A PROOF OF LEMMA 2

Proof. From the boundedness of the features (by L) and the rewards (by $\mathrm{R}_{\max}$), we have

$$\begin{aligned}
||A||_2 &= ||\mathbb{E}[\rho_t \phi_t \Delta\phi_t^\top]||_2 \\
&\leq \max_s ||\rho(s)\phi(s)(\Delta\phi(s))^\top||_2 \\
&\leq \rho_{\max} \max_s ||\phi(s)||_2 \max_s ||\phi(s) - \gamma\phi'(s)||_2 \\
&\leq \rho_{\max} \max_s ||\phi(s)||_2 \max_s \left(||\phi(s)||_2 + \gamma||\phi'(s)||_2\right) \\
&\leq (1+\gamma)\rho_{\max} L^2 d.
\end{aligned}$$

The second inequality is obtained by the consistent inequality of matrix norm, the third inequality comes from the triangular norm inequality, and the fourth inequality comes from the vector norm inequality $||\phi(s)||_2 \leq ||\phi(s)||_\infty \sqrt{d} \leq L\sqrt{d}$. The bound on $||b||_2$ can be derived in a similar way as follows.

$$\begin{aligned}
||b||_2 &= ||\mathbb{E}[\rho_t \phi_t r_t]||_2 \\
&\leq \max_s ||\rho(s)\phi(s)r(s)||_2 \\
&\leq \rho_{\max} \max_s ||\phi(s)||_2 \max_s ||r(s)||_2 \\
&\leq \rho_{\max} L \mathrm{R}_{\max}.
\end{aligned}$$

It completes the proof. □

B PROOF OF PROPOSITION 3

Proof. The proof of Proposition 3 mainly relies on Proposition 3.2 in Nemirovski *et al.* [2009]. We just need to map our convex-concave *stochastic* saddle-point problem in Eq. 14, i.e.,

$$\min_{\theta\in\Theta} \max_{y\in Y} \left(L(\theta, y) = \langle b - A\theta, y\rangle - \frac{1}{2}||y||_M^2 \right)$$

to the one in Section 3 of Nemirovski *et al.* [2009] and show that it satisfies all the conditions necessary for their Proposition 3.2. Assumption 2 guarantees that our feasible sets Θ and Y satisfy the conditions in Nemirovski *et al.* [2009], as they are non-empty bounded closed convex subsets of $\mathbb{R}^d$. We also see that our objective function $L(\theta, y)$ is *convex* in $\theta \in \Theta$ and *concave* in $y \in Y$, and also *Lipschitz continuous* on $\Theta \times Y$. It is known that in the above setting, our saddle-point problem in Eq. 14 is solvable, i.e., the corresponding *primal* and *dual* optimization problems: $\min_{\theta\in\Theta}\left[\max_{y\in Y} L(\theta, y)\right]$ and $\max_{y\in Y}\left[\min_{\theta\in\Theta} L(\theta, y)\right]$ are solvable with equal optimal values, denoted L^*, and pairs (θ^*, y^*) of optimal solutions to the respective problems from the set of saddle points of $L(\theta, y)$ on $\Theta \times Y$.

For our problem, the *stochastic sub-gradient vector* G is defined as

$$G(\theta, y) = \begin{bmatrix} G_\theta(\theta, y) \\ -G_y(\theta, y) \end{bmatrix} = \begin{bmatrix} -\hat{A}_t^\top y \\ -(\hat{b}_t - \hat{A}_t\theta - \hat{M}_t y) \end{bmatrix}.$$

This guarantees that the *deterministic sub-gradient vector*

$$g(\theta, y) = \begin{bmatrix} g_\theta(\theta, y) \\ -g_y(\theta, y) \end{bmatrix} = \begin{bmatrix} \mathbb{E}[G_\theta(\theta, y)] \\ -\mathbb{E}[G_y(\theta, y)] \end{bmatrix}$$

is well-defined, i.e., $g_\theta(\theta, y) \in \partial_\theta L(\theta, y)$ and $g_y(\theta, y) \in \partial_y L(\theta, y)$.

We also consider the Euclidean stochastic approximation (E-SA) setting in Nemirovski *et al.* [2009] in which the *distance generating functions* $\omega_\theta : \Theta \to \mathbb{R}$ and $\omega_y : Y \to \mathbb{R}$ are simply defined as

$$\omega_\theta = \frac{1}{2}||\theta||_2^2, \qquad \omega_y = \frac{1}{2}||y||_2^2,$$

modulus 1 w.r.t. $||\cdot||_2$, and thus, $\Theta^o = \Theta$ and $Y^o = Y$ (see pp. 1581 and 1582 in Nemirovski *et al.* 2009). This allows us to equip the set $Z = \Theta \times Y$ with the distance generating function

$$\omega(z) = \frac{\omega_\theta(\theta)}{2D_\theta^2} + \frac{\omega_y(y)}{2D_y^2},$$

where D_θ and D_y defined in Assumption 2.

Now that we consider the Euclidean case and set the norms to ℓ_2-norm, we can compute upper-bounds on the expectation of the dual norm of the stochastic sub-gradients

$$\mathbb{E}\left[||G_\theta(\theta, y)||_{*,\theta}^2\right] \leq M_{*,\theta}^2, \quad \mathbb{E}\left[||G_y(\theta, y)||_{*,y}^2\right] \leq M_{*,y}^2,$$

where $||\cdot||_{*,\theta}$ and $||\cdot||_{*,y}$ are the dual norms in Θ and Y, respectively. Since we are in the Euclidean setting and use the ℓ_2-norm, the dual norms are also ℓ_2-norm, and thus, to compute $M_{*,\theta}$, we need to upper-bound $\mathbb{E}\left[||G_\theta(\theta, y)||_2^2\right]$ and $\mathbb{E}\left[||G_y(\theta, y)||_2^2\right]$.

To bound these two quantities, we use the following equality that holds for any random variable x:

$$\mathbb{E}[||x||_2^2] = \mathbb{E}[||x - \mu_x||_2^2] + ||\mu_x||_2^2,$$

where $\mu_x = \mathbb{E}[x]$. Here how we bound $\mathbb{E}\left[||G_\theta(\theta, y)||_2^2\right]$,

$$\begin{aligned}
\mathbb{E}\left[||G_\theta(\theta, y)||_2^2\right] &= \mathbb{E}[||\hat{A}_t^\top y||_2^2] \\
&= \mathbb{E}[||\hat{A}_t^\top y - A^\top y||_2^2] + ||A^\top y||_2^2 \\
&\leq \sigma_2^2 + \left(||A||_2 ||y||_2\right)^2 \\
&\leq \sigma_2^2 + ||A||_2^2 R^2,
\end{aligned}$$

where the first inequality is from the definition of σ_3 in Eq. 20 and the consistent inequality of the matrix norm, and the second inequality comes from the boundedness of the feasible sets in Assumption 2. Similarly we bound $\mathbb{E}\left[||G_y(\theta, y)||_2^2\right]$ as follows:

$$\begin{aligned}
\mathbb{E}[||G_y(\theta, y)||_2^2] &= \mathbb{E}[||\hat{b}_t - \hat{A}_t\theta - \hat{M}_t y||_2^2] \\
&= ||b - A\theta + My||_2^2 \\
&+ \mathbb{E}[||\hat{b}_t - \hat{A}_t\theta - \hat{M}_t y - (b - A\theta - My)||_2^2] \\
&\leq \left(||b||_2 + ||A||_2||\theta||_2 + \tau||y||_2\right)^2 + \sigma_1^2 \\
&\leq \left(||b||_2 + (||A||_2 + \tau)R\right)^2 + \sigma_1^2,
\end{aligned}$$

where these inequalities come from the definition of σ_1 in Eq. 20 and the boundedness of the feasible sets in Assumption 2. This means that in our case we can compute $M^2_{*,\theta}, M^2_{*,y}$ as

$$M^2_{*,\theta} = \sigma_2^2 + ||A||_2^2 R^2,$$
$$M^2_{*,y} = \big(||b||_2 + (||A||_2 + \tau)R\big)^2 + \sigma_1^2,$$

and as a result

$$\begin{aligned} M^2_* &= 2D^2_\theta M^2_{*,\theta} + 2D^2_y M^2_{*,y} = 2R^2(M^2_{*,\theta} + M^2_{*,y}) \\ &= R^2\left(\sigma^2 + ||A||_2^2 R^2 + \big(||b||_2 + (||A||_2 + \tau)R\big)^2\right) \\ &\leq \big(R^2\,(2||A||_2 + \tau) + R(\sigma + ||b||_2)\big)^2, \end{aligned}$$

where the inequality comes from the fact that $\forall a, b, c \geq 0, a^2 + b^2 + c^2 \leq (a + b + c)^2$. Thus, we may write M_* as

$$M_* = R^2\,(2||A||_2 + \tau) + R(\sigma + ||b||_2). \tag{39}$$

Now we have all the pieces ready to apply Proposition 3.2 in Nemirovski *et al.* [2009] and obtain a high-probability bound on $\mathrm{Err}(\bar{\theta}_n, \bar{y}_n)$, where $\bar{\theta}_n$ and $\bar{y}_n$ (see Eq. 18) are the outputs of the revised GTD algorithm in Algorithm 1. From Proposition 3.2 in Nemirovski *et al.* [2009], if we set the step-size in Algorithm 1 (our revised GTD algorithm) to $\alpha_t = \frac{2c}{M_*\sqrt{5n}}$, where $c > 0$ is a positive constant, M_* is defined by Eq. 39, and n is the number of training samples in $\mathcal{D}$, with probability of at least $1 - \delta$, we have

$$\mathrm{Err}(\bar{\theta}_n, \bar{y}_n) \leq \sqrt{\frac{5}{n}}(8 + 2\log\frac{2}{\delta})R^2\left(2||A||_2 + \tau + \frac{||b||_2 + \sigma}{R}\right). \tag{40}$$

Note that we obtain Eq. 40 by setting $c = 1$ and the "light-tail" assumption in Eq. 22 guarantees that we satisfy the condition in Eq. 3.16 in Nemirovski *et al.* [2009], which is necessary for the high-probability bound in their Proposition 3.2 to hold. The proof is complete by replacing $||A||_2$ and $||b||_2$ from Lemma 2. □

C PROOF OF PROPOSITION 4

Proof. From Lemma 3, we have

$$\begin{aligned} V - \bar{v}_n = (I - \gamma\Pi P)^{-1} \times \\ \big[(V - \Pi V) + \Phi C^{-1}(b - A\bar{\theta}_n)\big]. \end{aligned}$$

Applying ℓ_2-norm w.r.t. the distribution ξ to both sides of this equation, we obtain

$$\begin{aligned} ||V - \bar{v}_n||_\xi \leq & ||(I - \gamma\Pi P)^{-1}||_\xi \times \\ & \big(||V - \Pi V||_\xi + ||\Phi C^{-1}(b - A\bar{\theta}_n)||_\xi\big). \end{aligned} \tag{41}$$

Since P is the kernel matrix of the target policy π and Π is the orthogonal projection w.r.t. ξ, the stationary distribution of π, we may write

$$||(I - \gamma\Pi P)^{-1}||_\xi \leq \frac{1}{1 - \gamma}.$$

Moreover, we may upper-bound the term $||\Phi C^{-1}(b - A\bar{\theta}_n)||_\xi$ in (41) using the following inequalities:

$$\begin{aligned} &||\Phi C^{-1}(b - A\bar{\theta}_n)||_\xi \\ &\quad \leq ||\Phi C^{-1}(b - A\bar{\theta}_n)||_2\sqrt{\xi_{\max}} \\ &\quad \leq ||\Phi||_2||C^{-1}||_2||(b - A\bar{\theta}_n)||_{M^{-1}}\sqrt{\tau\xi_{\max}} \\ &\quad \leq (L\sqrt{d})(\frac{1}{\nu})\sqrt{2\mathrm{Err}(\bar{\theta}_n, \bar{y}_n)}\sqrt{\tau\xi_{\max}} \\ &\quad = \frac{L}{\nu}\sqrt{2d\tau\xi_{\max}\mathrm{Err}(\bar{\theta}_n, \bar{y}_n)}, \end{aligned}$$

where the third inequality is the result of upper-bounding $||(b - A\bar{\theta}_n)||^{-1}_M$ using Eq. 28 and the fact that $\nu = 1/||C^{-1}||_2^2 = 1/\lambda_{\max}(C^{-1}) = \lambda_{\min}(C)$ (ν is the smallest eigenvalue of the covariance matrix C). □

D PROOF OF PROPOSITION 5

Proof. Using the triangle inequality, we may write

$$||V - \bar{v}_n||_\xi \leq ||\bar{v}_n - \Phi\theta^*||_\xi + ||V - \Phi\theta^*||_\xi. \tag{42}$$

The second term on the right-hand side of Eq. 42 can be upper-bounded by Lemma 4. Now we upper-bound the first term as follows:

$$\begin{aligned} &||\bar{v}_n - \Phi\theta^*||_\xi^2 \\ &\quad = ||\Phi\bar{\theta}_n - \Phi\theta^*||_\xi^2 \\ &\quad = ||\bar{\theta}_n - \theta^*||_C^2 \\ &\quad \leq ||\bar{\theta}_n - \theta^*||^2_{A^\top M^{-1}A}||(A^\top M^{-1}A)^{-1}||_2||C||_2 \\ &\quad = ||A(\bar{\theta}_n - \theta^*)||^2_{M^{-1}}||(A^\top M^{-1}A)^{-1}||_2||C||_2 \\ &\quad = ||A\bar{\theta}_n - b||^2_{M^{-1}}\frac{\tau_C}{\sigma_{\min}(A^\top M^{-1}A)}, \end{aligned}$$

where $\tau_C = \sigma_{\max}(C)$ is the largest singular value of C, and $\sigma_{\min}(A^\top M^{-1}A)$ is the smallest singular value of $A^\top M^{-1}A$. Using the result of Theorem 1, with probability at least $1 - \delta$, we have

$$\frac{1}{2}||A\bar{\theta}_n - b||^2_{M^{-1}} \leq \tau\xi_{\max}\mathrm{Err}(\bar{\theta}_n, \bar{y}_n). \tag{43}$$

Thus,

$$||\bar{v}_n - \Phi\theta^*||_\xi^2 \leq \frac{2\tau_C\tau\xi_{\max}}{\sigma_{\min}(A^\top M^{-1}A)}\mathrm{Err}(\bar{\theta}_n, \bar{y}_n) \tag{44}$$

From Eqs. 42, 32, and 44, the result of Eq. 33 can be derived, which completes the proof. □