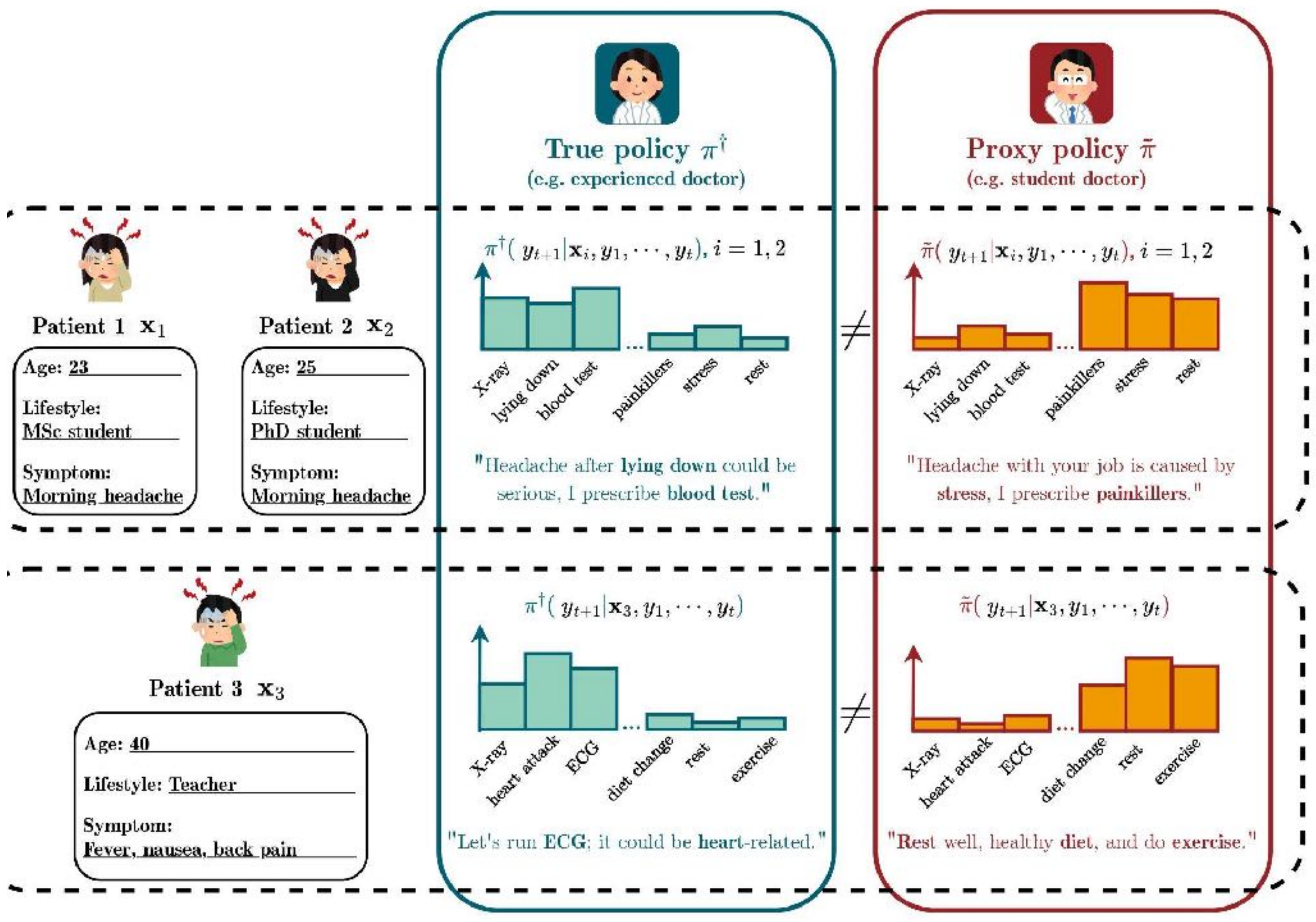
When Can Proxies Improve the Sample Complexity of Preference Learning?

Yuchen Zhu, Daniel Augusto de Souza, Zhengyan Shi, Mengyue Yang, Pasquale Minervini, Matt J. Kusner, Alexander D'Amour

ICML International Conference On Machine Learning 2025

Sufficient Conditions for Low-Rank/Dimensional Adaptation

I. Motivating Example



II. Mathematical set-up

$$\mathbf{x} \sim p_{\mathcal{X}}; \quad \mathbf{y}_1, \mathbf{y}_2 \underset{\text{i.i.d.}}{\sim} \pi_{\text{ref}}(\cdot \mid \mathbf{x}); \qquad (\mathbf{y}_w, \mathbf{y}_l) = (\mathbf{y}_1, \mathbf{y}_2) \quad \text{if } b = 1$$

$$b \sim \text{Bern}\left(\sigma\left(r(\mathbf{x}, \mathbf{y}_1) - r(\mathbf{x}, \mathbf{y}_2)\right)\right); \quad (\mathbf{y}_w, \mathbf{y}_l) = (\mathbf{y}_2, \mathbf{y}_1) \quad \text{if } b = 0$$

$$r(\mathbf{x}, \mathbf{y}) = \beta \log \frac{\pi(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})}.$$

$$\arg\max_{\pi} \mathbb{E}_{(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l) \sim G} \left[\log \sigma \left(\beta \log \frac{\pi(\mathbf{y}_w \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_w \mid \mathbf{x})} - \beta \log \frac{\pi(\mathbf{y}_l \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_l \mid \mathbf{x})}\right)\right]$$

III. Sufficient Conditions

Condition 1. For a given metric $d_{\mathcal{P}_\mathcal{Y}}$ on the space of distributions on $\mathcal{Y}$, there is some positive scalar L such that

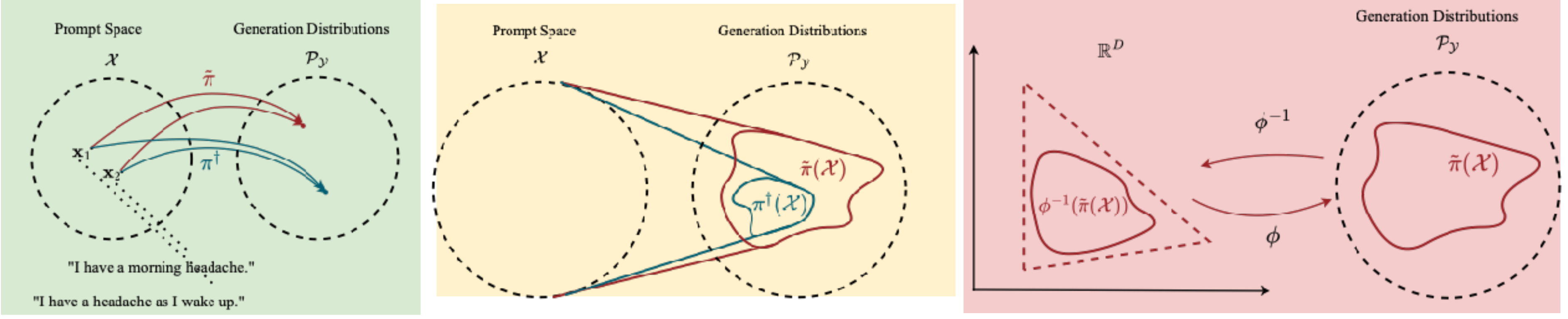
$$d_{\mathcal{P}_\mathcal{Y}}(\pi^\dagger(\cdot|x_1), \pi^\dagger(\cdot|x_2)) \leq L d_{\mathcal{P}_\mathcal{Y}}(\tilde{\pi}(\cdot|x_1), \tilde{\pi}(\cdot|x_2)) \tag{1}$$

Condition 2 (Image Inclusion).

$$\pi^\dagger(\mathcal{X}) \subseteq \tilde{\pi}(\mathcal{X}) \tag{2}$$

Condition 3 (Low-dimensional encoding). There exists an injective function $\phi\colon \mathcal{V} \to \mathcal{P}_\mathcal{Y}$, where $\mathcal{V} \subset \mathbb{R}^D$ is a bounded convex polytope with $D+1$ vertices, such that:

1. The image $\phi(\mathcal{V})$ contains the image of the policies: $\pi^\dagger(\mathcal{X}) \subseteq \tilde{\pi}(\mathcal{X}) \subseteq \phi(\mathcal{V})$;
2. It is $(L_\phi, L_{\phi^{-1}})$-bi-Lipschitz with its left inverse $\phi^{-1}\colon \mathcal{P}_\mathcal{Y} \to \mathcal{V}$: $\frac{1}{L_{\phi^{-1}}}\|v_1 - v_2\|_p \leq d(\phi(v_1), \phi(v_2)) \leq L_\phi \|v_1 - v_2\|_p$, where d is a metric on $\mathcal{P}_\mathcal{Y}$.



IV. Low-dimensional Adaptation

Theorem. We work under Conditions 1–3. For some D, there exists a Lipschitz invertible function $\tilde{\phi} : \mathcal{V} \to \tilde{\pi}(\mathcal{X})$ satisfying Condition 4, $\tilde{\Theta} \in \mathbb{R}^{N \times (D+1)}$ and $\tilde{\tau}^\circ : \mathcal{X} \to \Delta^D$ such that $\tilde{\pi} = \tilde{\phi} \circ \tilde{\Theta}\tilde{\tau}^\circ$. Moreover, for any $(\tilde{\phi}, \tilde{\Theta}, \tilde{\tau}^\circ)$ such that $\tilde{\pi} = \tilde{\phi} \circ \tilde{\Theta}\tilde{\tau}^\circ$, there exists a Lipschitz continuous function $\bar{\pi}^\dagger : \Delta^D \to \Delta^D$ such that $\pi^\dagger = \tilde{\phi} \circ \tilde{\Theta} \circ \bar{\pi}^\dagger \circ \tilde{\tau}^\circ$.

$$\tilde{\pi} = \tilde{\phi} \circ \tilde{\Theta} \circ \tilde{\tau}^\circ \qquad \pi^\dagger = \tilde{\phi} \circ \tilde{\Theta} \circ \bar{\pi}^\dagger \circ \tilde{\tau}^\circ$$

V. Improved Sample Complexity

Theorem. Assuming we have trained to convergence on proxy data, $\cdots$

we need

$$n(\epsilon, \omega) = \Omega\left(\frac{D}{\epsilon^2}\left(\frac{96 L_\phi \|\tilde{\Theta}\|_p L_{\tilde{\pi}} \sqrt{D}}{\epsilon}\right)^D \log\left(\frac{96 L_\phi \|\tilde{\Theta}\|_p \sqrt{D}}{\epsilon}\right) - \log \omega\right)$$

samples to generalise.

Theorem. Otherwise ...

We need

$$n(\epsilon, \omega) = \Omega\left(\frac{D'}{\epsilon^2}\left(\frac{48 L_\phi \|\tilde{\Theta}\|_p L_{\tilde{\pi}} E'(p, D')\sqrt{D'}}{\epsilon}\right)^{D'} \log\left(\frac{48 L_\phi \|\tilde{\Theta}\|_p L_{\tilde{\pi}} E'(p, D')\sqrt{D'}}{\epsilon}\right) - \log \omega\right)$$

VI. Real world application

Tempered softmax gives a real-world example which satisfies our conditions.

- $\mathcal{X} = \mathbb{R}^5$, $\mathcal{P}_\mathcal{X} = \mathcal{N}(\mathbf{0}, I_5)$
- $\mathcal{Y} = \{1, 2, 3\}$, so $\mathcal{P}_\mathcal{Y} = \Delta^2$, a two-simplex.
- $D = 1$. Condition 3 is satisfied.
- $\pi^\dagger : \mathbb{R}^5 \to \mathbb{R} \to \Delta^2$.
- $\log(\tilde{\pi}(y_k \mid x)) = \frac{\log(\pi^\dagger(y_k|x))}{T}$. Temperature $T = 5$.
- $\pi_{\text{ref}} = \text{Uniform}\{1, 2, 3\}$

Division by a constant T satisfies conditions 1 and 2 where $d_{\mathcal{P}_\mathcal{Y}}$ is the log difference.

	π_{ref}	$\pi^\dagger$	$\tilde{\pi}$	$\tilde{\pi}_\theta$	$\pi^\dagger_\theta$
mean	0.63	0.0	0.33	0.34	**0.32**
std	0.00	0.0	0.00	0.014	0.096

Table 1: Results for the Tempered Reward experiment

UCL · Microsoft Research · University of Bristol · The University of Edinburgh · Polytechnique Montréal · Google DeepMind