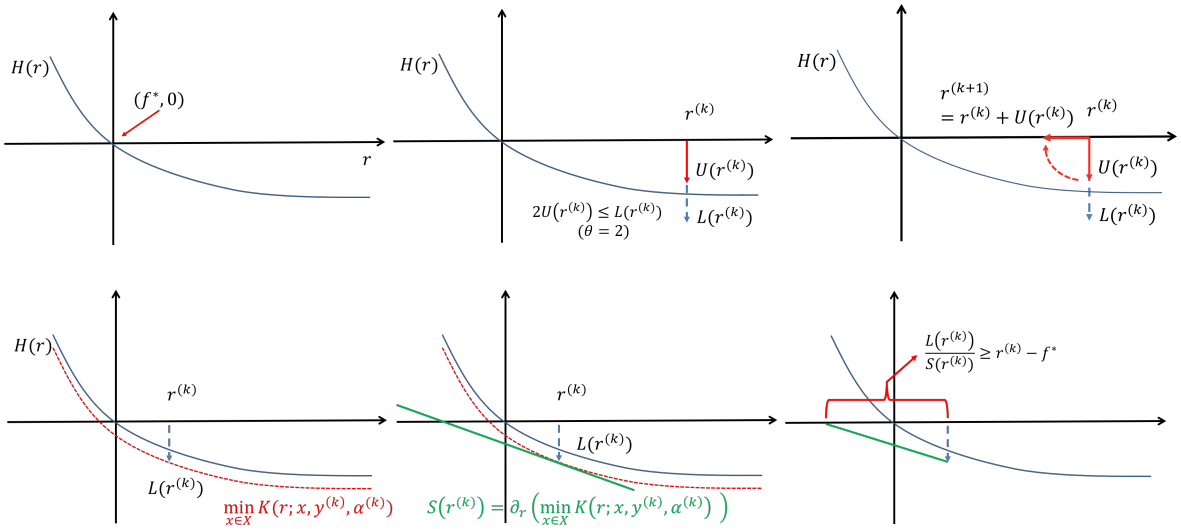

Figure 3. Illustration of AM-FLS Method. The figures on the top row depict the procedure to update $r^{(k)}$ using upper bound $U(r^{(k)})$. The figures on the bottom row show when to stop the algorithm.

The geometric illustration of Algorithm 1 has already been given in Aravkin et al. (2016). In Figure 3, we illustrate the intuition behind Algorithm 2. We choose $\theta = 2$ as an example. In the top-left picture in Figure 3, we plot the curve of a level function $H(r)$ that has all the properties in Lemma 1. Moreover, the x-axis represents the value of r and the point where the x-axis intersecting with the y-axis is $(f^*, H(f^*)) = (f^*, 0)$. In the top-middle picture, we consider a level parameter $r^{(k)} > f^*$ such that $H(r^{(k)}) < 0$, and use an oracle to find $U(r^{(k)})$ and $L(r^{(k)})$ such that $2U(r^{(k)}) \leq L(r^{(k)}) \leq H(r^{(k)}) \leq U(r^{(k)})$ (Property 4 in Definition 1 of an oracle with $\theta = 2$). In the top-right figure, we perform the update $r^{(k+1)} \leftarrow r^{(k)} + U(r^{(k)})$ such that $r^{(k)}$ moves towards the root f^* of $H(r)$ as k increases. Note that, in Algorithm 2, we use a slightly different updating step which is $r^{(k+1)} \leftarrow r^{(k)} + U(r^{(k)})/2$. This is because the multiplier $\frac{1}{2}$ (or any multiplier less than 1) applied to $U(r^{(k)})$ can avoid the extreme scenario where $r^{(k+1)} = f^*$. We want to avoid this scenario because, if it happens, we can no longer find $\bar{\mathbf{x}}$ such that $\mathcal{P}(r^{(k+1)}; \bar{\mathbf{x}}) < 0$ and thus cannot ensure the feasibility of the returned solution. The impact of this multiplier to the complexity of a feasible level-set method is analyzed by Lin et al. (2017).

In the bottom-left figure, we plot the curve (of r) $\min_{\mathbf{x}\in\mathcal{X}} K(r; \mathbf{x}, \mathbf{y}^{(k)}, \boldsymbol{\alpha}^{(k)})$ in red where $(\mathbf{y}^{(k)}, \boldsymbol{\alpha}^{(k)}) = \mathbf{w}^{(k)}$ is the dual solution found by the oracle when it solves (7). According to (7), $\min_{\mathbf{x}\in\mathcal{X}} K(r; \mathbf{x}, \mathbf{y}^{(k)}, \boldsymbol{\alpha}^{(k)})$ is a global lower bound of $H(r)$ and $L(r^{(k)}) = \min_{\mathbf{x}\in\mathcal{X}} K(r^{(k)}; \mathbf{x}, \mathbf{y}^{(k)}, \boldsymbol{\alpha}^{(k)})$. In the bottom-middle figure, we construct the tangent line for the curve $\min_{\mathbf{x}\in\mathcal{X}} K(r; \mathbf{x}, \mathbf{y}^{(k)}, \boldsymbol{\alpha}^{(k)})$ at $r^{(k)}$, namely, $L(r^{(k)}) + \partial_r(\min_{\mathbf{x}\in\mathcal{X}} K(r^{(k)}; \mathbf{x}, \mathbf{y}^{(k)}, \boldsymbol{\alpha}^{(k)}))(r - r^{(k)})$ which is the green line in this figure. Therefore, we can choose $S(r^{(k)}) = \partial_r(\min_{\mathbf{x}\in\mathcal{X}} K(r^{(k)}; \mathbf{x}, \mathbf{y}^{(k)}, \boldsymbol{\alpha}^{(k)}))$ as the slope in the output of the oracle, which will satisfy Property 5 in Definition 1. Finally, in the bottom-right picture, we show a line segment in the x-axis whose length is $\frac{L(r^{(k)})}{S(r^{(k)})}$ which is no shorter than $r^{(k)} - f^*$. Hence, to ensure $r^{(k)} - f^* \leq \varepsilon$, it suffices to stop Algorithm 2 when $\frac{L(r^{(k)})}{S(r^{(k)})} \leq \varepsilon$, or equivalently, $L(r^{(k)}) \geq \varepsilon S(r^{(k)})$.

B. Proof of Lemma 3

Proof. According to the update step in Algorithm 4, we have, for $t \geq 0$,

$$\mathbf{w}^{(t+1)} = (\mathbf{y}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}) \in \arg\min_{\mathbf{w}\in\mathcal{W}} -\boldsymbol{\alpha}^\top \mathbf{v}^{(t)} + G_\mu(\mathbf{w}) + \frac{D(\mathbf{w}, \mathbf{w}^{(t)})}{\tau}. \tag{15}$$

By Proposition 1, we have $\mathbf{y}^{(t+1)} \in \text{int}\Delta$ and $\boldsymbol{\alpha}^{(t+1)} = y_i^{(t+1)}\tilde{\boldsymbol{\alpha}}_i^{(t+1)}$ where

$$\tilde{\boldsymbol{\alpha}}_i^{(t+1)} \in \underset{\tilde{\boldsymbol{\alpha}}_i \in \mathbb{R}^{n_i}}{\arg\min} \left\{ \nu \left\|\tilde{\boldsymbol{\alpha}}_i\right\|_2^2 + \frac{1}{\tau}\left\|\tilde{\boldsymbol{\alpha}}_i - \tilde{\boldsymbol{\alpha}}_i^{(t)}\right\|_2^2 + \sum_{j=1}^{n_i} \frac{1}{n_i}\phi_{ij}^*\big(\tilde{\alpha}_{ij}\big) - \tilde{\boldsymbol{\alpha}}_i^\top \mathbf{v}_i^{(t)} \right\}. \tag{16}$$

Therefore, to prove this lemma, it suffices to prove $\|\tilde{\boldsymbol{\alpha}}_i^{(t)}\|_2 \le B$ for all $t \ge 0$ and $i = 0, 1, \ldots, m$. We prove this result under each of the two scenarios in Assumption 2.

Suppose scenario (b) in Assumption 2 holds such that $B \ge \max\limits_{\tilde{\alpha}_{ij} \in \text{dom}\phi_{ij}} \|\tilde{\boldsymbol{\alpha}}_i\|_2$. Since $\tilde{\boldsymbol{\alpha}}_i^{(t)}$ must stay in the domain of ϕ_{ij}^* according to (16), we have $\|\tilde{\boldsymbol{\alpha}}_i^{(t)}\|_2 \le B$ for all $t \ge 0$ and $i = 0, 1, \ldots, m$.

In the next, we prove this result by assuming scenario (a) in Assumption 2 holds such that B is a constant that satisfies

$$B \ge \max\left\{ 2\left\|\tilde{\boldsymbol{\alpha}}_i^*\right\|_2, \frac{8d\max_k \|\Theta_{ik}\|_2 B_{\mathbf{x}}}{\gamma}, 2\left\| \frac{\bar{\boldsymbol{\alpha}}_i^{(0)}}{\bar{y}_i^{(0)}} - \tilde{\boldsymbol{\alpha}}_i^* \right\|_2 \right\}.$$

Let $\tilde{\boldsymbol{\alpha}}_i^{(t)} = \frac{\boldsymbol{\alpha}_i^{(t)}}{y_i^{(t)}}$ and $\tilde{\boldsymbol{\alpha}}_i^* = \frac{\boldsymbol{\alpha}_i^*}{y_i^*}$ for $i = 0, 1, \ldots, m$. We will first prove

$$\|\tilde{\boldsymbol{\alpha}}_i^{(t)} - \tilde{\boldsymbol{\alpha}}_i^*\|_2 \le \max\left\{ \left\|\tilde{\boldsymbol{\alpha}}_i^*\right\|_2, \frac{4d\max_k \|\Theta_{ik}\|_2 B_{\mathbf{x}}}{\gamma}, \|\bar{\boldsymbol{\alpha}}_i^{(0)}/\bar{y}_i^{(0)} - \tilde{\boldsymbol{\alpha}}_i^*\|_2 \right\} \tag{17}$$

for all $t \ge 0$ by induction over the index t. Equation (17) holds trivially for $t = 0$ because $\tilde{\boldsymbol{\alpha}}_i^{(0)} = \bar{\boldsymbol{\alpha}}_i^{(0)}/\bar{y}_i^{(0)}$. Now, we assume (17) holds for iteration t and prove it also holds for iteration $t+1$.

According to (16), we can independently update each coordinate of $\tilde{\boldsymbol{\alpha}}_i^{(t+1)}$, denoted by $\tilde{\alpha}_{ij}^{(t+1)}$, by solving

$$\tilde{\alpha}_{ij}^{(t+1)} \in \underset{\tilde{\alpha}_{ij} \in \mathbb{R}}{\arg\min} \left\{ \nu(\tilde{\alpha}_{ij})^2 + \frac{1}{\tau}(\tilde{\alpha}_{ij} - \tilde{\alpha}_{ij}^{(t)})^2 + \frac{1}{n_i}\phi_{ij}^*\big(\tilde{\alpha}_{ij}\big) - \tilde{\alpha}_{ij} v_{ij}^{(t)} \right\}$$

whose optimality condition implies

$$0 \in 2\nu\tilde{\alpha}_{ij}^{(t+1)} + \frac{2}{\tau}(\tilde{\alpha}_{ij}^{(t+1)} - \tilde{\alpha}_{ij}^{(t)}) + \frac{1}{n_i}\partial\phi_{ij}^*\big(\tilde{\alpha}_{ij}^{(t+1)}\big) - v_{ij}^{(t)}. \tag{18}$$

By the definition of the saddle point $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\alpha}^*)$ of (9), the value $\tilde{\alpha}_{ij}^* := \frac{\alpha_{ij}^*}{y_i^*}$ satisfies

$$\tilde{\alpha}_{ij}^* \in \underset{\tilde{\alpha}_{ij} \in \mathbb{R}}{\arg\min} \left\{ -\frac{1}{n_i}\tilde{\alpha}_{ij}\xi_{ij}^\top \mathbf{x}^* + \frac{1}{n_i}\phi_{ij}^*\big(\tilde{\alpha}_{ij}\big) \right\}$$

whose optimality condition implies

$$0 \in -\frac{1}{n_i}\xi_{ij}^\top \mathbf{x}^* + \frac{1}{n_i}\partial\phi_{ij}^*\big(\tilde{\alpha}_{ij}^*\big). \tag{19}$$

Since ϕ_{ij} is smooth with its gradient being $\frac{1}{\gamma}$-Lipschitz continuous with respect to ℓ_2-norm, ϕ_{ij}^* is γ strongly convex with respect to ℓ_2-norm. Hence, the function $\nu(\alpha)^2 + \frac{1}{\tau}(\alpha - \tilde{\alpha}_{ij}^t)^2 + \frac{1}{n_i}\phi_{ij}^*(\alpha)$ is $(2\nu + \frac{2}{\tau} + \frac{\gamma}{n_i})$-strongly convex. Therefore, the strong monotonicity property of the subdifferential of this function implies

$$\begin{aligned} & \left[2\nu\tilde{\alpha}_{ij}^* + \frac{2}{\tau}(\tilde{\alpha}_{ij}^* - \tilde{\alpha}_{ij}^{(t)}) + \frac{1}{n_i}\partial\phi_{ij}^*\big(\tilde{\alpha}_{ij}^*\big) - 2\nu\tilde{\alpha}_{ij}^{(t+1)} - \frac{2}{\tau}(\tilde{\alpha}_{ij}^{(t+1)} - \tilde{\alpha}_{ij}^{(t)}) - \frac{1}{n_i}\partial\phi_{ij}^*\big(\tilde{\alpha}_{ij}^{(t+1)}\big)\right][\tilde{\alpha}_{ij}^* - \tilde{\alpha}_{ij}^{(t+1)}] \\ \ge & \left(2\nu + \frac{2}{\tau} + \frac{\gamma}{n_i}\right)(\tilde{\alpha}_{ij}^* - \tilde{\alpha}_{ij}^{(t+1)})^2, \end{aligned}$$

which implies

$$\begin{aligned}&\left|2\nu\tilde{\alpha}_{ij}^{*}+\frac{2}{\tau}(\tilde{\alpha}_{ij}^{*}-\tilde{\alpha}_{ij}^{(t)})+\frac{1}{n_i}\partial\phi_{ij}^{*}(\tilde{\alpha}_{ij}^{*})-2\nu\tilde{\alpha}_{ij}^{(t+1)}-\frac{2}{\tau}(\tilde{\alpha}_{ij}^{(t+1)}-\tilde{\alpha}_{ij}^{(t)})-\frac{1}{n_i}\partial\phi_{ij}^{*}(\tilde{\alpha}_{ij}^{(t+1)})\right|\\ \geq\ &\left(2\nu+\frac{2}{\tau}+\frac{\gamma}{n_i}\right)|\tilde{\alpha}_{ij}^{*}-\tilde{\alpha}_{ij}^{(t+1)}|.\end{aligned}$$

Applying the relationship (18) and (19) to the inequality above gives

$$\left|2\nu\tilde{\alpha}_{ij}^{*}+\frac{2}{\tau}(\tilde{\alpha}_{ij}^{*}-\tilde{\alpha}_{ij}^{(t)})+\frac{1}{n_i}\xi_{ij}^{\top}\mathbf{x}^{*}-v_{ij}^{(t)}\right|\geq\left(2\nu+\frac{2}{\tau}+\frac{\gamma}{n_i}\right)|\tilde{\alpha}_{ij}^{*}-\tilde{\alpha}_{ij}^{(t+1)}|,$$

which, by the triangle's inequality, further implies

$$\frac{2\nu\left\|\tilde{\boldsymbol{\alpha}}_i^{*}\right\|_2+\frac{2}{\tau}\left\|\tilde{\boldsymbol{\alpha}}_i^{*}-\tilde{\boldsymbol{\alpha}}_i^{(t)}\right\|_2+\frac{\gamma}{n_i}\left\|\frac{\Theta_i\mathbf{x}^{*}}{\gamma}-\frac{n_i\mathbf{v}_i^{(t)}}{\gamma}\right\|_2}{2\nu+\frac{2}{\tau}+\frac{\gamma}{n_i}}\geq\|\tilde{\boldsymbol{\alpha}}_i^{*}-\tilde{\boldsymbol{\alpha}}_i^{(t+1)}\|_2.\tag{20}$$

Note that the relationship $\frac{1}{n_i}\xi_{ij}^{\top}\mathbf{x}^{*}=\frac{1}{n_i}\partial\phi_{ij}^{*}(\tilde{\alpha}_{ij}^{*})$ implies $\nabla\phi_{ij}(\xi_{ij}^{\top}\mathbf{x}^{*})=\tilde{\alpha}_{ij}^{*}$. Moreover, the definition of $\mathbf{v}^{(t)}$ in Algorithm 4 indicates that

$$\|\Theta_i\mathbf{x}^{*}-n_i\mathbf{v}_i^{(t)}\|_2\leq 2\|\Theta_i\|_2B_{\mathbf{x}}+d\|\Theta_{ik}\|_2\|\bar{\mathbf{x}}_k^{(s)}-\mathbf{x}_k^{(t)}\|\leq 4d\max_k\|\Theta_{ik}\|_2B_{\mathbf{x}}$$

where Θ_{ik} is the kth column of Θ_i. By the induction hypothesis (17) and (20), we conclude that

$$\|\tilde{\boldsymbol{\alpha}}_i^{*}-\tilde{\boldsymbol{\alpha}}_i^{(t+1)}\|_2\leq\max\left\{\|\tilde{\boldsymbol{\alpha}}_i^{*}\|_2,\frac{4d\max_k\|\Theta_{ik}\|_2B_{\mathbf{x}}}{\gamma},\|\bar{\boldsymbol{\alpha}}_i^{(0)}/\bar{y}_i^{(0)}-\tilde{\boldsymbol{\alpha}}_i^{*}\|_2\right\}$$

so that the result (17) holds for $t+1$.

Finally, using (17) and the fact that $\|\tilde{\boldsymbol{\alpha}}_i^{(t)}\|_2\leq\|\tilde{\boldsymbol{\alpha}}_i^{*}\|_2+\|\tilde{\boldsymbol{\alpha}}_i^{*}-\tilde{\boldsymbol{\alpha}}_i^{(t)}\|_2$, we can show

$$\|\tilde{\boldsymbol{\alpha}}_i^{(t)}\|_2\leq\max\left\{2\|\tilde{\boldsymbol{\alpha}}_i^{*}\|_2,\frac{8d\max_k\|\Theta_{ik}\|_2B_{\mathbf{x}}}{\gamma},2\|\bar{\boldsymbol{\alpha}}_i^{(0)}/\bar{y}_i^{(0)}-\tilde{\boldsymbol{\alpha}}_i^{*}\|_2\right\}\leq B$$

which completes the proof. □

C. Proof of Theorem 1

Proof. The complexity of Algorithm 1 can be analyzed with a similar argument as in Section 2.1 in Aravkin et al. (2016) by incorporating the complexity of oracle $\mathcal{A}$. Consider an iteration k that is not the last iteration of Algorithm 1, i.e., $U(r^{(k)})>\varepsilon$. The property of $\mathcal{A}$ guarantees that $\theta H(r^{(k)})\geq\theta L(r^{(k)})\geq U(r^{(k)})>\varepsilon$ so that the complexity of $\mathcal{A}$ in iteration k is at most

$$\mathcal{C}(\max\{H(r^{(k)}),\varepsilon\})\leq\mathcal{C}(\max\{\theta^{-1}\varepsilon,\varepsilon\})=\mathcal{C}(\varepsilon).$$

Here, we use the facts that $\theta>1$ and that $\mathcal{C}(\cdot)$ is non-increasing by Definition 1. On the other hand, in the last iteration of Algorithm 1 where $U(r^{(k)})\leq\varepsilon$, we have $H(r^{(k)})\leq U(r^{(k)})\leq\varepsilon$ so that the complexity of $\mathcal{A}$ here is still at most $\mathcal{C}(\varepsilon)$. According to Theorem 2.4 in Aravkin et al. (2016), Algorithm 1 terminates after at most $\max\{1+\log_{2/\theta}(\frac{2\max\{|S(r^{(0)})||f^{*}-r^{(0)}|,L(r^{(0)})\}}{\varepsilon}),2\}$ iterations so that the total expected complexity of Algorithm 1 is $\mathcal{C}(\varepsilon)\max\{1+\log_{2/\theta}(\frac{2\max\{|S(r^{(0)})||f^{*}-r^{(0)}|,L(r^{(0)})\}}{\varepsilon}),2\}$. At the last iteration, we have $\mathcal{P}(r^{(k)};\mathbf{x}^{(k)})\leq U(r^{(k)})\leq\varepsilon$, which means the output solution $\mathbf{x}^{(k)}$ is ε-optimal and ε-feasible by the definition of $\mathcal{P}$ and the fact that $r^{(k)}\leq f^{*}$ during Algorithm 1. Then, we have verified the conclusion (1) and the first part of conclusion (c).

In the next, we analyze the complexity of Algorithm 2. The most part of the proof is from the proof of Theorem 2 in Lin et al. (2017). However, one major difference in our proof from Lin et al. (2017) is that we analyze the complexity for Algorithm 2 under a termination condition different from the one used in Lin et al. (2017). This difference is essential because it is the

main reason for Algorithm 2 to ensure an absolute ϵ-optimal solution while Lin et al. (2017) ensures a relative ϵ-optimal solution.

First of all, we claim that $S(r) \leq 0$ for any r. In fact, for any $r' > r$, the property of $S(r)$ promised by oracle $\mathcal{A}$ and the non-increasing property of $H(\cdot)$ by Lemma 1 guarantees $H(r) \geq H(r') \geq L(r) + S(r)(r' - r)$, which implies $S(r) \leq \frac{H(r)-L(r)}{r'-r}$. Letting r' goes to infinity leads to this conclusion.

According to the definition of β, Lemma 1(d) and convexity of $H(r)$, we can show that

$$\beta(r - f^*) \leq -H(r) \leq r - f^*, \quad \forall r \in (f^*, r^{(0)}]. \tag{21}$$

Suppose Algorithm 2 terminates at iteration $k = K$. From (21), the updating equation for $r^{(k+1)}$ and the fact that $H(r^{(k)}) \leq U(r^{(k)}) \leq L(r^{(k)})/\theta \leq H(r^{(k)})/\theta < 0$ (according to the property of $\mathcal{A}$), we have

$$r^{(k+1)} - f^* = r^{(k)} - f^* + U(r^{(k)})/2 \geq r^{(k)} - f^* + \frac{H(r^{(k)})}{2} \geq \frac{1}{2}(r^{(k)} - f^*) \tag{22}$$

$$r^{(k+1)} - f^* = r^{(k)} - f^* + U(r^{(k)})/2 \leq r^{(k)} - f^* + \frac{H(r^{(k)})}{2\theta} \leq \left(1 - \frac{\beta}{2\theta}\right)(r^{(k)} - f^*), \tag{23}$$

for $k = 0, 1, \ldots, K$. Recursively applying both inequalities gives

$$0 < \frac{1}{2^k}(r^{(0)} - f^*) \leq r^{(k)} - f^* \leq \left(1 - \frac{\beta}{2\theta}\right)^k (r^{(0)} - f^*), \quad \text{for } k = 0, 1, 2, \ldots, K. \tag{24}$$

Choosing $r = r^{(k)}$ in the inequality (21), and applying (24) and the properties of $L(r^{(k)})$ and $U(r^{(k)})$, we can show that

$$-L(r^{(k)}) \leq -\theta U(r^{(k)}) \leq -\theta H(r^{(k)}) \leq \theta(r^{(k)} - f^*) \leq \theta\left(1 - \frac{\beta}{2\theta}\right)^k (r^{(0)} - f^*) \leq -\frac{H(r^{(0)})}{2}$$

for $k \geq \frac{2\theta}{\beta} \log\left(\frac{2\theta(r^{(0)}-f^*)}{|H(r^{(0)})|}\right)$. With k satisfying this inequality, using the fact that $f^* < r^{(k)}$ by (24), the definition of $S(r^{(k)})$, and the fact that $S(r^{(k)}) \leq 0$, we can prove that

$$H(r^{(0)}) \geq L(r^{(k)}) + S(r^{(k)})(r^{(0)} - r^{(k)}) \geq \frac{H(r^{(0)})}{2} + S(r^{(k)})(r^{(0)} - f^*),$$

or equivalently, $S(r^{(k)}) \leq \frac{H(r^{(0)})}{2(r^{(0)}-f^*)} = -\frac{\beta}{2} < 0$. Therefore, if we simultaneously require $k \geq \frac{2\theta}{\beta} \log\left(\frac{2\theta(r^{(0)}-f^*)^2}{|H(r^{(0)})|\varepsilon}\right)$, we will further ensure

$$-L(r^{(k)}) \leq \frac{-H(r^{(0)})\varepsilon}{2(r^{(0)} - f^*)} = \frac{\beta\varepsilon}{2} \leq -\varepsilon S(r^{(k)}).$$

Therefore, the total number of main iterations K of Algorithm 2 satisfies

$$K \leq \frac{2\theta}{\beta} \log\left(\frac{2\theta(r^{(0)} - f^*)}{|H(r^{(0)})|} \max\left\{\frac{r^{(0)} - f^*}{\varepsilon}, 1\right\}\right) = \frac{2\theta}{\beta} \log\left(\frac{2\theta}{\beta} \max\left\{\frac{r^{(0)} - f^*}{\varepsilon}, 1\right\}\right)$$

and, by the assumption on the complexity of $\mathcal{A}$ in Definition 1, the the overall expected complexity of Algorithm 2 is $\sum_{k=0}^{K} \mathcal{C}(|H(r^{(k)})|)$.

To further analyze the overall complexity of Algorithm 2, consider an iteration $k < K$, i.e., $L(r^{(k)}) < \varepsilon S(r^{(k)})$. The property of $\mathcal{A}$ guarantees that $\theta H(r^{(k)}) \leq L(r^{(k)}) < \varepsilon S(r^{(k)})$ which, together with the definition of $S(r^{(k)})$ and the fact $f^* < r^{(k)} < r^{(0)}$, implies that

$$H(r^{(0)}) \geq L(r^{(k)}) + S(r^{(k)})(r^{(0)} - r^{(k)}) \geq \theta H(r^{(k)}) + \frac{\theta H(r^{(k)})}{\varepsilon}(r^{(0)} - f^*).$$

This inequality further implies, for $0 \leq k < K$,

$$|H(r^{(k)})| \geq \frac{|H(r^{(0)})|}{\theta(1 + (r^{(0)} - f^*)/\varepsilon)}. \tag{25}$$

Hence, the expected complexity of $\mathcal{A}$ in iteration $k < K$ (non-terminating iteration) of Algorithm 2 is at most

$$\mathcal{C}(|H(r^{(k)})|) \leq \mathcal{C}\left(\frac{|H(r^{(0)})|}{\theta(1+(r^{(0)}-f^*)/\varepsilon)}\right) \leq \mathcal{C}\left(\frac{\beta|H(r^{(0)})|}{2\theta(1+(r^{(0)}-f^*)/\varepsilon)}\right).$$

On the other hand, according to (21), (22) and (25), we have

$$-H(r^{(K)}) \geq \beta(r^{(K)}-f^*) \geq \frac{\beta}{2}(r^{(K-1)}-f^*) \geq \frac{\beta|H(r^{(K-1)})|}{2} \geq \frac{\beta|H(r^{(0)})|}{2\theta(1+(r^{(0)}-f^*)/\varepsilon)},$$

so the complexity of $\mathcal{A}$ at the last iteration (i.e., $k = K$) is at most

$$\mathcal{C}(|H(r^{(k)})|) \leq \mathcal{C}\left(\frac{\beta|H(r^{(0)})|}{2\theta(1+(r^{(0)}-f^*)/\varepsilon)}\right).$$

Hence, the total expected complexity of Algorithm 2 is $\mathcal{C}\left(\frac{\beta|H(r^{(0)})|}{2\theta(1+(r^{(0)}-f^*)/\varepsilon)}\right)\frac{2\theta}{\beta}\log\left(\frac{2\theta}{\beta}\max\{\frac{r^{(0)}-f^*}{\varepsilon},1\}\right)$.

Lastly, we analyze the quality of the solutions from Algorithm 2. Recall that $f^* < r^{(k)}$ by (24), so that, at any iteration, $\mathcal{P}(r^{(k)};\mathbf{x}_k) \leq U(r^{(k)}) \leq L(r^{(k)})/\theta \leq H(r^{(k)})/\theta < 0$, which implies that $\max_{i=1,\ldots,m}[f_i(\mathbf{x}^k) - r_i] \leq 0$ according to the definition of $\mathcal{P}$. Hence, we have proved that $\mathbf{x}^{(k)}$ is a strictly feasible solution for any k.

Furthermore, we note that the affine-minorant property of $S(r^{(K)})$ implies

$$H\left(r^{(K)} - \frac{L(r^{(K)})}{S(r^{(K)})}\right) \geq L(r^{(K)}) + S(r^{(K)})\left(r^{(K)} - \frac{L(r^{(K)})}{S(r^{(K)})} - r^{(K)}\right) = 0,$$

so we must have $r^{(K)} - L(r^{(K)})/S(r^{(K)}) \leq f^*$, which further ensures $r^{(K)} - f^* \leq L(r^{(K)})/S(r^{(K)}) \leq \varepsilon$ according to the terminating condition of Algorithm 2. At the last iteration, again, we have $\mathcal{P}(r^{(K)};\mathbf{x}^{(K)}) \leq U(r^{(K)}) \leq L(r^{(K)})/\theta \leq H(r^{(K)})/\theta < 0$. Because $0 \leq r^{(K)} - f^* \leq \varepsilon$ and $\mathcal{P}(r^{(K)};\mathbf{x}^{(K)}) < 0$, we have $f_0(\mathbf{x}^{(K)}) - f^* \leq r^{(K)} - f^* \leq \varepsilon$ and $\max_{i=1,\ldots,m}[f_i(\mathbf{x}^K) - r_i] \leq 0$ according to the definition of $\mathcal{P}$. Hence, Algorithm 2 returns an ε-optimal and feasible solution at termination. Then, we have verified the conclusion (2) and the second part of conclusion (c).

□

D. Proof of Proposition 1

Proof of Proposition 1. By the definition of G_ν, D and h_B, after organizing terms, (12) can be formulated as

$$\min_{\mathbf{w}\in\mathcal{W}}\left\{\begin{array}{l} 2(1+B)^2\nu\sum_{i=0}^m y_i\ln y_i + \frac{2(1+B)^2}{\tau}\sum_{i=0}^m y_i\ln\left(\frac{y_i}{y_i'}\right) + \mathbf{y}^\top\mathbf{r} \\ +\sum_{i=0}^m \nu y_i\left\|\frac{\boldsymbol{\alpha}_i}{y_i}\right\|_2^2 + \sum_{i=0}^m\frac{y_i}{\tau}\left\|\frac{\boldsymbol{\alpha}_i}{y_i}-\frac{\boldsymbol{\alpha}_i'}{y_i'}\right\|_2^2 + \sum_{i=0}^m\sum_{j=1}^{n_i}\frac{y_i}{n_i}\phi_{ij}^*\left(\frac{\alpha_{ij}}{y_i}\right) - \sum_{i=0}^m y_i\left(\frac{\boldsymbol{\alpha}_i}{y_i}\right)^\top\mathbf{v}_i \end{array}\right\}. \tag{26}$$

We first fix $\mathbf{y} \in \text{int}\Delta$ and only optimize $\boldsymbol{\alpha} \in \mathbb{R}^n$ in (26). By changing variables with $\tilde{\boldsymbol{\alpha}}_i = \frac{\boldsymbol{\alpha}_i}{y_i}$ and $\tilde{\boldsymbol{\alpha}}_i' = \frac{\boldsymbol{\alpha}_i'}{y_i'}$, (26) becomes

$$\begin{aligned} &\min_{\mathbf{y}\in\Delta,\tilde{\boldsymbol{\alpha}}\in\mathbb{R}^n}\left\{\begin{array}{l} 2(1+B)^2\nu\sum_{i=0}^m y_i\ln y_i + \frac{2(1+B)^2}{\tau}\sum_{i=0}^m y_i\ln\left(\frac{y_i}{y_i'}\right) + \mathbf{y}^\top\mathbf{r} \\ +\sum_{i=0}^m \nu y_i\|\tilde{\boldsymbol{\alpha}}_i\|_2^2 + \sum_{i=0}^m\frac{y_i}{\tau}\|\tilde{\boldsymbol{\alpha}}_i-\tilde{\boldsymbol{\alpha}}_i'\|_2^2 + \sum_{i=0}^m\sum_{j=1}^{n_i}\frac{y_i}{n_i}\phi_{ij}^*(\tilde{\alpha}_{ij}) - \sum_{i=0}^m y_i(\tilde{\boldsymbol{\alpha}}_i)^\top\mathbf{v}_i \end{array}\right\} \\ = &\min_{\mathbf{y}\in\Delta}\left\{\begin{array}{l} 2(1+B)^2\nu\sum_{i=0}^m y_i\ln y_i + \frac{2(1+B)^2}{\tau}\sum_{i=0}^m y_i\ln\left(\frac{y_i}{y_i'}\right) + \mathbf{y}^\top\mathbf{r} \\ +\sum_{i=0}^m y_i\min_{\tilde{\boldsymbol{\alpha}}_i\in\mathbb{R}^{n_i}}\left[\nu\|\tilde{\boldsymbol{\alpha}}_i\|_2^2 + \frac{1}{\tau}\|\tilde{\boldsymbol{\alpha}}_i-\tilde{\boldsymbol{\alpha}}_i'\|_2^2 + \sum_{j=1}^{n_i}\frac{1}{n_i}\phi_{ij}^*(\tilde{\alpha}_{ij}) - \tilde{\boldsymbol{\alpha}}_i^\top\mathbf{v}_i\right] \end{array}\right\} \qquad (27) \\ = &\min_{\mathbf{y}\in\Delta}\left\{2(1+B)^2\nu\sum_{i=0}^m y_i\ln y_i + \frac{2(1+B)^2}{\tau}\sum_{i=0}^m y_i\ln\left(\frac{y_i}{y_i'}\right) + \mathbf{y}^\top(\mathbf{r}+\boldsymbol{\rho})\right\}. \qquad (28) \end{aligned}$$

The equality (27) above indicates that the minimization over $\boldsymbol{\alpha}_i$ in (26) for a given $\mathbf{y}$ is equivalent to the inner minimization over $\tilde{\boldsymbol{\alpha}}_i$ in (27), which is independent of $\mathbf{y}$ and can be solved for each i separately. Note that the ith inner minimization is exactly (13), whose optimal solution, i.e., $\tilde{\boldsymbol{\alpha}}_i^\#$, has a closed form for many commonly used loss function ϕ_{ij}. The equality (28) indicates that, after obtaining the optimal $\tilde{\boldsymbol{\alpha}}_i$, we can solve the optimal $\mathbf{y}$ by solving the outer minimization problem (28) whose solution is exactly $\mathbf{y}^\#$ defined in Proposition 1 which can be verified from the optimality condition. According to the relationship that $\tilde{\boldsymbol{\alpha}}_i = \frac{\boldsymbol{\alpha}_i}{y_i}$, the optimal value of the original variable $\boldsymbol{\alpha}_i$ should be $\boldsymbol{\alpha}_i^\# = \tilde{\boldsymbol{\alpha}}_i^\# y_i^\#$. □

E. Proof of Theorem 2 and Theorem 3

In this section, we provide the proofs for Theorem 2 and Theorem 3.

Proof of Theorem 2. With a little abuse of notation, in this proof, we denote by $(\mathbf{x}^*, \mathbf{w}^*)$ the saddle point of (9) but hide their dependency on μ and ν. For simplicity of notation, we define $F_\mu(\mathbf{x}) := \frac{\mu\|\mathbf{x}\|_2^2}{2}$.

We first analyze the convergence property of the sth outer iteration of Algorithm 4. Let $\mathbb{E}_t$ represent the conditional expectation conditioning on $(\mathbf{x}^{(0)}, \mathbf{w}^{(0)}) = (\bar{\mathbf{x}}^{(s)}, \bar{\mathbf{w}}^{(s)})$ as well as all the stochastic outcomes up to the end of inner iteration t of the outer iteration s of Algorithm 4.

The definition of $(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)})$ and the optimality conditions of $(\mathbf{x}^*, \mathbf{w}^*)$ imply that, for any $\mathbf{x} \in \mathcal{X}$ and $\mathbf{w} = (\mathbf{y}, \boldsymbol{\alpha}) \in \mathcal{W}$,

$$\left(\mu + \frac{1}{\sigma}\right)\frac{\|\mathbf{x} - \mathbf{x}^{(t+1)}\|_2^2}{2} + (\mathbf{x}^{(t+1)})^\top \mathbf{u}^{(t)} + F_\mu(\mathbf{x}^{(t+1)}) + \frac{\|\mathbf{x}^{(t)} - \mathbf{x}^{(t+1)}\|_2^2}{2\sigma} \le \mathbf{x}^\top \mathbf{u}^{(t)} + F_\mu(\mathbf{x}) + \frac{\|\mathbf{x} - \mathbf{x}^{(t)}\|_2^2}{2\sigma} \quad (29)$$

$$\left(\nu + \frac{1}{\tau}\right) D(\mathbf{w}, \mathbf{w}^{(t+1)}) - (\boldsymbol{\alpha}^{(t+1)})^\top \mathbf{v}^{(t)} + G_\nu(\mathbf{w}^{(t+1)}) + \frac{D(\mathbf{w}^{(t+1)}, \mathbf{w}^{(t)})}{\tau} \le -\boldsymbol{\alpha}^\top \mathbf{v}^{(t)} + G_\nu(\mathbf{w}) + \frac{D(\mathbf{w}, \mathbf{w}^{(t)})}{\tau} \quad (30)$$

Motived by (88) and (89) in Xiao et al. (2017), we define

$$\tilde{\mathcal{P}}(\mathbf{x}) := \boldsymbol{\alpha}^* A\mathbf{x} + F_\mu(\mathbf{x}) - \boldsymbol{\alpha}^* A\mathbf{x}^* - F_\mu(\mathbf{x}^*) \quad \text{and} \quad \tilde{\mathcal{D}}(\mathbf{w}) := \boldsymbol{\alpha} A\mathbf{x}^* - G_\nu(\mathbf{w}) - \boldsymbol{\alpha}^* A\mathbf{x}^* + G_\nu(\mathbf{w}^*).$$

Note that $\min_{\mathbf{x}\in\mathcal{X}} \tilde{\mathcal{P}}(\mathbf{x}) = \tilde{\mathcal{P}}(\mathbf{x}^*) = 0$ and $\max_{\mathbf{w}\in\mathcal{W}} \tilde{\mathcal{D}}(\mathbf{w}) = \tilde{\mathcal{D}}(\mathbf{w}^*) = 0$. By the μ-strong convexity of F_μ with respect to Euclidean distance and the ν-strong convexity of G_ν with respect to Bregman divergence D, we can show that

$$\tilde{\mathcal{P}}(\mathbf{x}) \ge \frac{\mu\|\mathbf{x} - \mathbf{x}^*\|_2^2}{2} \quad \text{and} \quad -\tilde{\mathcal{D}}(\mathbf{w}) \ge \nu D(\mathbf{w}, \mathbf{w}^*) \quad (31)$$

We choose $\mathbf{x} = \mathbf{x}^*$ in (29) and $\mathbf{w} = \mathbf{w}^*$ in (30), and the add (29) and (30) together. After organizing terms, we obtain

$$\begin{aligned}
&\left(\mu + \frac{1}{\sigma}\right)\frac{\|\mathbf{x}^* - \mathbf{x}^{(t+1)}\|_2^2}{2} + \frac{\|\mathbf{x}^{(t)} - \mathbf{x}^{(t+1)}\|_2^2}{2\sigma} + \left(\nu + \frac{1}{\tau}\right) D(\mathbf{w}^*, \mathbf{w}^{(t+1)}) + \frac{D(\mathbf{w}^{(t+1)}, \mathbf{w}^{(t)})}{\tau} \\
&+\tilde{\mathcal{P}}(\mathbf{x}^{(t+1)}) - \tilde{\mathcal{D}}(\mathbf{w}^{(t+1)}) \\
\le\ & (\mathbf{x}^* - \mathbf{x}^{(t+1)})^\top \mathbf{u}^{(t)} + \frac{\|\mathbf{x}^* - \mathbf{x}^{(t)}\|_2^2}{2\sigma} - (\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^{(t+1)})^\top \mathbf{v}^{(t)} + \frac{D(\mathbf{w}^*, \mathbf{w}^{(t)})}{\tau} + \boldsymbol{\alpha}^* A\mathbf{x}^{(t+1)} - \boldsymbol{\alpha}^{(t+1)} A\mathbf{x}^* \\
=\ & (\mathbf{x}^* - \mathbf{x}^{(t)})^\top [\mathbf{u}^{(t)} - A^\top \boldsymbol{\alpha}^{(t)}] + (\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^{(t)})^\top [A\mathbf{x}^{(t)} - \mathbf{v}^{(t)}] + \frac{\|\mathbf{x}^* - \mathbf{x}^{(t)}\|_2^2}{2\sigma} + \frac{D(\mathbf{w}^*, \mathbf{w}^{(t)})}{\tau} \\
&+(\mathbf{x}^* - \mathbf{x}^{(t)})^\top A^\top \boldsymbol{\alpha}^{(t)} - (\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^{(t)})^\top A\mathbf{x}^{(t)} - (\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)})^\top A^\top \boldsymbol{\alpha}^{(t)} + (\boldsymbol{\alpha}^{(t+1)} - \boldsymbol{\alpha}^{(t)})^\top A\mathbf{x}^{(t)} \\
&+(\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)})^\top [A^\top \boldsymbol{\alpha}^{(t)} - \mathbf{u}^{(t)}] - (\boldsymbol{\alpha}^{(t+1)} - \boldsymbol{\alpha}^{(t)})^\top [A\mathbf{x}^{(t)} - \mathbf{v}^{(t)}] + \boldsymbol{\alpha}^* A\mathbf{x}^{(t+1)} - \boldsymbol{\alpha}^{(t+1)} A\mathbf{x}^* \\
=\ & (\mathbf{x}^* - \mathbf{x}^{(t)})^\top [\mathbf{u}^{(t)} - A^\top \boldsymbol{\alpha}^{(t)}] + (\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^{(t)})^\top [A\mathbf{x}^{(t)} - \mathbf{v}^{(t)}] + \frac{\|\mathbf{x}^* - \mathbf{x}^{(t)}\|_2^2}{2\sigma} + \frac{D(\mathbf{w}^*, \mathbf{w}^{(t)})}{\tau} \\
&-(\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)})^\top A^\top (\boldsymbol{\alpha}^{(t)} - \boldsymbol{\alpha}^*) + (\boldsymbol{\alpha}^{(t+1)} - \boldsymbol{\alpha}^{(t)})^\top A(\mathbf{x}^{(t)} - \mathbf{x}^*) \\
&+(\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)})^\top [A^\top \boldsymbol{\alpha}^{(t)} - \mathbf{u}^{(t)}] - (\boldsymbol{\alpha}^{(t+1)} - \boldsymbol{\alpha}^{(t)})^\top [A\mathbf{x}^{(t)} - \mathbf{v}^{(t)}]
\end{aligned} \quad (32)$$

Since the random indexes k and l are independent of $\mathbf{x}^{(t)}$ and $\mathbf{w}^{(t)}$, we have

$$\mathbb{E}_t[(\mathbf{x}^* - \mathbf{x}^{(t)})^\top (\mathbf{u}^{(t)} - A^\top \boldsymbol{\alpha}^{(t)})] = 0 \quad \text{and} \quad \mathbb{E}_t[(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^{(t)})^\top (A\mathbf{x}^{(t)} - \mathbf{v}^{(t)})] = 0 \quad (33)$$

by the definition of $\mathbf{u}^{(t)}$ and $\mathbf{v}^{(t)}$.

Next, we study the three lines on the right hand side of (32), respectively. By the definition of $\mathbf{u}^{(t)}$, Cauchy-Schwarz

inequality and Young's inequality, we have

$$\begin{aligned}
&\mathbb{E}_t\left[(\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)})^\top(\mathbf{u}^{(t)}-A^\top\boldsymbol{\alpha}^{(t)})\right]\\
\leq\ & \frac{1}{2a_t}\mathbb{E}_t\|\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}\|_2^2+\frac{a_t}{2}\mathbb{E}_t\|A^\top\bar{\boldsymbol{\alpha}}^{(s)}+nA_{l:}^\top\boldsymbol{\alpha}_l^{(t)}-nA_{l:}^\top\bar{\boldsymbol{\alpha}}_l^{(s)}-A^\top\boldsymbol{\alpha}^{(t)}\|_2^2\\
\leq\ & \frac{1}{2a_t}\mathbb{E}_t\|\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}\|_2^2+a_tn\max_l\|A_{l:}\|_2^2\|\boldsymbol{\alpha}^{(t)}-\boldsymbol{\alpha}^*\|_2^2+a_tn\max_l\|A_{l:}\|_2^2\|\bar{\boldsymbol{\alpha}}^{(s)}-\boldsymbol{\alpha}^*\|_2^2\\
\leq\ & \frac{1}{2a_t}\mathbb{E}_t\|\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}\|_2^2+2a_tn\max_l\|A_{l:}\|_2^2D(\mathbf{w}^*,\mathbf{w}^{(t)})+2a_tn\max_l\|A_{l:}\|_2^2D(\mathbf{w}^*,\bar{\mathbf{w}}^{(s)})
\end{aligned}\tag{34}$$

Similarly, we can prove that

$$\begin{aligned}
&\mathbb{E}_t\left[(\boldsymbol{\alpha}^{(t)}-\boldsymbol{\alpha}^{(t+1)})^\top(A\mathbf{x}^{(t)}-\mathbf{v}^{(t)})\right]\\
\leq\ & \frac{1}{2b_t}\mathbb{E}_t\|\boldsymbol{\alpha}^{(t)}-\boldsymbol{\alpha}^{(t+1)}\|_2^2+b_td\max_k\|A_{:k}\|_2^2\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2+b_td\max_k\|A_{:k}\|_2^2\|\bar{\mathbf{x}}^{(s)}-\mathbf{x}^*\|_2^2\\
\leq\ & \frac{1}{b_t}\mathbb{E}_tD(\mathbf{w}^{(t+1)},\mathbf{w}^{(t)})+b_td\max_k\|A_{:k}\|_2^2\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2+b_td\max_k\|A_{:k}\|_2^2\|\bar{\mathbf{x}}^{(s)}-\mathbf{x}^*\|_2^2
\end{aligned}\tag{35}$$

Applying Cauchy-Schwarz inequality and Young's inequality in a similar way gives

$$\mathbb{E}_t\left[(\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)})^\top A^\top(\boldsymbol{\alpha}^{(t)}-\boldsymbol{\alpha}^*)\right]\ \leq\ \frac{1}{2a_t}\mathbb{E}_t\|\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}\|_2^2+a_t\|A\|_2^2D(\mathbf{w}^*,\mathbf{w}^{(t)})\tag{36}$$

$$\mathbb{E}_t\left[(\boldsymbol{\alpha}^{(t+1)}-\boldsymbol{\alpha}^{(t)})^\top A(\mathbf{x}^{(t)}-\mathbf{x}^*)\right]\ \leq\ \frac{1}{b_t}\mathbb{E}_tD(\mathbf{w}^{(t+1)},\mathbf{w}^{(t)})+\frac{b_t\|A\|_2^2}{2}\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2\tag{37}$$

Choosing $a_t=2\sigma$ and $b_t=2\tau$ and applying (33), (34), (35), (36) and (37) to (32) lead to

$$\begin{aligned}
&\left(\mu+\frac{1}{\sigma}\right)\frac{\mathbb{E}_t\|\mathbf{x}^*-\mathbf{x}^{(t+1)}\|_2^2}{2}+\left(\nu+\frac{1}{\tau}\right)\mathbb{E}_tD(\mathbf{w}^*,\mathbf{w}^{(t+1)})+\tilde{\mathcal{P}}(\mathbf{x}^{(t+1)})-\tilde{\mathcal{D}}(\mathbf{w}^{(t+1)})\\
\leq\ & \left(2\tau\|A\|_2^2+4\tau d\max_k\|A_{:k}\|_2^2+\frac{1}{\sigma}\right)\frac{\|\mathbf{x}^*-\mathbf{x}^{(t)}\|_2^2}{2}+\left(2\sigma\|A\|_2^2+4\sigma n\max_l\|A_{l:}\|_2^2+\frac{1}{\tau}\right)D(\mathbf{w}^*,\mathbf{w}^{(t)})\\
&+2\tau d\max_k\|A_{:k}\|_2^2\|\mathbf{x}^*-\bar{\mathbf{x}}^{(s)}\|_2^2+4\sigma n\max_l\|A_{l:}\|_2^2D(\mathbf{w}^*,\bar{\mathbf{w}}^{(s)})\\
\leq\ & \left(\tau\kappa\mu\nu+2\tau\kappa\mu\nu+\frac{1}{\sigma}\right)\frac{\|\mathbf{x}^*-\mathbf{x}^{(t)}\|_2^2}{2}+\left(\sigma\kappa\mu\nu+2\sigma\kappa\mu\nu+\frac{1}{\tau}\right)D(\mathbf{w}^*,\mathbf{w}^{(t)})\\
&+\tau\kappa\mu\nu\|\mathbf{x}^*-\bar{\mathbf{x}}^{(s)}\|_2^2+2\sigma\kappa\mu\nu D(\mathbf{w}^*,\bar{\mathbf{w}}^{(s)}),
\end{aligned}\tag{38}$$

where in the last inequality we use the fact that the operator norm of A, i.e., $\|A\|_2$, satisfies $\|A\|_2^2\leq\|A\|_{\max}^2=\frac{\kappa\mu\nu}{2}$ and the fact that $\max\{d\max_k\|A_{:k}\|_2^2,n\max_l\|A_{l:}\|_2^2\}\leq\|A\|_{\max}^2=\frac{\kappa\mu\nu}{2}$.

Let η be a constant to be determined later. Choosing $\sigma=\frac{\eta}{\kappa\mu}$ and $\tau=\frac{\eta}{\kappa\nu}$ in (38), we obtain the following inequality

$$\begin{aligned}
&\left(1+\frac{\kappa}{\eta}\right)\mu\mathbb{E}_t\frac{\|\mathbf{x}^*-\mathbf{x}^{(t+1)}\|_2^2}{2}+\left(1+\frac{\kappa}{\eta}\right)\nu\mathbb{E}_tD(\mathbf{w}^*,\mathbf{w}^{(t+1)})+\mathbb{E}_t\tilde{\mathcal{P}}(\mathbf{x}^{(t+1)})-\mathbb{E}_t\tilde{\mathcal{D}}(\mathbf{w}^{(t+1)})\\
\leq\ & \left(3\eta+\frac{\kappa}{\eta}\right)\mu\frac{\|\mathbf{x}^*-\mathbf{x}^{(t)}\|_2^2}{2}+\left(3\eta+\frac{\kappa}{\eta}\right)\nu D(\mathbf{w}^*,\mathbf{w}^{(t)})+\eta\mu\|\mathbf{x}^*-\bar{\mathbf{x}}^{(s)}\|_2^2+2\eta\nu D(\mathbf{w}^*,\bar{\mathbf{w}}^{(s)}),
\end{aligned}$$

which, if divided by $\left(1+\frac{\kappa}{\eta}\right)$, further implies

$$\frac{1}{1+\frac{\kappa}{\eta}}[\tilde{\mathcal{P}}(\mathbf{x}^{(t+1)})-\tilde{\mathcal{D}}(\mathbf{w}^{(t+1)})]+\mathbb{E}_t\delta^{(t+1)}\ \leq\ \left(1-\frac{1-3\eta}{1+\frac{\kappa}{\eta}}\right)\mathbb{E}_t\delta^{(t)}+\frac{2\eta}{1+\frac{\kappa}{\eta}}\mathbb{E}_t\bar{\delta}^{(s)},\tag{39}$$

where

$$\delta^{(t)}:=\frac{\mu\|\mathbf{x}^*-\mathbf{x}^{(t)}\|_2^2}{2}+\nu D(\mathbf{w}^*,\mathbf{w}^{(t)})\quad\text{and}\quad\bar{\delta}^{(s)}:=\frac{\mu\|\mathbf{x}^*-\bar{\mathbf{x}}^{(s)}\|_2^2}{2}+\nu D(\mathbf{w}^*,\bar{\mathbf{w}}^{(s)}).$$

Since $\delta^{(0)} = \bar{\delta}^{(s)}$ and $\delta^{(T)} = \bar{\delta}^{(s+1)}$, applying (39) recursively for $t = 0, 1, \ldots, T-1$ yields

$$\frac{1}{1+\frac{\kappa}{\eta}}[\tilde{\mathcal{P}}(\bar{\mathbf{x}}^{(s+1)}) - \tilde{\mathcal{D}}(\bar{\mathbf{w}}^{(s+1)})] + \bar{\delta}^{(s+1)} \leq \left\{\left(1 - \frac{1-3\eta}{1+\frac{\kappa}{\eta}}\right)^T + \frac{2\eta}{1-3\eta}\right\}\bar{\delta}^{(s)}$$

Choosing $\eta = \frac{1}{12}$ in this inequality gives

$$\frac{1}{1+12\kappa}[\tilde{\mathcal{P}}(\bar{\mathbf{x}}^{(s+1)}) - \tilde{\mathcal{D}}(\bar{\mathbf{w}}^{(s+1)})] + \bar{\delta}^{(s+1)} \leq \left\{\left(1 - \frac{1}{4/3+16\kappa}\right)^T + \frac{2}{9}\right\}\bar{\delta}^{(s)}$$

Hence, by choosing $T = (4/3+16\kappa)\log(\frac{18}{5})$, we can ensure $\left(1 - \frac{1}{4/3+16\kappa}\right)^T \leq \frac{5}{18}$ which implies

$$\frac{1}{1+12\kappa}[\tilde{\mathcal{P}}(\bar{\mathbf{x}}^{(s+1)}) - \tilde{\mathcal{D}}(\bar{\mathbf{w}}^{(s+1)})] + \bar{\delta}^{(s+1)} \leq \frac{1}{2}\bar{\delta}^{(s)}. \tag{40}$$

Because $\tilde{\mathcal{P}}(\mathbf{x}) - \tilde{\mathcal{D}}(\mathbf{w}) \geq 0$ for any $\mathbf{x} \in \mathcal{X}$ and $\mathbf{w} \in \mathcal{W}$, the inequality above, if applied recursively for $s = 0, 1, \ldots, S-1$, implies

$$\bar{\delta}^{(s)} \leq \left(\frac{1}{2}\right)^s \bar{\delta}^{(0)}. \tag{41}$$

The inequality (41) only establishes the convergence of Algorithm 4 in terms of the solution's distance to the saddle point of (9). In the next, we will prove the convergence of $\mathcal{P}_{\mu,\nu}(r; \bar{\mathbf{x}}^{(s+1)}) - \mathcal{D}_{\mu,\nu}(r; \bar{\mathbf{w}}^{(s+1)})$ to zero.

By the μ-strong convexity of F_μ and the ν-strong convexity of G_ν with respect to Euclidean distance, we can show that $\mathcal{P}_{\mu,\nu}(r; \mathbf{x})$ and $\mathcal{D}_{\mu,\nu}(r; \mathbf{w})$ are smooth with Lipschitz continuous gradients. Therefore, according to Lemma 8 in Xiao et al. (2017), we have

$$\begin{aligned}
\mathcal{P}_{\mu,\nu}(r; \mathbf{x}) - \mathcal{D}_{\mu,\nu}(r; \mathbf{w}) &\leq \tilde{\mathcal{P}}(\mathbf{x}) - \tilde{\mathcal{D}}(\mathbf{w}) + \frac{\|A\|^2}{2\nu}\|\mathbf{x} - \mathbf{x}^*\|_2^2 + \frac{\|A\|^2}{2\mu}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_2^2 \\
&\leq \tilde{\mathcal{P}}(\mathbf{x}) - \tilde{\mathcal{D}}(\mathbf{w}) + \frac{\|A\|^2}{2\nu}\|\mathbf{x} - \mathbf{x}^*\|_2^2 + \frac{\|A\|^2}{\mu}D(\boldsymbol{\alpha}^*, \boldsymbol{\alpha})
\end{aligned}$$

for any $\mathbf{x} \in \mathcal{X}$ and $\mathbf{w} \in \mathcal{W}$, which implies

$$\mathcal{P}_{\mu,\nu}(r; \bar{\mathbf{x}}^{(s+1)}) - \mathcal{D}_{\mu,\nu}(r; \bar{\mathbf{w}}^{(s+1)}) \leq \tilde{\mathcal{P}}(\bar{\mathbf{x}}^{(s+1)}) - \tilde{\mathcal{D}}(\bar{\mathbf{w}}^{(s+1)}) + \kappa\bar{\delta}^{(s+1)} \leq (1+12\kappa)\left\{\frac{1}{1+12\kappa}[\tilde{\mathcal{P}}(\bar{\mathbf{x}}^{(s+1)}) - \tilde{\mathcal{D}}(\bar{\mathbf{w}}^{(s+1)})] + \bar{\delta}^{(s+1)}\right\}.$$

Applying this inequality to (40) and combining it with (41) yield

$$\mathcal{P}_{\mu,\nu}(r; \bar{\mathbf{x}}^{(s)}) - \mathcal{D}_{\mu,\nu}(r; \bar{\mathbf{w}}^{(s)}) \leq (1+12\kappa)\left\{\frac{1}{1+12\kappa}[\tilde{\mathcal{P}}(\bar{\mathbf{x}}^{(s)}) - \tilde{\mathcal{D}}(\bar{\mathbf{w}}^{(s)})] + \bar{\delta}^{(s)}\right\} \leq \left(\frac{1}{2}\right)^s (1+12\kappa)\,\bar{\delta}^{(0)}.$$

The first conclusion of this theorem comes from this inequality and the fact that $\bar{\delta}^{(0)} \leq \mathcal{P}_{\mu,\nu}(r; \bar{\mathbf{x}}^{(0)}) - \mathcal{D}_{\mu,\nu}(r; \bar{\mathbf{w}}^{(0)})$.

In the next, we prove the second conclusion of Theorem 2, namely, the expected number of stages before Algorithm 4 terminates. The argument in this proof is originally developed in Section C in the Appendix of (Lin et al., 2015). Let $\mathcal{S}(\zeta)$ be the (stage) index of outer iteration (i.e., s) when Algorithm 4 terminates. By Markov's inequality and (14), we have

$$\begin{aligned}
\text{Prob}(\mathcal{S}(\zeta) \geq s+1) &\leq \text{Prob}(\mathcal{P}_{\mu,\nu}(r; \bar{\mathbf{x}}^{(s)}) - \mathcal{D}_{\mu,\nu}(r; \bar{\mathbf{w}}^{(s)}) > \zeta) \\
&\leq \frac{\mathbb{E}[\mathcal{P}_{\mu,\nu}(r; \bar{\mathbf{x}}^{(s)}) - \mathcal{D}_{\mu,\nu}(r; \bar{\mathbf{w}}^{(s)})]}{\zeta} \\
&\leq (1+12\kappa)\left(\frac{1}{2}\right)^s \frac{\mathcal{P}_{\mu,\nu}(r; \bar{\mathbf{x}}^{(0)}) - \mathcal{D}_{\mu,\nu}(r; \bar{\mathbf{w}}^{(0)})}{\zeta}.
\end{aligned} \tag{42}$$

Let $\mathcal{S}_0 := 1 + 2\log\left(\frac{(2+24\kappa)[\mathcal{P}_{\mu,\nu}(r;\bar{\mathbf{x}}^{(0)})-\mathcal{D}_{\mu,\nu}(r;\bar{\mathbf{w}}^{(0)})]}{\zeta}\right)$. Using (42), we can show that

$$\begin{aligned}
\mathbb{E}S(\zeta) &= \sum_{s=0}^{\infty} \text{Prob}(S(\zeta) \geq s) \\
&\leq \mathcal{S}_0 + \sum_{s=\mathcal{S}_0}^{\infty} \text{Prob}(S(\zeta) \geq s) \\
&\leq \mathcal{S}_0 + \left(\frac{1}{2}\right)^{\mathcal{S}_0-1} \left(\sum_{s=0}^{\infty}\left(\frac{1}{2}\right)^s\right)(1+12\kappa)\,\frac{\mathcal{P}_{\mu,\nu}(r;\bar{\mathbf{x}}^{(0)}) - \mathcal{D}_{\mu,\nu}(r;\bar{\mathbf{w}}^{(0)})}{\zeta} \\
&\leq \mathcal{S}_0 + \left(\frac{1}{2}\right)^{\mathcal{S}_0-1} (2+24\kappa)\,\frac{\mathcal{P}_{\mu,\nu}(r;\bar{\mathbf{x}}^{(0)}) - \mathcal{D}_{\mu,\nu}(r;\bar{\mathbf{w}}^{(0)})}{\zeta} \\
&\leq \mathcal{S}_0 + 1
\end{aligned}$$

and the second conclusion follows. □

Proof of Theorem 3. Suppose CheckGap$(\hat{\mathbf{x}}^{(p)}, \hat{\mathbf{w}}^{(p)}, \varepsilon, \theta)$ = "Success", and thus, Algorithm 5 terminates when iteration index $p = P$. We first prove by induction that

$$\mathcal{P}(r;\hat{\mathbf{x}}^{(p)}) - \mathcal{D}(r;\hat{\mathbf{w}}^{(p)}) \leq \frac{\mathcal{P}(r;\hat{\mathbf{x}}^{(0)}) - \mathcal{D}(r;\hat{\mathbf{w}}^{(0)})}{2^p} = \frac{\zeta_0}{2^p} \quad \text{for } p = 0, 1, \ldots, P-2. \tag{43}$$

This inequality holds trivially for $p = 0$. Suppose it holds for iteration $p-1$ with $p-1 \leq P-3$. We then consider iteration $p \leq P-2$ in Algorithm 5 where SVRG$(\bar{\mathbf{x}}^{(0)}, \bar{\mathbf{w}}^{(0)}, \mu, \nu, \zeta, \varepsilon, \theta)$ is called with $\mathbf{x}^{(0)} = \hat{\mathbf{x}}^{(p)}$, $\mathbf{w}^{(0)} = \hat{\mathbf{w}}^{(p)}$, $\mu = \frac{\zeta_0}{2^{p+3}Q_{\mathbf{x}}}$, $\nu = \frac{\zeta_0}{2^{p+3}Q_{\mathbf{w}}}$, and $\zeta = \frac{\zeta_0}{2^{p+2}}$. Because $p+1 \leq P-1$ (i.e., this is not the last called of SVRG), we must have CheckGap$(\hat{\mathbf{x}}^{(p+1)}, \hat{\mathbf{w}}^{(p+1)}, \varepsilon, \theta)$ = "Continue". In other word, Algorithm 4 (SVRG) in this call is terminated because

$$\mathcal{P}_{\mu,\nu}(r;\hat{\mathbf{x}}^{(p+1)}) - \mathcal{D}_{\mu,\nu}(r;\hat{\mathbf{w}}^{(p+1)}) \leq \frac{\zeta_0}{2^{p+2}}.$$

According to this inequality and Lemma 4, we have

$$\mathcal{P}(r;\hat{\mathbf{x}}^{(p+1)}) - \mathcal{D}(r;\hat{\mathbf{w}}^{(p+1)}) \leq \mathcal{P}_{\mu,\nu}(r;\mathbf{x}) - \mathcal{D}_{\mu,\nu}(r;\mathbf{w}) + \mu Q_{\mathbf{x}} + \nu Q_{\mathbf{w}} \leq \frac{\zeta_0}{2^{p+2}} + \frac{\zeta_0}{2^{p+3}Q_{\mathbf{x}}}Q_{\mathbf{x}} + \frac{\zeta_0}{2^{p+3}Q_{\mathbf{w}}}Q_{\mathbf{w}} = \frac{\zeta_0}{2^{p+1}}$$

which implies our claim (43) by induction.

In the next, we want to show that Algorithm 5 satisfies the property of an affine minorant oracle. Suppose $r > f^*$ so that $H(r) < 0$. According to (43), with $p \geq \log_2\left(\frac{2\zeta_0\theta}{(\theta-1)|H(r)|}\right) \geq \log_2\left(\frac{\zeta_0\theta}{(\theta-1)|H(r)|}\right)$, Algorithm 5 can ensure $\mathcal{P}(r;\hat{\mathbf{x}}^{(p)}) - \mathcal{D}(r;\hat{\mathbf{w}}^{(p)}) \leq \frac{\theta-1}{\theta}|H(r)| \leq \frac{\theta-1}{\theta}|\mathcal{D}(r;\hat{\mathbf{w}}^{(p)})|$ which implies $\theta\mathcal{P}(r;\hat{\mathbf{x}}^{(p)}) \leq \mathcal{D}(r;\hat{\mathbf{w}}^{(p)})$.

Suppose $r \leq f^*$ so that $H(r) \geq 0$. We apply (43) to two cases, $H(r) \geq \frac{\varepsilon}{2}$ and $H(r) < \frac{\varepsilon}{2}$. In the case where $H(r) \geq \frac{\varepsilon}{2}$, with $p \geq \log_2\left(\frac{2\zeta_0\theta}{(\theta-1)\max\{|H(r)|,\varepsilon\}}\right) \geq \log_2\left(\frac{\zeta_0\theta}{(\theta-1)|H(r)|}\right)$, Algorithm 5 can ensure $\mathcal{P}(r;\hat{\mathbf{x}}^{(p)}) - \mathcal{D}(r;\hat{\mathbf{w}}^{(p)}) \leq \frac{\theta-1}{\theta}|H(r)| \leq \frac{\theta-1}{\theta}\mathcal{P}(r;\hat{\mathbf{w}}^{(p)})$ which implies $\mathcal{P}(r;\hat{\mathbf{x}}^{(p)}) \leq \theta\mathcal{D}(r;\hat{\mathbf{w}}^{(p)})$. In the case where $H(r) < \frac{\varepsilon}{2}$, with $p \geq \log_2\left(\frac{2\zeta_0\theta}{(\theta-1)\max\{|H(r)|,\varepsilon\}}\right) \geq \log_2\left(\frac{2\zeta_0}{\varepsilon}\right)$, Algorithm 5 can ensure $\mathcal{P}(r;\hat{\mathbf{x}}^{(p)}) - \mathcal{D}(r;\hat{\mathbf{w}}^{(p)}) \leq \frac{\varepsilon}{2}$ which implies $\mathcal{P}(r;\hat{\mathbf{x}}^{(p)}) \leq \mathcal{D}(r;\hat{\mathbf{w}}^{(p)}) + \frac{\varepsilon}{2} \leq H(r) + \frac{\varepsilon}{2} \leq \varepsilon$. Overall,

Based on these arguments, at least one of the three conditions in Algorithm 3 will be satisfied, and Algorithm 5 will terminate and return the desired $L(r)$, $U(r)$ and $S(r)$ within P iterations with

$$P \leq \begin{cases} \tilde{\mathcal{C}}(|H(r)|) & \text{if } r > f^* \\ \tilde{\mathcal{C}}(\max\{|H(r)|, \varepsilon\}) & \text{if } r \leq f^*. \end{cases} \tag{44}$$

where $\tilde{\mathcal{C}}(z) := \log_2\left(\frac{2\zeta_0\theta}{(\theta-1)z}\right)$. Note that the two upper bounds on P here correspond to the two cases of the complexity of the affine minorant oracle $\mathcal{A}$ in Definition 1.

Lastly, we analyze the total complexity of Algorithm 5 by analyzing the complexity of each call of SVRG. When calling SVRG in iteration p, the parameters are chosen as $\mu = \frac{\zeta_0}{2^{p+3}Q_{\mathbf{x}}}$, $\nu = \frac{\zeta_0}{2^{p+3}Q_{\mathbf{w}}}$ and $\zeta = \frac{\zeta_0}{2^{p+2}}$. Hence, by Lemma 4 and (43), we have

$$\mathcal{P}_{\mu,\nu}(r;\hat{\mathbf{x}}^{(p)}) - \mathcal{D}_{\mu,\nu}(r;\hat{\mathbf{w}}^{(p)}) \leq \mathcal{P}(r;\hat{\mathbf{x}}^{(p)}) - \mathcal{D}(r;\hat{\mathbf{w}}^{(p)}) + \frac{\zeta_0}{2^{p+3}Q_{\mathbf{x}}}Q_{\mathbf{x}} + \frac{\zeta_0}{2^{p+3}Q_{\mathbf{w}}}Q_{\mathbf{w}} \leq \frac{\zeta_0}{2^{p-1}}.$$

According to Theorem 2, the expected number of outer iterations in the pth call of SVRG is at most

$$\mathcal{S} \leq 2 + 2\log\left(\frac{(2+24\kappa)\left[\mathcal{P}_{\mu,\nu}(r;\hat{\mathbf{x}}^{(p)}) - \mathcal{D}_{\mu,\nu}(r;\hat{\mathbf{w}}^{(p)})\right]}{\zeta}\right) \leq 2 + 2\log\left(16 + 2^{2p+13}\frac{3\|A\|_{\max}^2 Q_{\mathbf{x}}Q_{\mathbf{w}}}{\zeta_0^2}\right)$$

and the number of inner iterations is

$$T = \left(\frac{4}{3} + 16\kappa\right)\log\left(\frac{18}{5}\right) = \left(\frac{4}{3} + 2^{2p+11}\frac{\|A\|_{\max}^2 Q_{\mathbf{x}}Q_{\mathbf{w}}}{\zeta_0^2}\right)\log\left(\frac{18}{5}\right)$$

This indicates the expected complexity of the pth call of SVRG is at most

$$\mathcal{C}_{SVRG}^{(p)} := nd\left[2 + 2\log\left(16 + 2^{2p+13}\frac{3\|A\|_{\max}^2 Q_{\mathbf{x}}Q_{\mathbf{w}}}{\zeta_0^2}\right)\right] + (n+d)\left(\frac{4}{3} + 2^{2p+11}\frac{\|A\|_{\max}^2 Q_{\mathbf{x}}Q_{\mathbf{w}}}{\zeta_0^2}\right)\log\left(\frac{18}{5}\right).$$

Given $\mathcal{C}_{SVRG}^{(p)}$ above and the upper bound (44) for P for two different cases ($r > f^*$ and $r \leq f^*$), the total expected complexity of Algorithm 5, when used as an affine minorant oracle $\mathcal{A}$ in Definition 1, is at most

$$\begin{aligned}
\sum_{p=0}^{P-1}\mathcal{C}_{SVRG}^{(p)} &\leq \sum_{p=0}^{P-1} nd\left[2 + 2\log\left(16 + 2^{2p+13}\frac{3\|A\|_{\max}^2 Q_{\mathbf{x}}Q_{\mathbf{w}}}{\zeta_0^2}\right)\right] + \sum_{p=0}^{P-1}(n+d)\left(\frac{4}{3} + 2^{2p+11}\frac{\|A\|_{\max}^2 Q_{\mathbf{x}}Q_{\mathbf{w}}}{\zeta_0^2}\right)\log\left(\frac{18}{5}\right) \\
&\leq 2Pnd + nd\log\left(1 + \frac{3\|A\|_{\max}^2 Q_{\mathbf{x}}Q_{\mathbf{w}}}{\zeta_0^2}\right)\left(\sum_{p=0}^{P-1}(4p+26)\right) \\
&\quad + \frac{4P(n+d)}{3} + \frac{\|A\|_{\max}^2 Q_{\mathbf{x}}Q_{\mathbf{w}}}{\zeta_0^2}(n+d)\log\left(\frac{18}{5}\right)\left(\sum_{p=0}^{P-1}2^{2p+11}\right) \\
&\leq 2Pnd + nd\log\left(1 + \frac{3\|A\|_{\max}^2 Q_{\mathbf{x}}Q_{\mathbf{w}}}{\zeta_0^2}\right)(2P(P-1) + 26P)) \\
&\quad + \frac{4P(n+d)}{3} + \frac{\|A\|_{\max}^2 Q_{\mathbf{x}}Q_{\mathbf{w}}}{\zeta_0^2}(n+d)\log\left(\frac{18}{5}\right)2^{11}\frac{4^P - 1}{3} \\
&\leq \begin{cases} \mathcal{C}(|H(r)|) & \text{if } r > f^* \\ \mathcal{C}(\max\{|H(r)|, \varepsilon\}) & \text{if } r \leq f^*, \end{cases}
\end{aligned}$$

where (after replacing P by (44))

$$\begin{aligned}
\mathcal{C}(z) &:= 2\tilde{\mathcal{C}}(z)nd + nd\log\left(1 + \frac{\|A\|_{\max}^2 Q_{\mathbf{x}}Q_{\mathbf{w}}}{\zeta_0^2}\right)\left(2\tilde{\mathcal{C}}(z)(\tilde{\mathcal{C}}(z)-1) + 26\tilde{\mathcal{C}}(z)\right) \\
&\quad + \frac{4\tilde{\mathcal{C}}(z)(n+d)}{3} + \frac{\|A\|_{\max}^2 Q_{\mathbf{x}}Q_{\mathbf{w}}}{\zeta_0^2}(n+d)\log\left(\frac{18}{5}\right)2^{11}\frac{4^{\tilde{\mathcal{C}}(z)} - 1}{3}.
\end{aligned}$$

Using the fact that $\tilde{\mathcal{C}}(z)$ is the logarithmic function $\tilde{\mathcal{C}}(z) = \log_2\left(\frac{2\zeta_0\theta}{(\theta-1)z}\right)$ such that $\tilde{O}\left(4^{\tilde{\mathcal{C}}(z)}\right) = \tilde{O}(\frac{1}{z^2})$, we conclude that $\mathcal{C}(z) = \tilde{O}\left(nd + (n+d)\frac{\|A\|_{\max}^2}{z^2}\right)$, which completes the proof. □

F. Proof of Lemma 4

For any $\mathbf{x} \in \mathcal{X}$, let

$$\mathbf{w}_{\mathbf{x}} \in \arg\max_{\mathbf{w}\in\mathcal{W}} K(r;\mathbf{x},\mathbf{w}) = \arg\min_{\mathbf{w}\in\mathcal{W}} -\boldsymbol{\alpha}^\top A\mathbf{x} + G_0(\mathbf{w})$$

where $G_0(\mathbf{w}) = \sum_{i=0}^{m} \sum_{j=1}^{n_i} \frac{y_i}{n_i} \phi_{ij}^*\left(\frac{\alpha_{ij}}{y_i}\right) + \mathbf{y}^\top \mathbf{r}$ which is $G_\nu(\mathbf{w})$ when $\nu = 0$. Note that, the minimization above has a similar form as the minimization in Line 12 of Algorithm 4, i.e., the updating equation for $\mathbf{w}^{(t+1)}$, if we replace τ by $+\infty$, ν by 0, and $\mathbf{v}^{(t)}$ by $A\mathbf{x}$. Therefore, by a similar proof as Lemma 3, we can show that $\mathbf{w}_\mathbf{x} \in \mathcal{W}_B$ with B defined as in Lemma 3.

Similarly, for any $\mathbf{x} \in \mathcal{X}$, let

$$\mathbf{w}_{\mathbf{x},\nu} \in \arg\max_{\mathbf{w}\in\mathcal{W}} K(r;\mathbf{x},\mathbf{w}) - \nu h_B(\mathbf{w}) = \arg\min_{\mathbf{w}\in\mathcal{W}} -\boldsymbol{\alpha}^\top A\mathbf{x} + G_\nu(\mathbf{w}).$$

The minimization above has a similar form as the minimization in the updating equation for $\mathbf{w}^{(t+1)}$ in Algorithm 4, if we replace τ by $+\infty$ and $\mathbf{v}^{(t)}$ by $A\mathbf{x}$. Then, by a similar proof as Lemma 3, we can also show that $\mathbf{w}_{\mathbf{x},\nu} \in \mathcal{W}_B$ with B defined as in Lemma 3.

By the definitions of $\mathcal{P}$ and $\mathcal{P}_{\mu,\nu}$, we can show that

$$\mathcal{P}(r;\mathbf{x}) - \mathcal{P}_{\mu,\nu}(r;\mathbf{x}) \le K(r;\mathbf{x},\mathbf{w}_\mathbf{x}) - K(r;\mathbf{x},\mathbf{w}_\mathbf{x}) + \nu h_B(\mathbf{w}_\mathbf{x}) - \frac{\mu\|\mathbf{x}\|_2^2}{2} \le \nu \max_{\mathbf{w}\in\mathcal{W}_B} h_B(\mathbf{w}) = \frac{\nu Q_\mathbf{w}}{2}$$

and

$$\mathcal{P}_{\mu,\nu}(r;\mathbf{x}) - \mathcal{P}(r;\mathbf{x}) \le K(r;\mathbf{x},\mathbf{w}_{\mathbf{x},\nu}) - \nu h_B(\mathbf{w}_{\mathbf{x},\nu}) + \frac{\mu\|\mathbf{x}\|_2^2}{2} - K(r;\mathbf{x},\mathbf{w}_\mathbf{x}) \le \mu \max_{\mathbf{x}\in\mathcal{X}} \frac{\|\mathbf{x}\|_2^2}{2} = \frac{\mu Q_\mathbf{x}}{2}$$

where we used the facts that h_B is non-negative over $\mathcal{W}_B$ and that $\mathbf{w}_\mathbf{x} \in \mathcal{W}_B$. Combining these two inequalities gives $|\mathcal{P}(r;\mathbf{x}) - \mathcal{P}_{\mu,\nu}(r;\mathbf{x})| \le \frac{\mu Q_\mathbf{x}}{2} + \frac{\nu Q_\mathbf{w}}{2}$. By a similar argument, we can also show that $|\mathcal{D}(r;\mathbf{x}) - \mathcal{D}_{\mu,\nu}(r;\mathbf{x})| \le \frac{\mu Q_\mathbf{x}}{2} + \frac{\nu Q_\mathbf{w}}{2}$ which, together with the previous inequality, leads to the conclusion of this Lemma.