Quadratically Regularized Subgradient Methods for Weakly Convex Optimization with Weakly Convex Constraints Supplementary Materials

1. Appendix

In this section, we provide the proofs for the theoretical results in the paper.

1.1. Proof of Lemma 1

Proof. By KKT conditions, it holds that $\lambda_t \geq 0$ and $\lambda_t \left(g(\widehat{\mathbf{x}}_t) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|^2\right) = 0$. If $\lambda_t = 0$, there is nothing to show. So, we focus on the case that $\lambda_t > 0$ and $g(\widehat{\mathbf{x}}_t) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|^2 = 0$. Note that $\mathbf{x}_0$ is an ϵ^2-feasible solution. Using the definitions of $\mathcal{A}(\mathbf{x}_t, \hat{\rho}, \hat{\epsilon}, \delta/T)$ and $\hat{\epsilon}$ and the union bound, we can show that the iterate $\mathbf{x}_t$ generated by Algorithm 1 is an ϵ^2-feasible solution for any t with a probability of at least $1 - \delta$.

Let $\tilde{\mathbf{x}}_t \equiv \underset{\mathbf{x}\in\mathcal{X}}{\arg\min}\{g(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \mathbf{x}_t\|^2\}$. According to Assumption 1B, the fact that $\mathbf{x}_t$ is ϵ^2-feasible, and the fact that $\hat{\rho} \leq \rho + \rho_\epsilon$, we have

$$-\sigma_\epsilon \geq \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) + \frac{\rho + \rho_\epsilon}{2}\|\mathbf{x} - \mathbf{x}_t\|^2 \geq \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \mathbf{x}_t\|^2 = g(\tilde{\mathbf{x}}_t) + \frac{\hat{\rho}}{2}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t\|^2. \tag{1}$$

As a result, the Lagrangian multiplier λ_t is well-defined and satisfies the optimality condition below together with $\widehat{\mathbf{x}}_t$:

$$\mathbf{0} \in \partial f(\widehat{\mathbf{x}}_t) + \hat{\rho}(\widehat{\mathbf{x}}_t - \mathbf{x}_t) + \lambda_t(\partial g(\widehat{\mathbf{x}}_t) + \hat{\rho}(\widehat{\mathbf{x}}_t - \mathbf{x}_t)) + \widehat{\boldsymbol{\zeta}}_t, \tag{2}$$

for some $\widehat{\boldsymbol{\zeta}}_t \in \mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}}_t)$.

Since $g(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \mathbf{x}_t\|^2 + \mathbf{1}_{\mathcal{X}}(\mathbf{x})$ is $(\hat{\rho} - \rho)$-strongly convex in $\mathbf{x}$ and $\frac{\widehat{\boldsymbol{\zeta}}_t}{\lambda_t} \in \mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}}_t) = \partial\mathbf{1}_{\mathcal{X}}(\widehat{\mathbf{x}}_t)$, we have

$$\begin{aligned} g(\tilde{\mathbf{x}}_t) + \frac{\hat{\rho}}{2}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t\|^2 &\geq g(\widehat{\mathbf{x}}_t) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + \langle \partial g(\widehat{\mathbf{x}}_t) + \hat{\rho}(\widehat{\mathbf{x}}_t - \mathbf{x}_t) + \frac{\widehat{\boldsymbol{\zeta}}_t}{\lambda_t}, \tilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\rangle + \frac{\hat{\rho} - \rho}{2}\|\tilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2 \\ &= \langle \partial g(\widehat{\mathbf{x}}_t) + \hat{\rho}(\widehat{\mathbf{x}}_t - \mathbf{x}_t) + \widehat{\boldsymbol{\zeta}}_t/\lambda_t, \tilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\rangle + \frac{\hat{\rho} - \rho}{2}\|\tilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2. \end{aligned}$$

Applying (1) to the inequality above and arranging terms give

$$\begin{aligned} -\sigma_\epsilon - \frac{(\hat{\rho} - \rho)\|\tilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2}{2} &\geq \langle \partial g(\widehat{\mathbf{x}}_t) + \hat{\rho}(\widehat{\mathbf{x}}_t - \mathbf{x}_t) + \widehat{\boldsymbol{\zeta}}_t/\lambda_t, \tilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\rangle \\ &\geq -\frac{\|\partial g(\widehat{\mathbf{x}}_t) + \hat{\rho}(\widehat{\mathbf{x}}_t - \mathbf{x}_t) + \widehat{\boldsymbol{\zeta}}_t/\lambda_t\|^2}{2(\hat{\rho} - \rho)} - \frac{(\hat{\rho} - \rho)\|\tilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2}{2}, \end{aligned}$$

which implies $\|\partial g(\widehat{\mathbf{x}}_t) + \hat{\rho}(\widehat{\mathbf{x}}_t - \mathbf{x}_t) + \widehat{\boldsymbol{\zeta}}_t/\lambda_t\|^2 \geq 2\sigma_\epsilon(\hat{\rho} - \rho)$.

Using this lower bound on $\|\partial g(\widehat{\mathbf{x}}_t) + \hat{\rho}(\widehat{\mathbf{x}}_t - \mathbf{x}_t) + \widehat{\boldsymbol{\zeta}}_t/\lambda_t\|^2$ and (2), we have that

$$\lambda_t = \frac{\|\partial f(\widehat{\mathbf{x}}_t) + \hat{\rho}(\widehat{\mathbf{x}}_t - \mathbf{x}_t)\|}{\|\partial g(\widehat{\mathbf{x}}_t) + \hat{\rho}(\widehat{\mathbf{x}}_t - \mathbf{x}_t) + \widehat{\boldsymbol{\zeta}}_t/\lambda_t\|} \leq \frac{M + \hat{\rho}D}{\sqrt{2\sigma_\epsilon(\hat{\rho} - \rho)}}$$

for all t with a probability of at least $1 - \delta$, where we have used Assumption 1C and Assumption 1F in the inequality. □

1.2. Proof of Theorem 1

Proof. Since $\mathbf{x}_{t+1} = \mathcal{A}(\mathbf{x}_t, \hat{\rho}, \hat{\epsilon}, \delta/T)$, the definition of $\mathcal{A}$ and the union bound imply that the following inequalities hold for $t = 0, \dots, T-1$ with a probability of at least $1-\delta$.

$$f(\mathbf{x}_{t+1}) + \frac{\hat{\rho}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - f(\widehat{\mathbf{x}}_t) - \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \leq \hat{\epsilon}^2, \quad g(\mathbf{x}_{t+1}) + \frac{\hat{\rho}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \leq \hat{\epsilon}^2. \tag{3}$$

Let λ_t be the optimal Lagrangian multiplier corresponding to $\widehat{\mathbf{x}}_t$. Then $\widehat{\mathbf{x}}_t$ is also the optimal solution of the Lagrangian function $\mathcal{L}(\mathbf{x}) \equiv f(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \mathbf{x}_t\|^2 + \lambda_t(g(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \mathbf{x}_t\|^2)$. Since $\mathcal{L}(\mathbf{x})$ is $(1+\lambda_t)(\hat{\rho} - \rho)$-strongly convex, we have

$$\begin{aligned}
\frac{(1+\lambda_t)(\hat{\rho}-\rho)}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 \quad &\leq \quad f(\mathbf{x}_t) + \frac{\hat{\rho}}{2}\|\mathbf{x}_t - \mathbf{x}_t\|^2 + \lambda_t(g(\mathbf{x}_t) + \frac{\hat{\rho}}{2}\|\mathbf{x}_t - \mathbf{x}_t\|^2) \\
&\quad - \left[f(\widehat{\mathbf{x}}_t) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + \lambda_t(g(\widehat{\mathbf{x}}_t) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|^2)\right] \\
&= \quad f(\mathbf{x}_t) - f(\widehat{\mathbf{x}}_t) + \lambda_t g(\mathbf{x}_t) - \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|^2,
\end{aligned} \tag{4}$$

where we use the complementary slackness, i.e., $\lambda_t(g(\widehat{\mathbf{x}}_t) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|^2) = 0$ in the equality above. Organizing the terms in the first inequality of (3), we get

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\widehat{\mathbf{x}}_t) + \hat{\epsilon}^2 + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|^2 - \frac{\hat{\rho}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&\leq f(\widehat{\mathbf{x}}_t) + \hat{\epsilon}^2 + f(\mathbf{x}_t) - f(\widehat{\mathbf{x}}_t) + \lambda_t g(\mathbf{x}_t) - \frac{(1+\lambda_t)(\hat{\rho}-\rho)}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 \\
&= f(\mathbf{x}_t) + \lambda_t g(\mathbf{x}_t) - \frac{(1+\lambda_t)(\hat{\rho}-\rho)}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \hat{\epsilon}^2
\end{aligned}$$

where second inequality is because of (4). The inequality above can be written as

$$\frac{(1+\lambda_t)(\hat{\rho}-\rho)}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \lambda_t g(\mathbf{x}_t) + \hat{\epsilon}^2 \tag{5}$$

Summing up inequality (5) from $t = 0, 1, \dots, T-1$, we have

$$\sum_{t=0}^{T-1} \frac{(1+\lambda_t)(\hat{\rho}-\rho)}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 \leq f(\mathbf{x}_0) - f_{\text{lb}} + \sum_{t=0}^{T-1}\lambda_t g(\mathbf{x}_t) + T\hat{\epsilon}^2,$$

where f_{lb} is introduced in Assumption 1D. Note that $g(\mathbf{x}_t) \leq g(\mathbf{x}_t) + \frac{\hat{\rho}}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \leq \hat{\epsilon}^2$ because of the property of $\mathcal{A}$. So we have

$$\sum_{t=0}^{T-1} \frac{(\hat{\rho}-\rho)}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 \leq \sum_{t=0}^{T-1} \frac{(1+\lambda_t)(\hat{\rho}-\rho)}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 \leq f(\mathbf{x}_0) - f_{\text{lb}} + \sum_{t=0}^{T-1}\lambda_t\hat{\epsilon}^2 + T\hat{\epsilon}^2.$$

Dividing both sides by $T(\hat{\rho}-\rho)/2$, we have

$$\begin{aligned}
\mathbb{E}_R\|\mathbf{x}_R - \widehat{\mathbf{x}}_R\|^2 = \frac{1}{T}\sum_{t=0}^{T-1}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 &\leq \frac{2(f(\mathbf{x}_0) - f_{\text{lb}})}{T(\hat{\rho}-\rho)} + \frac{2}{T(\hat{\rho}-\rho)}\sum_{t=0}^{T-1}(1+\lambda_t)\hat{\epsilon}^2 \\
&\leq \frac{2(f(\mathbf{x}_0) - f_{\text{lb}})}{T(\hat{\rho}-\rho)} + \frac{2\hat{\epsilon}^2}{(\hat{\rho}-\rho)}\left(\frac{M + \hat{\rho}D}{\sqrt{2\sigma_\epsilon(\hat{\rho}-\rho)}} + 1\right) \\
&\leq \frac{\epsilon^2}{2} + \frac{\epsilon^2}{2} = \epsilon^2
\end{aligned}$$

with a probability of at least $1-\delta$, where the second inequality is by Lemma 1 and the last inequality follows the definitions of T and $\hat{\epsilon}$. □

1.3. Proof of Theorem 2

Proof. For simplicity of notation, we defined $\mu := \hat{\rho} - \rho$. Let $J := \{0, 1, \ldots, K-1\} \backslash I$ where I is generated in Algorithm 2 when it terminates.

Suppose $k \in I$, namely, $G(\mathbf{z}_k) \leq \hat{\epsilon}^2$ is satisfied in iteration k. Algorithm 2 will update $\mathbf{z}_{k+1}$ using $F'(\mathbf{z}_k)$. Following the standard analysis of subgradient decent method, we can get

$$\begin{aligned} F(\mathbf{z}_k) - F(\widehat{\mathbf{x}}_t) &\leq \gamma_k (M^2 + \hat{\rho}^2 D^2) + (\frac{1}{2\gamma_k} - \frac{\mu}{2}) \|\mathbf{z}_k - \widehat{\mathbf{x}}_t\|^2 - \frac{\|\mathbf{z}_{k+1} - \widehat{\mathbf{x}}_t\|^2}{2\gamma_k} \\ &= \frac{2(M^2 + \hat{\rho}^2 D^2)}{\mu(k+2)} + (\frac{\mu(k+2)}{4} - \frac{2\mu}{4}) \|\mathbf{z}_k - \widehat{\mathbf{x}}_t\|^2 - \frac{\mu(k+2)}{4} \|\mathbf{z}_{k+1} - \widehat{\mathbf{x}}_t\|^2 \\ &= \frac{2(M^2 + \hat{\rho}^2 D^2)}{\mu(k+2)} + \frac{\mu k}{4} \|\mathbf{z}_k - \widehat{\mathbf{x}}_t\|^2 - \frac{\mu(k+2)}{4} \|\mathbf{z}_{k+1} - \widehat{\mathbf{x}}_t\|^2 \end{aligned} \tag{6}$$

Multiplying $k+1$ to the both sides of 6, we can get

$$\begin{aligned} (k+1)(F(\mathbf{z}_k) - F(\widehat{\mathbf{x}}_t)) &\leq \frac{2(M^2 + \hat{\rho}^2 D^2)(k+1)}{\mu(k+2)} + \frac{\mu k(k+1)}{4} \|\mathbf{z}_k - \widehat{\mathbf{x}}_t\|^2 - \frac{\mu(k+1)(k+2)}{4} \|\mathbf{z}_{k+1} - \widehat{\mathbf{x}}_t\|^2 \\ &\leq \frac{2(M^2 + \hat{\rho}^2 D^2)}{\mu} + \frac{\mu k(k+1)}{4} \|\mathbf{z}_k - \widehat{\mathbf{x}}_t\|^2 - \frac{\mu(k+1)(k+2)}{4} \|\mathbf{z}_{k+1} - \widehat{\mathbf{x}}_t\|^2 \end{aligned} \tag{7}$$

Suppose $k \in J$, namely, $G(\mathbf{z}_k) \leq \hat{\epsilon}^2$ is not satisfied in iteration k. Algorithm 2 will update $\mathbf{z}_{k+1}$ using $G'(\mathbf{z}_k)$. Similarly, we can get

$$(k+1)(G(\mathbf{z}_k) - G(\widehat{\mathbf{x}}_t)) \leq \frac{2(M^2 + \hat{\rho}^2 D^2)}{\mu} + \frac{\mu k(k+1)}{4} \|\mathbf{z}_k - \widehat{\mathbf{x}}_t\|^2 - \frac{\mu(k+1)(k+2)}{4} \|\mathbf{z}_{k+1} - \widehat{\mathbf{x}}_t\|^2 \tag{8}$$

Summing up inequalities (7) and (8) from $k = 0, \ldots, K-1$ and dropping the non-negative terms, we obtain

$$\sum_{k \in I} (k+1)(F(\mathbf{z}_k) - F(\widehat{\mathbf{x}}_t)) + \sum_{k \in J} (k+1)(G(\mathbf{z}_k) - G(\widehat{\mathbf{x}}_t)) \leq \frac{2K(M^2 + \hat{\rho}^2 D^2)}{\mu} \tag{9}$$

Because $G(\mathbf{z}_k) > \hat{\epsilon}^2$ when $k \in J$ and $G(\widehat{\mathbf{x}}_t) \leq 0$, the inequality above implies

$$\sum_{k \in I} (k+1)(F(\mathbf{z}_k) - F(\widehat{\mathbf{x}}_t)) + \sum_{k \in J} (k+1)\hat{\epsilon}^2 \leq \frac{2K(M^2 + \hat{\rho}^2 D^2)}{\mu} \tag{10}$$

Rearranging terms gives

$$\begin{aligned} \sum_{k \in I} (k+1)(F(\mathbf{z}_k) - F(\widehat{\mathbf{x}}_t)) &\leq \sum_{k \in I} (k+1)\hat{\epsilon}^2 - \sum_{k=0}^{K-1} (k+1)\hat{\epsilon}^2 + \frac{2K(M^2 + \hat{\rho}^2 D^2)}{\mu} \\ &\leq \sum_{k \in I} (k+1)\hat{\epsilon}^2 - \frac{K(K+1)}{2}\hat{\epsilon}^2 + \frac{2K(M^2 + \hat{\rho}^2 D^2)}{\mu}. \end{aligned}$$

Given that $K \geq \frac{4(M^2 + \hat{\rho}^2 D^2)}{\mu \hat{\epsilon}^2}$, the summation of the last two terms in the inequality above is non-positive. As a result, we have

$$\sum_{k \in I} (k+1)(F(\mathbf{z}_k) - F(\widehat{\mathbf{x}}_t)) \leq \sum_{k \in I} (k+1)\hat{\epsilon}^2$$

Dividing both sides by $\sum_{k \in I}(k+1)$ and using the convexity of F, we obtain $F(\mathbf{x}_{t+1}) - F(\widehat{\mathbf{x}}_t) \leq \hat{\epsilon}^2$. As the same time, the convexity of G ensures $G(\mathbf{x}_{t+1}) \leq \frac{\sum_{k \in I}(k+1)G(\mathbf{z}_k)}{\sum_{k \in I}(k+1)} \leq \hat{\epsilon}^2$.

Hence, Algorithm 2 can be used as an oracle to solve (9) and the complexity of Algorithm 1 will be

$$TK = O\left(\frac{(f(\mathbf{x}_0) - f_{\text{lb}})(M^2 + \hat{\rho}^2 D^2)}{\epsilon^4 (\hat{\rho} - \rho)^3} \Big(\frac{M + \hat{\rho} D}{\sqrt{\sigma_\epsilon (\hat{\rho} - \rho)}} + 1 \Big) \right).$$

Note that, Algorithm 2 is deterministic so that the complexity above does not depend on δ. □

1.4. Proof of Theorem 3

Proof. According to Assumption 1B and the factor that $\mathbf{x}_t$ is ϵ^2-feasible with a high probability, Assumption 2 (The Slater's condition) in (Yu et al., 2017) holds for the subproblem (9) with a high probability. According to Theorem 4 in (Yu et al., 2017), Algorithm 3 guarantees

$$F(\mathbf{x}_{t+1}) - F(\widehat{\mathbf{x}}_t) \le \mathcal{B}_1(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta) \tag{11}$$

with a probability of at least $1-\delta$, where

$$\begin{aligned} & \mathcal{B}_1(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta) \\ \equiv\ & \frac{D^2 + \tilde{M}_1^2/4 + (\tilde{M}_0 + \sqrt{m}\tilde{M}_1 D)^2/2 + \log^{0.5}\left(\frac{1}{\delta}\right)\tilde{M}_0 \Lambda(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta)}{\sqrt{K}}, \end{aligned} \tag{12}$$

$$\begin{aligned} \Lambda(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta) \equiv\ & \frac{\sigma_\epsilon}{2} + (\tilde{M}_0 + \sqrt{m}\tilde{M}_1 D) + \frac{2D^2}{\sigma_\epsilon} + \frac{2\tilde{M}_1 D + (\tilde{M}_0 + \sqrt{m}\tilde{M}_1 D)^2}{\sigma_\epsilon} \\ & + \tilde{\Lambda}(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta) + \frac{8(\tilde{M}_0 + \sqrt{m}\tilde{M}_1 D)^2}{\sigma_\epsilon}\log\left(\frac{2K}{\delta}\right) = O(\log(K/\delta)), \end{aligned} \tag{13}$$

and

$$\tilde{\Lambda}(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta) \equiv \frac{8(\tilde{M}_0 + \sqrt{m}\tilde{M}_1 D)^2}{\sigma_\epsilon}\log\left[1 + \frac{32(\tilde{M}_0 + \sqrt{m}\tilde{M}_1 D)^2}{\sigma_\epsilon^2}\exp\left(\frac{\sigma_\epsilon}{8(\tilde{M}_0 + \sqrt{m}\tilde{M}_1 D)}\right)\right].$$

According to equation (22) in (Yu et al., 2017), Algorithm 3 guarantees

$$F_i(\mathbf{x}_{t+1}) \le \frac{\|(Q_K^1, Q_K^2, \ldots, Q_K^m)\|}{K} + \frac{\tilde{M}_1^2}{\sqrt{K}} + \frac{\sqrt{m}\tilde{M}_1^2}{2K^2}\sum_{k=0}^{K-1}\|(Q_k^1, Q_k^2, \ldots, Q_k^m)\| \tag{14}$$

for $i = 1, \ldots, m$. It is also shown in Theorem 3 in (Yu et al., 2017) that

$$\|(Q_k^1, Q_k^2, \ldots, Q_k^m)\| \le \sqrt{K}\Lambda(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta) \tag{15}$$

for $k = 0, 1, \ldots, K$ with a probability of at least $1-\delta$. Applying (15) to (14) and organizing terms, we obtain

$$F_i(\mathbf{x}_{t+1}) \le \mathcal{B}_2(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta) \tag{16}$$

with a probability of at least $1-\delta$, where

$$\begin{aligned} & \mathcal{B}_2(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta) \\ \equiv\ & \frac{\Lambda(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta) + \tilde{M}_1^2 + \Lambda(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta)\sqrt{m}\tilde{M}_1^2/2}{\sqrt{K}} \end{aligned} \tag{17}$$

To ensure Algorithm 3 is an oracle for (9), it suffices to choose the K large enough so that the left hand sides of (11) and (16) are both no more than $\hat{\epsilon}^2$. Because $\Lambda(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta) = O(\log(K/\delta))$. It suffices to choose $K = \tilde{O}(\frac{1}{\hat{\epsilon}^4}\log(\frac{1}{\delta}))$. Hence, Algorithm 3 can be used as an oracle to solve (9) and the complexity of Algorithm 1 will be

$$TK = \tilde{O}\left(\frac{1}{\epsilon^6}\right).$$

$\square$

References

Yu, H., Neely, M., and Wei, X. Online convex optimization with stochastic constraints. In *Advances in Neural Information Processing Systems*, pp. 1428–1438, 2017.