X-MEN: Guaranteed XOR-Maximum Entropy Constrained Inverse Reinforcement Learning (Supplementary Material)

Fan Ding[1] **Yexiang Xue**[1]

[1]Department of Computer Science, Purdue University, West Lafayette, Indiana, USA

1 PROOF OF THEOREM 4

Before proving Theorem 4, we need to bound two important terms shown in Lemma 1.

Lemma 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function and $\theta^* = argmin_\theta f(\theta)$. In iteration t, g_t is the estimated gradient. Suppose $||\mathbb{E}[g_t^+]||_2 \leq G$, $||\mathbb{E}[g_t^-]||_2 \leq G$, and $||\theta_t - \theta^*||_2 \leq R$. If there exists a constant $c \geq 1$ s.t. $\frac{1}{c}[\nabla f(\theta_t)]^+ \leq \mathbb{E}[g_t^+] \leq c[\nabla f(\theta_t)]^+$ and $c[\nabla f(\theta_t)]^- \leq \mathbb{E}[g_t^-] \leq \frac{1}{c}[\nabla f(\theta_t)]^-$, then we have*

$$\frac{1}{c}||\mathbb{E}[g_t]||_2^2 \leq \langle \nabla f(\theta_t), \mathbb{E}[g_t]\rangle + 2(c - \frac{1}{c})G^2. \quad (1)$$

$$\langle \nabla f(\theta_t), \theta_t - \theta^*\rangle \leq c\langle \mathbb{E}[g_t], \theta_t - \theta^*\rangle + 2(c - \frac{1}{c})GR. \quad (2)$$

1.1 PROOF OF LEMMA 1

Proof. (Lemma 1) Since we have the constant bound that

$$\frac{1}{c}[\nabla f(\theta_t)]^+ \leq \mathbb{E}[g_t^+] \leq c[\nabla f(\theta_t)]^+. \quad (3)$$

$$c[\nabla f(\theta_t)]^- \leq \mathbb{E}[g_t^-] \leq \frac{1}{c}[\nabla f(\theta_t)]^-. \quad (4)$$

and because of $g_t^+ \geq \mathbf{0}$ and $g_t^- \leq \mathbf{0}$ we can obtain

$$\begin{aligned}\frac{1}{c}||\mathbb{E}[g_t^+]||_2^2 = \frac{1}{c}\langle \mathbb{E}[g_t^+], \mathbb{E}[g_t^+]\rangle &\leq \langle [\nabla f(\theta_t)]^+, \mathbb{E}[g_t^+]\rangle \\ &\leq c\langle \mathbb{E}[g_t^+], \mathbb{E}[g_t^+]\rangle = c||\mathbb{E}[g_t^+]||_2^2. \\ \frac{1}{c}||\mathbb{E}[g_t^-]||_2^2 = \frac{1}{c}\langle \mathbb{E}[g_t^-], \mathbb{E}[g_t^-]\rangle &\leq \langle [\nabla f(\theta_t)]^-, \mathbb{E}[g_t^-]\rangle \\ &\leq c\langle \mathbb{E}[g_t^-], \mathbb{E}[g_t^-]\rangle = c||\mathbb{E}[g_t^-]||_2^2.\end{aligned}$$

For cross terms, we have:

$$\langle [\nabla f(\theta_t)]^+, \mathbb{E}[g_t^-]\rangle \geq c\langle [\nabla \mathbb{E}[g_t^+], \mathbb{E}[g_t^-]\rangle$$
$$\langle [\nabla f(\theta_t)]^-, \mathbb{E}[g_t^+]\rangle \geq c\langle [\nabla \mathbb{E}[g_t^-], \mathbb{E}[g_t^+]\rangle$$

Notice that:

$$\begin{aligned}&\frac{1}{c}||\mathbb{E}[g_t]||_2^2 \\ =&\frac{1}{c}||\mathbb{E}[g_t^+] + \mathbb{E}[g_t^-]||_2^2 \\ =&\frac{1}{c}(||\mathbb{E}[g_t^+]||_2^2 + ||\mathbb{E}[g_t^-]||_2^2 + 2\langle \mathbb{E}[g_t^+], \mathbb{E}[g_t^-]\rangle)\end{aligned}$$

Then we can further derive:

$$\begin{aligned}&\frac{1}{c}||\mathbb{E}(g_t)||_2^2 \\ \leq&\langle [\nabla f(\theta_t)]^+, \mathbb{E}[g_t^+]\rangle + \langle [\nabla f(\theta_t)]^-, \mathbb{E}[g_t^-]\rangle + \\ &\frac{1}{c^2}\langle [\nabla f(\theta_t)]^+, \mathbb{E}[g_t^-]\rangle + \frac{1}{c^2}\langle [\nabla f(\theta_t)]^-, \mathbb{E}[g_t^+]\rangle \\ =&\langle [\nabla f(\theta_t)]^+, \mathbb{E}[g_t^+]\rangle + \langle [\nabla f(\theta_t)]^-, \mathbb{E}[g_t^-]\rangle + \\ &\langle [\nabla f(\theta_t)]^+, \mathbb{E}[g_t^-]\rangle + \langle [\nabla f(\theta_t)]^-, \mathbb{E}[g_t^+]\rangle + \\ &(\frac{1}{c^2} - 1)\left(\langle [\nabla f(\theta_t)]^+, \mathbb{E}[g_t^-]\rangle + \langle [\nabla f(\theta_t)]^-, \mathbb{E}[g_t^+]\rangle\right) \\ =&\langle \nabla f(\theta_t), \mathbb{E}[g_t]\rangle + (\frac{1}{c^2} - 1)\langle [\nabla f(\theta_t)]^+, \mathbb{E}[g_t^-]\rangle \\ &+ (\frac{1}{c^2} - 1)\langle [\nabla f(\theta_t)]^-, \mathbb{E}[g_t^+]\rangle \\ \leq&\langle \nabla f(\theta_t), \mathbb{E}[g_t]\rangle + (\frac{1}{c} - c)\langle \mathbb{E}[g_t^+], \mathbb{E}[g_t^-]\rangle \\ &+ (\frac{1}{c} - c)\langle \mathbb{E}[g_t^-], \mathbb{E}[g_t^+]\rangle.\end{aligned}$$

According to Cauchy-Schwarz Inequality, there is $|\langle \mathbb{E}[g_t^+], \mathbb{E}[g_t^-]\rangle| \leq ||\mathbb{E}[g_t^+]||_2||\mathbb{E}[g_t^-]||_2 \leq G^2$. Combining the proof above, we can get Equation 1.
To prove Equation 2, first notice:

$$\begin{aligned}\frac{1}{c}\langle \mathbb{E}[g_t^+], [\theta_t - \theta^*]^+\rangle &\leq \langle [\nabla f(\theta_t)]^+, [\theta_t - \theta^*]^+\rangle \\ &\leq c\langle \mathbb{E}[g_t^+], [\theta_t - \theta^*]^+\rangle, \\ \frac{1}{c}\langle \mathbb{E}[g_t^-], [\theta_t - \theta^*]^-\rangle &\leq \langle [\nabla f(\theta_t)]^-, [\theta_t - \theta^*]^-\rangle \\ &\leq c\langle \mathbb{E}[g_t^-], [\theta_t - \theta^*]^-\rangle, \\ c\langle \mathbb{E}[g_t^+], [\theta_t - \theta^*]^-\rangle &\leq \langle [\nabla f(\theta_t)]^+, [\theta_t - \theta^*]^-\rangle\end{aligned}$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

$$\leq \frac{1}{c}\langle \mathbb{E}[g_t^+], [\theta_t - \theta^*]^-\rangle,$$

$$c\langle \mathbb{E}[g_t^-], [\theta_t - \theta^*]^+\rangle \leq \langle [\nabla f(\theta_t)]^-, [\theta_t - \theta^*]^+\rangle$$

$$\leq \frac{1}{c}\langle \mathbb{E}[g_t^-], [\theta_t - \theta^*]^+\rangle,$$

where $[\theta_t - \theta^*]^+ = \max\{\theta_t - \theta^*, \mathbf{0}\}$ and $[\theta_t - \theta^*]^- = \min\{\theta_t - \theta^*, \mathbf{0}\}$.
Then we have:

$$\begin{aligned}
&\langle \nabla f(\theta_t), \theta_t - \theta^*\rangle \\
=&\langle [\nabla f(\theta_t)]^+ + [\nabla f(\theta_t)]^-, [\theta_t - \theta^*]^+ + [\theta_t - \theta^*]^-\rangle \\
=&\langle [\nabla f(\theta_t)]^+, [\theta_t - \theta^*]^+\rangle + \langle [\nabla f(\theta_t)]^+, [\theta_t - \theta^*]^-\rangle + \\
&\langle [\nabla f(\theta_t)]^-, [\theta_t - \theta^*]^+\rangle + \langle [\nabla f(\theta_t)]^-, [\theta_t - \theta^*]^-\rangle \\
\leq& c\langle \mathbb{E}[g_t^+], [\theta_t - \theta^*]^+\rangle + c\langle \mathbb{E}[g_t^-], [\theta_t - \theta^*]^-\rangle + \\
&\frac{1}{c}\langle \mathbb{E}[g_t^-], [\theta_t - \theta^*]^+\rangle + \frac{1}{c}\langle \mathbb{E}[g_t^+], [\theta_t - \theta^*]^-\rangle \\
=& c\langle \mathbb{E}[g_t^+], [\theta_t - \theta^*]^+\rangle + c\langle \mathbb{E}[g_t^-], [\theta_t - \theta^*]^-\rangle + \\
&c\langle \mathbb{E}[g_t^-], [\theta_t - \theta^*]^+\rangle + c\langle \mathbb{E}[g_t^+], [\theta_t - \theta^*]^-\rangle + \\
&(\frac{1}{c} - c)(\langle \mathbb{E}[g_t^-], [\theta_t - \theta^*]^+\rangle + \langle \mathbb{E}[g_t^+], [\theta_t - \theta^*]^-\rangle) \\
=& c\langle \mathbb{E}[g_t], [\theta_t - \theta^*]\rangle + \\
&(\frac{1}{c} - c)(\langle \mathbb{E}[g_t^-], [\theta_t - \theta^*]^+\rangle + \langle \mathbb{E}[g_t^+], [\theta_t - \theta^*]^-\rangle)
\end{aligned}$$

In addition, $\langle \mathbb{E}[g_t^-], [\theta_t - \theta^*]^+\rangle$ and $\langle \mathbb{E}[g_t^+], [\theta_t - \theta^*]^-\rangle$ could be bounded by Cauchy-Schwarz Inequality:

$$\begin{aligned}
|\langle \mathbb{E}[g_t^+], [\theta_t - \theta^*]^-\rangle| &\leq ||\mathbb{E}[g_t^+]||_2||[\theta_t - \theta^*]^-||_2 \\
&= ||\mathbb{E}[g_t^+]||_2|| \min\{\theta_t - \theta^*, \mathbf{0}\}||_2 \\
&\leq ||\mathbb{E}[g_t^+]||_2||\theta_t - \theta^*||_2 \\
&\leq GR \\
|\langle \mathbb{E}[g_t^-], [\theta_t - \theta^*]^+\rangle| &\leq ||\mathbb{E}[g_t^-]||_2||[\theta_t - \theta^*]^+||_2 \\
&= ||\mathbb{E}[g_t^-]||_2|| \max\{\theta_t - \theta^*, \mathbf{0}\}||_2 \\
&\leq ||\mathbb{E}[g_t^-]||_2||\theta_t - \theta^*||_2 \\
&\leq GR
\end{aligned}$$

Therefore Equation 2 can be proved, and this completes the proof. □

Lemma 1 gives the new bounds of two terms assuming the constant bound on the gradient, which are essential to the proof of convergence rate. Based on Lemma 1, we can prove Theorem 4, which bounds the error of Stochastic Gradient Descent (SGD) on a convex optimization problem when the estimated gradient g_t in the t-th step resides in a constant bound of $\nabla f(\theta_t)$.

Proof. (Theorem 4) By L-smooth of f, for the t-th iteration,

$$f(\theta_{t+1}) \leq f(\theta_t) + \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{L}{2}||\theta_{t+1} - \theta_t||_2^2,$$

$$= f(\theta_t) - \eta\langle \nabla f(\theta_t), g_t\rangle + \frac{L\eta^2}{2}||g_t||^2.$$

Because of the constant bound on gradient and $||\mathbb{E}[g_t]||_2^2 = \mathbb{E}[||g_t||_2^2] - Var(g_t)$, by taking expectation on both sides w.r.t g_t we get from Lemma 1 that

$$\begin{aligned}
&\mathbb{E}[f(\theta_{t+1})] \leq f(\theta_t) - \eta\langle \nabla f(\theta_t), \mathbb{E}[g_t]\rangle + \frac{L\eta^2}{2}\mathbb{E}[||g_t||_2^2] \\
&\leq f(\theta_t) - \eta\left(\frac{1}{c}||\mathbb{E}[g_t]||_2^2 - 2(c - \frac{1}{c})G^2\right) + \frac{L\eta^2}{2}\mathbb{E}[||g_t||_2^2] \\
&= f(\theta_t) - \eta\left(\frac{1}{c}\left(\mathbb{E}[||g_t||_2^2] - Var(g_t)\right) - 2(c - \frac{1}{c})G^2\right) + \\
&\quad\frac{L\eta^2}{2}\mathbb{E}[||g_t||_2^2] \\
&\leq f(\theta_t) - \frac{\eta(2 - L\eta c)}{2c}\mathbb{E}[||g_t||_2^2] + \frac{\eta}{c}\sigma^2 + 2\eta(c - \frac{1}{c})G^2 \\
&\leq f(\theta_t) - \frac{\eta c}{2}\mathbb{E}[||g_t||_2^2] + \frac{\eta}{c}\sigma^2 + 2\eta(c - \frac{1}{c})G^2
\end{aligned}$$

where the last inequality follows as $L\eta c \leq 2 - c^2$. Because f is convex, still from Lemma 1 we get

$$\begin{aligned}
&\mathbb{E}[f(\theta_{t+1})] \\
&\leq f(\theta^*) + \langle \nabla f(\theta_t), \theta_t - \theta^*\rangle - \frac{\eta c}{2}\mathbb{E}[||g_t||_2^2] + \\
&\quad\frac{\eta}{c}\sigma^2 + 2\eta(c - \frac{1}{c})G^2, \\
&\leq f(\theta^*) + c\langle \mathbb{E}[g_t], \theta_t - \theta^*\rangle + 2(c - \frac{1}{c})GR - \frac{\eta c}{2}\mathbb{E}[||g_t||_2^2] + \\
&\quad\frac{\eta}{c}\sigma^2 + 2\eta(c - \frac{1}{c})G^2, \\
&= f(\theta^*) + c\mathbb{E}[\langle g_t, \theta_t - \theta^*\rangle - \frac{\eta}{2}||g_t||_2^2] + \frac{\eta}{c}\sigma^2 + \\
&\quad 2(c - \frac{1}{c})GR + 2\eta(c - \frac{1}{c})G^2.
\end{aligned}$$

Denote $\Lambda = \frac{\eta}{c}\sigma^2 + 2(c - \frac{1}{c})GR + 2\eta(c - \frac{1}{c})G^2$. We now repeat the calculations by completing the square for the middle two terms to get

$$\begin{aligned}
&\mathbb{E}[f(\theta_{t+1})] \\
&\leq f(\theta^*) + \frac{c}{2\eta}\mathbb{E}[2\eta\langle g_t, \theta_t - \theta^*\rangle - \eta^2||g_t||_2^2] + \Lambda, \\
&\leq f(\theta^*) + \frac{c}{2\eta}\mathbb{E}[||\theta_t - \theta^*||_2^2 - ||\theta_t - \theta^* - \eta g_t||_2^2] + \Lambda, \\
&= f(\theta^*) + \frac{c}{2\eta}\mathbb{E}[(||\theta_t - \theta^*||_2^2 - ||\theta_{t+1} - \theta^*||_2^2)] + \Lambda.
\end{aligned}$$

Summing the above equations for $t = 0, \ldots, T-1$, we get

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}[f(\theta_{t+1}) - f(\theta^*)] \\
&\leq \frac{c}{2\eta}(||\theta_0 - \theta^*||_2^2 - \mathbb{E}[||\theta_T - \theta^*||_2^2]) + T\Lambda \\
&\leq \frac{c||\theta_0 - \theta^*||_2^2}{2\eta} + T\Lambda.
\end{aligned}$$

Finally, by Jensen's inequality, $tf(\overline{\theta_T}) \leq \sum_{t=1}^{T} f(\theta_t)$,

$$\sum_{t=0}^{T-1} \mathbb{E}[f(\theta_{t+1}) - f(\theta^*)] = \mathbb{E}[\sum_{t=1}^{T} f(\theta_t)] - Tf(\theta^*)$$
$$\geq T\mathbb{E}[f(\overline{\theta_T})] - Tf(\theta^*).$$

Combining the above equations we get

$$\mathbb{E}[f(\overline{\theta_T})] \leq f(\theta^*) + \frac{c||\theta_0 - \theta^*||_2^2}{2\eta T} + \frac{\eta}{c}\sigma^2 + 2(c - \frac{1}{c})GR + 2\eta(c - \frac{1}{c})G^2.$$

This completes the proof. □

2 PROOF OF THEOREM 3

To prove Theorem 3, we first introduce a Lemma as follows:

Lemma 2. *If the total variation* $\max_\theta Var_P(f(\tau)) \leq \sigma_2^2$*, then* $L(\theta)$ *is* σ_2^2*-smooth w.r.t.* θ*.*

Proof. Since $L(\theta) = \frac{1}{N}\sum_{\tau \in \mathcal{D}} \log P(\tau|\theta, T)$, σ_2^2-smoothness requires that

$$||\nabla L(\theta_1) - \nabla L(\theta_2)||_2 \leq \sigma_2^2 ||\theta_1 - \theta_2||_2$$

where L is a constant. Because of the mean value theorem, there exists a point $\tilde{\theta} \in (\theta_1, \theta_2)$ such that

$$\nabla L(\theta_1) - \nabla L(\theta_2) = \nabla(\nabla L(\tilde{\theta}))(\theta_1 - \theta_2).$$

Taking the L_2 norm for both sides, we have

$$\begin{aligned} ||\nabla L(\theta_1) - \nabla L(\theta_2)||_2 =& ||\nabla(\nabla L(\tilde{\theta}))(\theta_1 - \theta_2)||_2 \\ \leq& ||\nabla(\nabla L(\tilde{\theta}))||_2 \, ||\theta_1 - \theta_2||_2 \end{aligned} \quad (5)$$

Then, the problem is to bound the matrix 2-norm $||\nabla(\nabla L(\tilde{\theta}))||_2$. Since we know the explicit form of $L(\theta)$, we know

$$\nabla L(\theta) = \frac{1}{|\mathcal{D}|}\sum_{\tau \in \mathcal{D}} f(\tau) - \nabla \log Z_\theta,$$

$$\begin{aligned} &\nabla(\nabla L(\theta)) \\ =& -\sum_{\tau \in \mathcal{T}} [f(\tau) - \nabla \log Z_\theta][f(\tau) - \nabla \log Z_\theta]^T P(\tau|\theta, T), \end{aligned} \quad (6)$$

where $\nabla(\nabla L(\theta))$ is the co-variance matrix. Denote $\text{Cov}_\theta[f(\tau)] = -\nabla(\nabla L(\theta))$, which is both symmetric and positive semi-definite. We have

$$||\nabla(\nabla L(\tilde{\theta}))||_2 = ||\text{Cov}_\theta[f(\tau)]||_2 = \lambda_{max},$$

where λ_{max} is the maximum eigenvalue of the matrix $\text{Cov}_\theta[f(\tau)]$. Then, because of the positive semi-definiteness of the co-variance matrix, all the eigenvalues are non-negative, and we can bound λ_{max} as

$$\lambda_{max} \leq \sum_i \lambda_i = Tr(\text{Cov}_\theta[f(\tau)]),$$

where $Tr(\text{Cov}_\theta[\phi(X)])$ is the trace of matrix $\text{Cov}_\theta[f(\tau)]$. Using the definition in Equation 6, $Tr(\text{Cov}_\theta[f(\tau)])$ can be further derived as:

$$Tr(\text{Cov}_\theta[f(\tau)]) = \mathbb{E}_P[||f(\tau)||_2^2] - ||\mathbb{E}_P[f(\tau)]||_2^2,$$

which is equal to the total variation $Var_P(f(\tau))$. Therefore, we have

$$||\nabla(\nabla L(\tilde{\theta}))||_2 \leq Var_P(f(\tau)) \leq \sigma_2^2.$$

Combining this with Equation 5, we know

$$||\nabla L(\theta_1) - \nabla L(\theta_2)||_2 \leq \sigma_2^2 \, ||\theta_1 - \theta_2||_2.$$

This completes the proof.

□

We give the full proof of Theorem 3 as follows:

Proof. (Theorem 3) Since we use M_1 samples from the training set $\{\tau_i\}_{i=1}^{M_1}$ and M_2 samples $\tau'_1, \ldots, \tau'_{M_2}$ from $P(\tau|T, \theta)$ using XOR-Sampling at each iteration, we have

$$g_k = \frac{1}{M_1}\sum_{\tau \in \mathcal{D}_{M_1}} f(\tau) - \frac{1}{M_2}\sum_{j=1}^{M_2} f(\tau'_j)$$

Denote $g_k^i = \frac{1}{M_1}\sum_{j=1}^{M_1} f(\tau_j) - f(\tau'_i)$, we have the expectation of g_k as

$$\mathbb{E}_{\mathcal{D},P}[g_k] = \mathbb{E}_{\mathcal{D},P}[g_k^i].$$

In each iteration k we can adjust the parameters in XOR-Sampling to give the constant factor approximation of both the denominator and the nominator, then for each g_k^i we can obtain from Theorem 2 that

$$\frac{1}{\delta}[\nabla L(\theta_k)]^+ \leq \mathbb{E}_{\mathcal{D},P}[g_k^{i+}] \leq \delta[\nabla L(\theta_k)]^+, \quad (7)$$

$$\delta[\nabla L(\theta_k)]^- \leq \mathbb{E}_{\mathcal{D},P}[g_k^{i-}] \leq \frac{1}{\delta}[\nabla L(\theta_k)]^-. \quad (8)$$

where we denote

$$g_k^{i+} = \max\{g_k^i, \mathbf{0}\}, \quad g_k^{i-} = \min\{g_k^i, \mathbf{0}\},$$
$$[\nabla L(\theta_k)]^+ = \mathbb{E}_P[[\mathbb{E}_\mathcal{D}[f(\tau)] - f(\tau')]^+]$$
$$[\nabla L(\theta_k)]^- = \mathbb{E}_P[[\mathbb{E}_\mathcal{D}[f(\tau)] - f(\tau')]^-].$$

Notice that $g_k^+ = \frac{1}{M_2}\sum_{i=1}^{M_2} g_k^{i+}$ and $g_k^- = \frac{1}{M_2}\sum_{i=1}^{M_2} g_k^{i-}$. Combined with Equation 7 and 8, we know,

$$\frac{1}{\delta}[\nabla L(\theta_k)]^+ \le \mathbb{E}[g_k^+] \le \delta[\nabla L(\theta_k)]^+,$$
$$\delta[\nabla L(\theta_k)]^- \le \mathbb{E}[g_k^-] \le \frac{1}{\delta}[\nabla L(\theta_k)]^-.$$

As required in Theorem 3, $||\mathbb{E}[g_k^+]||_2$ and $||\mathbb{E}[g_k^-]||_2$ can be bounded by

$$\begin{aligned}\mathbb{E}[g_k^+] &= \mathbb{E}_P[[\mathbb{E}_{\mathcal{D}}[f(\tau)] - f(\tau')]^+] \\ &\le \delta\mathbb{E}_P[[\mathbb{E}_{\mathcal{D}}[f(\tau)] - f(\tau')]^+] \\ &\le \delta\mathbb{E}_P[[\mathbb{E}_{\mathcal{D}}[f(\tau)]]^+ + [-f(\tau')]^+] \\ &= \delta\{[\mathbb{E}_{\mathcal{D}}[f(\tau)]]^+ - \mathbb{E}_P[[f(\tau')]^-]\}.\end{aligned}$$

$$\begin{aligned}\mathbb{E}[g_k^-] &= \mathbb{E}_P[[\mathbb{E}_{\mathcal{D}}[f(\tau)] - f(\tau')]^-] \\ &\ge \delta\mathbb{E}_P[[\mathbb{E}_{\mathcal{D}}[f(\tau)] - f(\tau')]^-] \\ &\ge \delta\mathbb{E}_P[[\mathbb{E}_{\mathcal{D}}[f(\tau)]]^- + [-f(\tau')]^-] \\ &= \delta\{[\mathbb{E}_{\mathcal{D}}[f(\tau)]]^- - \mathbb{E}_P[[f(\tau')]^+]\}.\end{aligned}$$

Therefore, we have $||\mathbb{E}[g_k^+]||_2^2 \le \delta^2(G+E)^2$ and $||\mathbb{E}[g_k^-]||_2^2 \le \delta^2(G+E)^2$.

In terms of variance, because $\mathbb{E}_{\mathcal{D},P}[g_k] = \mathbb{E}_{\mathcal{D},P}[g_k^i]$, the variance of g_k, denoted as $Var_{\mathcal{D},P}(g_k)$, can then be bounded as

$$\begin{aligned}&Var_{\mathcal{D},P}(g_k) \\ =&Var_D\left(\frac{1}{M_1}\sum_{j=1}^{M_1} f(\tau_j)\right) + Var_P\left(\frac{1}{M_2}\sum_{i=1}^{M_2} f(\tau_i')\right) \\ =&\frac{1}{M_1}Var_D(f(\tau_j)) + \frac{1}{M_2}Var_P(f(\tau_i')) \\ \le&\frac{\sigma_1^2}{M_1} + \frac{\sigma_2^2}{M_2}.\end{aligned}$$

The last inequality is because $Var_{\mathcal{D}}(f(\tau)) \le \sigma_1^2$ and $\max_\theta Var_P(f(\tau_j')) \le \sigma_2^2$.

Since $L(\theta)$ is convex and σ_2^2-smooth from Lemma 2, according to Theorem 4, when the learning rate η is bounded by:

$$\eta \le \frac{2-\delta^2}{\sigma^2\delta}, \tag{9}$$

we can then apply Theorem 4 to get the result in Theorem 3:

$$\begin{aligned}&\mathbb{E}[L(\overline{\theta_K})] - OPT \\ \le& \frac{\delta||\theta_0 - \theta^*||_2^2}{2\eta K} + \frac{\eta\sigma_1^2}{\delta M_1} + \frac{\eta\sigma_2^2}{\delta M_2} + \\ &\quad 2(\delta^2-1)(G+E)R + 2\eta(\delta^3-\delta)(G+E)^2.\end{aligned}$$

This completes the proof. □

3 PROOF OF THEOREM 5

Proof. (Theorem 5) Since we use flow constraints to ensure valid trajectories, the number of binary variables in XOR-Sampling in $O(|\mathcal{S}||\mathcal{A}|)$. From Theorem 2 we know that in each iteration of X-MEN, we need to access $O(-|\mathcal{S}||\mathcal{A}|\log(1-1/\sqrt{\delta})\log(-|\mathcal{S}||\mathcal{A}|/\gamma\log(1-1/\sqrt{\delta})))$ queries of NP oracles in order to generate one sample. However, as specified also in Ermon et al [2013b], only the first sample needs those many queries. Once we have the first sample, the number of XOR constraints to add can be known in generating future samples for this SGD iteration. Therefore, we fix the number of XOR constraints added starting the generation of the second sample. As a result, we only need one NP oracle query in generating each of the following (M_2-1) samples. Therefore, total queries in each iteration will be $O(-|\mathcal{S}||\mathcal{A}|\log(1-1/\sqrt{\delta})\log(-|\mathcal{S}||\mathcal{A}|/\gamma\log(1-1/\sqrt{\delta}))+M_2)$. To complete all K SGD iterations, X-MEN needs $O(-K|\mathcal{S}||\mathcal{A}|\log(1-1/\sqrt{\delta})\log(-|\mathcal{S}||\mathcal{A}|/\gamma\log(1-1/\sqrt{\delta}))+KM_2)$ NP oracle queries in total. □