A. Illustration of Algorithm 2

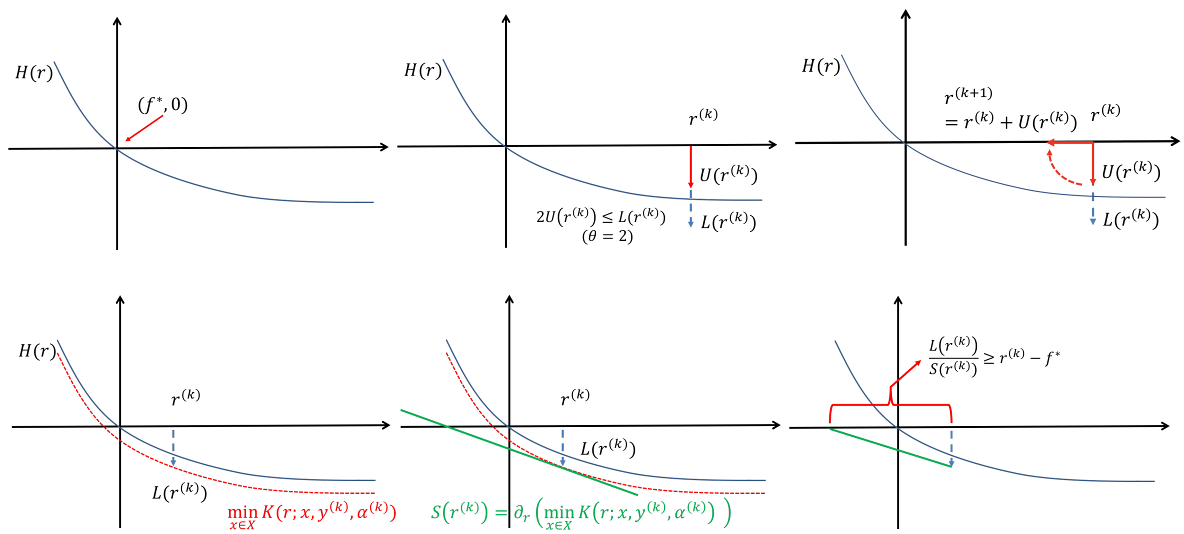

Figure 3. Illustration of AM-FLS Method. The figures on the top row depict the procedure to update $r^{(k)}$ using upper bound $U(r^{(k)})$. The figures on the bottom row show when to stop the algorithm.

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, $L(r^{(k)}) + \partial_r(\min_{\mathbf{x}\in\mathcal{X}} K(r; \mathbf{x}, \mathbf{y}^{(k)}, \boldsymbol{\alpha}^{(k)}))(r - r^{(k)})$ which is the green line in this figure. Therefore, we can choose $S(r^{(k)}) = \partial_r(\min_{\mathbf{x}\in\mathcal{X}} K(r; \mathbf{x}, \mathbf{y}^{(k)}, \boldsymbol{\alpha}^{(k)}))$ 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

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

Therefore, to prove this lemma, it suffices to prove $\|\tilde{\boldsymbol{\alpha}}_i^{(t)}\|_2 \leq B$ for all $t \geq 0$ 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 $B \geq \max_{\tilde{\alpha}_{ij} \in \text{dom}\phi_{ij}} \|\tilde{\boldsymbol{\alpha}}_i\|_2$. Since $\tilde{\boldsymbol{\alpha}}_i^{(t)}$ must stay in the domain of ϕ_{ij}^* according to (16), we have $\|\tilde{\boldsymbol{\alpha}}_i^{(t)}\|_2 \leq B$ for all $t \geq 0$ 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

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

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

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

for all $t \geq 0$ 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

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

whose optimality condition implies

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

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

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

whose optimality condition implies

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

or, equivalently,

$$2\nu\tilde{\alpha}_{ij}^* + \frac{2}{\tau}(\tilde{\alpha}_{ij}^* - \tilde{\alpha}_{ij}^{(t)}) \in 2\nu\tilde{\alpha}_{ij}^* + \frac{2}{\tau}(\tilde{\alpha}_{ij}^* - \tilde{\alpha}_{ij}^{(t)}) + \frac{1}{n_i}\partial\phi_{ij}^*(\tilde{\alpha}_{ij}^*) - \frac{1}{n_i}\xi_{ij}^\top \mathbf{x}^*. \tag{19}$$

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

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

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

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

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

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

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

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

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

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

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

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

On the other hand, in the last iteration 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 $\max\{1+\log_{2/\theta}(\frac{\max\{|S(r^{(0)})||f^*-r^{(0)}|,L(r^{(0)})\}}{\varepsilon}),2\}$ iterations so that the total complexity of Algorithm 1 is $\mathcal{C}(\varepsilon)\max\{1+\log_{2/\theta}(\frac{\max\{|S(r^{(0)})||f^*-r^{(0)}|,L(r^{(0)})\}}{\varepsilon}),2\}$. 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 $\mathcal{P}$.

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 $\mathcal{A}$ 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 Lemma 1(c) and convexity of $H(r)$, we can show that

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

From (21), the updating equation for $r^{(k+1)}$ and the fact that $H(r^{(k)}) \leq U(r^{(k)}) \leq L(r^{(k)})/\theta \leq H(r^{(k)})/\theta \leq 0$, we have

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

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

Recursively applying both inequalities gives

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

The inequality (21) for $r = r^{(k)}$, (24) and the property of $L(r^{(k)})$ together imply

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

for any given $k \geq \frac{2\theta}{\beta}\log\left(\frac{2\theta(r^{(0)}-f^*)}{|H(r^{(0)})|}\right)$. With the same k, the definition of $S(r^{(k)})$ and the fact that $S(r^{(k)}) \leq 0$ imply 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 ensure $-L(r^{(k)}) \leq \frac{-H(r^{(0)})\varepsilon}{2(r^{(0)}-f^*)} \leq -\varepsilon S(r^{(k)})$. Therefore, Algorithm 2 terminates after at most $\frac{2\theta}{\beta}\log\left(\frac{2\theta(r^{(0)}-f^*)}{|H(r^{(0)})|}\max\{\frac{r^{(0)}-f^*}{\varepsilon}, 1\}\right) = \frac{2\theta}{\beta}\log\left(\frac{2\theta}{\beta}\max\{\frac{r^{(0)}-f^*}{\varepsilon}, 1\}\right)$ iterations.

To obtain the overall complexity, consider an iteration k that is not the last iteration of Algorithm 2, i.e., $L(r^{(k)}) < \varepsilon S(r^{(k)})$. Without lose of generality, we assume $r^{(0)} - f^* > \varepsilon$. 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 $S(r^{(k)})$, 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 $|H(r^{(k)})| \geq \frac{|H(r^{(0)})|}{\theta(1+(r^{(0)}-f^*)/\varepsilon)} = \frac{\beta(r^{(0)}-f^*)}{\theta(1+(r^{(0)}-f^*)/\varepsilon)} \geq \frac{\varepsilon\beta}{2\theta}$ where the equality is by the definition of β and the inequality is by the fact that $r^{(0)} - f^* > \varepsilon$. Hence, the complexity of $\mathcal{A}$ in iteration k (non-terminating iteration) is at most

$$\mathcal{C}(|H(r^{(k)})|) \leq \mathcal{C}(\theta^{-1}\varepsilon\beta/2).$$

On the other hand, in the last iteration Algorithm 2, we have $-H(r^{(k)}) \geq \beta(r^{(k)} - f^*) \geq \frac{\beta}{2}(r^{(k-1)} - f^*) \geq \frac{\beta|H(r^{(k-1)})|}{2} \geq \frac{\beta^2\varepsilon}{4\theta}$ so that the complexity of $\mathcal{A}$ here is most

$$\mathcal{C}(|H(r^{(k)})|) \leq \mathcal{C}(\theta^{-1}\varepsilon\beta^2/4).$$

Hence, the total complexity Algorithm 2 is $\mathcal{C}(\theta^{-1}\varepsilon\beta^2/4)\frac{2\theta}{\beta}\log\left(\frac{2\theta}{\beta}\max\{\frac{r^{(0)}-f^*}{\varepsilon}, 1\}\right)$.

Lastly, we analyze the quality of the output solution from Algorithm 2. We note that the affine-minorant property of $S(r^{(k)})$ implies $H(r^{(k)} - L(r^{(k)})/S(r^{(k)})) \geq L(r^{(k)}) + S(r^{(k)})(r^{(k)} - L(r^{(k)})/S(r^{(k)}) - r^{(k)}) = 0$ such that we must have $r^{(k)} - L(r^{(k)})/S(r^{(k)}) \leq f^*$, which further ensures $r^{(k)} - f^* \leq L(r^{(k)})/S(r^{(k)}) \leq \varepsilon$ once Algorithm 2 terminates. At the last iteration, we then have $\mathcal{P}(r^{(k)}; \mathbf{x}_k) \leq U(r^{(k)}) \leq L(r^{(k)})/\theta \leq H(r^{(k)})/\theta < 0$ as $r^{(k)} > f^*$. 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.

□

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

$$\min_{\mathbf{w}\in\mathcal{W}} \left\{ \begin{array}{l} 2(1+B)^2\nu\sum_{i=0}^m y_i\ln y_i + \frac{2(1+B)^2}{\tau}\sum_{i=0}^m y_i \ln\left(\frac{y_i}{y_i'}\right) + \mathbf{y}^\top\mathbf{r} \\ +\sum_{i=0}^m \nu y_i \left\|\frac{\boldsymbol{\alpha}_i}{y_i}\right\|_2^2 + \sum_{i=0}^m \frac{y_i}{\tau}\left\|\frac{\boldsymbol{\alpha}_i}{y_i} - \frac{\boldsymbol{\alpha}_i'}{y_i'}\right\|_2^2 + \sum_{i=0}^m\sum_{j=1}^{n_i}\frac{y_i}{n_i}\phi_{ij}^*\left(\frac{\alpha_{ij}}{y_i}\right) - \sum_{i=0}^m y_i\left(\frac{\boldsymbol{\alpha}_i}{y_i}\right)^\top\mathbf{v}_i \end{array} \right\}. \tag{25}$$

We first fix $\mathbf{y}\in\Delta$ and only optimize $\boldsymbol{\alpha}\in\mathbb{R}^n$ in (25). It is easy to observe that each component $\boldsymbol{\alpha}_i$ in $\boldsymbol{\alpha}$ can be optimized independently. By changing variables with $\tilde{\boldsymbol{\alpha}}_i = \frac{\boldsymbol{\alpha}_i}{y_i}$ and $\tilde{\boldsymbol{\alpha}}_i' = \frac{\boldsymbol{\alpha}_i'}{y_i'}$, the minimization over $\boldsymbol{\alpha}_i$ can extracted from (25) and formulated as (13), which has a closed-form for many commonly used loss function ϕ_{ij}. Importantly, both the optimal value ρ_i and the optimal solution $\tilde{\boldsymbol{\alpha}}^*$ do not depend on y_i. Therefore, (25) is equivalent to

$$\min_{\mathbf{y}\in\Delta}\left\{2(1+B)^2\nu\sum_{i=0}^m y_i\ln y_i + \frac{2(1+B)^2}{\tau}\sum_{i=0}^m y_i\ln\left(\frac{y_i}{y_i'}\right) + \mathbf{y}^\top(\mathbf{r}+\boldsymbol{\rho})\right\}.$$

whose solution in a closed form is $y_i^\#$ defined in (14) which can be derived 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, only 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 $F_\mu(\mathbf{x}) := \frac{\mu\|\mathbf{x}\|_2^2}{2}$. Let $\mathbb{E}_t$ represent the conditional expectation conditioning on all the stochastic outcomes up to the end of iteration t. 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 $\mathbf{w} = (\mathbf{y}, \boldsymbol{\alpha})\in\mathcal{W}$,

$$\left(\mu + \frac{1}{\sigma}\right)\frac{\|\mathbf{x}-\mathbf{x}^{(t+1)}\|_2^2}{2} + (\mathbf{x}^{(t+1)})^\top\mathbf{u}^{(t)} + F_\mu(\mathbf{x}^{(t+1)}) + \frac{\|\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}\|_2^2}{2\sigma} \le \mathbf{x}^\top\mathbf{u}^{(t)} + F_\mu(\mathbf{x}) + \frac{\|\mathbf{x}-\mathbf{x}^{(t)}\|_2^2}{2\sigma} \tag{26}$$

$$\left(\nu + \frac{1}{\tau}\right)D(\mathbf{w}, \mathbf{w}^{(t+1)}) - (\boldsymbol{\alpha}^{(t+1)})^\top\mathbf{v}^{(t)} + G_\nu(\mathbf{w}^{(t+1)}) + \frac{D(\mathbf{w}^{(t+1)}, \mathbf{w}^{(t)})}{\tau} \le -\boldsymbol{\alpha}^\top\mathbf{v}^{(t)} + G_\nu(\mathbf{w}) + \frac{D(\mathbf{w}, \mathbf{w}^{(t)})}{\tau} \tag{27}$$

Let

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

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 strong convexity of F_μ with respect to Euclidean distance and the strong convexity of G_ν with respect to Bregman divergence D, we can show that

$$\tilde{\mathcal{P}}(\mathbf{x}) \ge \frac{\mu\|\mathbf{x}-\mathbf{x}^*\|_2^2}{2} \quad \text{and} \quad -\tilde{\mathcal{D}}(\mathbf{w}) \ge \nu D(\mathbf{w}, \mathbf{w}^*) \tag{28}$$

We choose $\mathbf{x} = \mathbf{x}^*$ in (26) and $\mathbf{w} = \mathbf{w}^*$ in (27) and add (26), and (27) 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)})\\
\leq\;& (\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}\tag{29}$$

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

$$\mathbb{E}_t[(\mathbf{x}^*-\mathbf{x}^{(t)})^\top(\mathbf{u}^{(t)}-A^\top\boldsymbol{\alpha}^{(t)})]=0 \quad\text{and}\quad \mathbb{E}_t[(\boldsymbol{\alpha}^*-\boldsymbol{\alpha}^{(t)})^\top(A\mathbf{x}^{(t)}-\mathbf{v}^{(t)})]=0 \tag{30}$$

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

Next, we study the three lines on the right hand side of (29), respectively. By the definition of $\mathbf{u}^{(t)}$, Cauchy-Schwarz inequality and Young's inequality, we have

$$\begin{aligned}
&\mathbb{E}_t\left[(\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)})^\top(\mathbf{u}^{(t)}-A^\top\boldsymbol{\alpha}^{(t)})\right]\\
\leq\;& \frac{1}{2a_t}\mathbb{E}_t\|\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}\|_2^2+\frac{a_t}{2}\mathbb{E}_t\|A^\top\bar{\boldsymbol{\alpha}}^{(s)}+nA_{l:}^\top\boldsymbol{\alpha}_l^{(t)}-nA_{l:}^\top\bar{\boldsymbol{\alpha}}_l^{(s)}-A^\top\boldsymbol{\alpha}^{(t)}\|_2^2\\
\leq\;& \frac{1}{2a_t}\mathbb{E}_t\|\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}\|_2^2+a_tn\max_l\|A_{l:}\|_2^2\|\boldsymbol{\alpha}^{(t)}-\boldsymbol{\alpha}^*\|_2^2+a_tn\max_l\|A_{l:}\|_2^2\|\bar{\boldsymbol{\alpha}}^{(s)}-\boldsymbol{\alpha}^*\|_2^2\\
\leq\;& \frac{1}{2a_t}\mathbb{E}_t\|\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}\|_2^2+2a_tn\max_l\|A_{l:}\|_2^2D(\mathbf{w}^*,\mathbf{w}^{(t)})+2a_tn\max_l\|A_{l:}\|_2^2D(\mathbf{w}^*,\bar{\mathbf{w}}^{(s)})
\end{aligned}\tag{31}$$

Similarly, we can prove that

$$\begin{aligned}
&\mathbb{E}_t\left[(\boldsymbol{\alpha}^{(t)}-\boldsymbol{\alpha}^{(t+1)})^\top(A\mathbf{x}^{(t)}-\mathbf{v}^{(t)})\right]\\
\leq\;& \frac{1}{2b_t}\mathbb{E}_t\|\boldsymbol{\alpha}^{(t)}-\boldsymbol{\alpha}^{(t+1)}\|_2^2+b_td\max_k\|A_{:k}\|_2^2\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2+b_td\max_k\|A_{:k}\|_2^2\|\bar{\mathbf{x}}^{(s)}-\mathbf{x}^*\|_2^2\\
\leq\;& \frac{1}{b_t}\mathbb{E}_tD(\mathbf{w}^{(t+1)},\mathbf{w}^{(t)})+b_td\max_k\|A_{:k}\|_2^2\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2+b_td\max_k\|A_{:k}\|_2^2\|\bar{\mathbf{x}}^{(s)}-\mathbf{x}^*\|_2^2
\end{aligned}\tag{32}$$

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

$$\mathbb{E}_t\left[(\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)})^\top A^\top(\boldsymbol{\alpha}^{(t)}-\boldsymbol{\alpha}^*)\right] \leq \frac{1}{2a_t}\mathbb{E}_t\|\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}\|_2^2+a_t\|A\|_2^2D(\mathbf{w}^*,\mathbf{w}^{(t)}) \tag{33}$$

$$\mathbb{E}_t\left[(\boldsymbol{\alpha}^{(t+1)}-\boldsymbol{\alpha}^{(t)})^\top A(\mathbf{x}^{(t)}-\mathbf{x}^*)\right] \leq \frac{1}{b_t}\mathbb{E}_tD(\mathbf{w}^{(t+1)},\mathbf{w}^{(t)})+\frac{b_t\|A\|_2^2}{2}\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2 \tag{34}$$

Choosing $a_t = 2\sigma$ and $b_t = 2\tau$ and applying (30), (31), (32), (33) and (34) to (29) lead to

$$\begin{aligned}
&\left(\mu+\frac{1}{\sigma}\right)\frac{\mathbb{E}_t\|\mathbf{x}^*-\mathbf{x}^{(t+1)}\|_2^2}{2}+\left(\nu+\frac{1}{\tau}\right)\mathbb{E}_tD(\mathbf{w}^*,\mathbf{w}^{(t+1)})+\tilde{\mathcal{P}}(\mathbf{x}^{(t+1)})-\tilde{\mathcal{D}}(\mathbf{w}^{(t+1)})\\
\leq\;& \left(2\tau\|A\|_2^2+4\tau d\max_k\|A_{:k}\|_2^2+\frac{1}{\sigma}\right)\frac{\|\mathbf{x}^*-\mathbf{x}^{(t)}\|_2^2}{2}+\left(2\sigma\|A\|_2^2+4\sigma n\max_l\|A_{l:}\|_2^2+\frac{1}{\tau}\right)D(\mathbf{w}^*,\mathbf{w}^{(t)})\\
&+2\tau d\max_k\|A_{:k}\|_2^2\|\mathbf{x}^*-\bar{\mathbf{x}}^{(s)}\|_2^2+4\sigma n\max_l\|A_{l:}\|_2^2D(\mathbf{w}^*,\bar{\mathbf{w}}^{(s)})
\end{aligned}\tag{35}$$

Note that the operator norm of A, i.e., $\|A\|_2$, satisfies $\|A\|_2 \leq \|A\|_{\max}$ so that $\kappa = \frac{2\|A\|_{\max}^2}{\mu\nu} = \frac{2\max\{d\max_k \|A_{:k}\|_2^2, n\max_l \|A_{l:}\|_2^2\}}{\mu\nu}$. Let η be a constant to be determined later. Choosing $\sigma = \frac{\eta}{\kappa\mu}$ and $\tau = \frac{\eta}{\kappa\nu}$ in (35), we obtain the following inequality

$$\begin{aligned} &\left(1+\frac{\kappa}{\eta}\right)\mu\mathbb{E}_t\frac{\|\mathbf{x}^*-\mathbf{x}^{(t+1)}\|_2^2}{2} + \left(1+\frac{\kappa}{\eta}\right)\nu\mathbb{E}_t D(\mathbf{w}^*,\mathbf{w}^{(t+1)}) + \mathbb{E}_t\tilde{\mathcal{P}}(\mathbf{x}^{(t+1)}) - \mathbb{E}_t\tilde{\mathcal{D}}(\mathbf{w}^{(t+1)}) \\ \leq\ & \left(4\eta+\frac{\kappa}{\eta}\right)\mu\frac{\|\mathbf{x}^*-\mathbf{x}^{(t)}\|_2^2}{2} + \left(4\eta+\frac{\kappa}{\eta}\right)\nu D(\mathbf{w}^*,\mathbf{w}^{(t)}) + 2\eta\mu\|\mathbf{x}^*-\bar{\mathbf{x}}^{(s)}\|_2^2 + 4\eta\nu D(\mathbf{w}^*,\bar{\mathbf{w}}^{(s)}), \end{aligned}$$

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

$$\frac{1}{1+\frac{\kappa}{\eta}}[\tilde{\mathcal{P}}(\mathbf{x}^{(t+1)}) - \tilde{\mathcal{D}}(\mathbf{w}^{(t+1)})] + \mathbb{E}\delta^{(t+1)} \leq \left(1-\frac{1-4\eta}{1+\frac{\kappa}{\eta}}\right)\mathbb{E}\delta^{(t)} + \frac{4\eta}{1+\frac{\kappa}{\eta}}\mathbb{E}\bar{\delta}^{(s)}, \tag{36}$$

where

$$\delta^{(t)} = \frac{\mu\mathbb{E}\|\mathbf{x}^*-\mathbf{x}^{(t)}\|_2^2}{2} + \nu\mathbb{E}D(\mathbf{w}^*,\mathbf{w}^{(t)})$$

and

$$\bar{\delta}^{(s)} = \frac{\mu\mathbb{E}\|\mathbf{x}^*-\bar{\mathbf{x}}^{(s)}\|_2^2}{2} + \nu\mathbb{E}D(\mathbf{w}^*,\bar{\mathbf{w}}^{(s)}).$$

Since $\delta^{(0)} = \bar{\delta}^{(s)}$ and $\delta^{(T)} = \bar{\delta}^{(s+1)}$, applying (36) recursively for $t = 0, 1, \ldots, T-1$ yields

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

Choosing $\eta = \frac{1}{20}$ in this inequality gives

$$\frac{1}{1+20\kappa}[\tilde{\mathcal{P}}(\bar{\mathbf{x}}^{(s+1)}) - \tilde{\mathcal{D}}(\bar{\mathbf{w}}^{(s+1)})] + \bar{\delta}^{(s+1)} \leq \left\{\left(1-\frac{1}{5/4+20\kappa}\right)^T + \frac{1}{4}\right\}\bar{\delta}^{(s)}$$

The following inequality is then obtained when $T = (5/4+20\kappa)\log(2)$ so that $\left(1-\frac{1}{5/4+20\kappa}\right)^T \leq \frac{1}{2}$:

$$\frac{1}{1+20\kappa}[\tilde{\mathcal{P}}(\bar{\mathbf{x}}^{(s+1)}) - \tilde{\mathcal{D}}(\bar{\mathbf{w}}^{(s+1)})] + \bar{\delta}^{(s+1)} \leq \frac{1}{2}\bar{\delta}^{(s)}. \tag{37}$$

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

$$\bar{\delta}^{(s)} \leq \left(\frac{1}{2}\right)^s \bar{\delta}^{(0)}. \tag{38}$$

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

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

Applying this inequality to (37) and combining it with (38) yield

$$\mathcal{P}_{\mu,\nu}(r;\bar{\mathbf{x}}^{(s)}) - \mathcal{D}_{\mu,\nu}(r;\bar{\mathbf{w}}^{(s)}) \leq (1+\kappa)\left\{\frac{1}{1+20\kappa}[\tilde{\mathcal{P}}(\bar{\mathbf{x}}^{(s)}) - \tilde{\mathcal{D}}(\bar{\mathbf{w}}^{(s)})] + \bar{\delta}^{(s)}\right\} \leq \left(\frac{1}{2}\right)^{s}(1+\kappa)\,\bar{\delta}^{(0)}$$

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 when Algorithm 4 terminates. By Markov's inequality, we have

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

Therefore, let $\mathcal{S}_0 = 2\log\left(\frac{(2+2\kappa)[\mathcal{P}_{\mu,\nu}(r;\bar{\mathbf{x}}^{(0)}) - \mathcal{D}_{\mu,\nu}(r;\bar{\mathbf{w}}^{(0)})]}{\zeta}\right)$. We can show that

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

and the second conclusion follows. □

Proof of Theorem 3. We first claim

$$\mathcal{P}(r;\hat{\mathbf{x}}^{(p)}) - \mathcal{D}(r;\hat{\mathbf{w}}^{(p)}) \leq \frac{\mathcal{P}(r;\hat{\mathbf{x}}^{(0)}) - \mathcal{D}(r;\hat{\mathbf{w}}^{(0)})}{2^p} = \frac{\zeta_0}{2^p} \tag{39}$$

Obviously, this is true for $p = 0$ by the definition of ζ_0. Suppose it holds for iteration p. According to Lemma 4 and Theorem 2, we have

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

which implies our claim (39) 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 (39), with $p = \log_2\left(\frac{\zeta_0\theta}{(\theta-1)|H(r)|}\right)$, 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 must consider two cases, $H(r) \geq \frac{\varepsilon}{2}$ and $H(r) < \frac{\varepsilon}{2}$, separately. In the case where $H(r) \geq \frac{\varepsilon}{2}$, with $p = \log_2\left(\frac{\zeta_0\theta}{(\theta-1)|H(r)|}\right)$, 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 $p = \log_2\left(\frac{2\zeta_0}{\varepsilon}\right)$, Algorithm 5 can ensure $\mathcal{P}(r;\hat{\mathbf{x}}^{(p)}) - \mathcal{D}(r;\hat{\mathbf{w}}^{(p)}) \leq \frac{\varepsilon}{2}$ which implies $\mathcal{P}(r;\hat{\mathbf{x}}^{(p)}) \leq \mathcal{D}(r;\hat{\mathbf{w}}^{(p)}) + \frac{\varepsilon}{2} \leq H(r) + \frac{\varepsilon}{2} \leq \varepsilon$. Based on the argument above, 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 $S(r)$, in no more than

$$P = \log_2 \left(\frac{2\zeta_0 \theta}{(\theta - 1) \max\{|H(r)|, \varepsilon\}} \right) \tag{40}$$

iterations.

In the pth call of SVRG in Algorithm 5 , the parameters are set 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,

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

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

$$\begin{aligned} \mathcal{S} &\leq 1 + 2\log\left(\frac{(2+2\kappa)\left[\mathcal{P}_{\mu,\nu}(r; \hat{\mathbf{x}}^{(p)}) - \mathcal{D}_{\mu,\nu}(r; \hat{\mathbf{w}}^{(p)})\right]}{\zeta} \right) \\ &\leq O\left(\log\left(\frac{[\|A\|_2^2 + \max\{d \max_k \|A_{:k}\|_2^2, n \max_l \|A_{l:}\|_2^2\}]\zeta_0}{2^p \mu\nu} \right)\right) \\ &= \tilde{O}\left(p\right). \end{aligned}$$

Given the upper bound (40) for the total number of calls of SVRG, the total expected complexity of Algorithm 5 is at most

$$\begin{aligned} & \sum_{p=0}^{P} \tilde{O}\left(\left(nd + (n+d) \frac{[\|A\|_2^2 + \max\{d \max_k \|A_{:k}\|_2^2, n \max_l \|A_{l:}\|_2^2\}]}{\mu\nu} \right) p \right) \\ \leq\ & \sum_{p=0}^{P} \tilde{O}\left(\left(nd + (n+d) Q_{\mathbf{x}} Q_{\mathbf{w}} [\|A\|_2^2 + \max\{d \max_k \|A_{:k}\|_2^2, n \max_l \|A_{l:}\|_2^2\}] 2^{2p} \right) p \right) \\ \leq\ & \tilde{O}\left(ndP\right) + \tilde{O}\left((n+d) Q_{\mathbf{x}} Q_{\mathbf{w}} [\|A\|_2^2 + \max\{d \max_k \|A_{:k}\|_2^2, n \max_l \|A_{l:}\|_2^2\}] P \right) \times \tilde{O}\left(\sum_{p=0}^{P} 2^{2p} \right) \\ =\ & \tilde{O}\left(nd + \frac{(n+d) Q_{\mathbf{x}} Q_{\mathbf{w}} [\|A\|_2^2 + \max\{d \max_k \|A_{:k}\|_2^2, n \max_l \|A_{l:}\|_2^2\}]}{\max\{|H(r)|^2, \varepsilon^2\}} \right) \\ =\ & \tilde{O}\left(nd + (n+d) \frac{\|A\|_{\max}^2}{\varepsilon^2} \right), \end{aligned}$$

where, in the first equality, we use the fact that P is a logarithmic term and $\tilde{O}\left(\sum_{p=0}^{P} 2^{2p}\right) = \tilde{O}\left(2^{2P}\right) = \tilde{O}\left(\frac{1}{\max\{|H(r)|^2, \varepsilon^2\}}\right)$. □