Rob Wass and Clinton Golding. Sharpening a tool for teaching: the zone of proximal development. *Teaching in Higher Education*, 19:671 – 684, 2014. 2

Yilei Zeng, Jiali Duan, Yang Li, Emilio Ferrara, Lerrel Pinto, C. C. Jay Kuo, and Stefanos Nikolaidis. Human decision makings on curriculum reinforcement learning with difficulty adjustment, 2022. URL https://arxiv.org/abs/2208.02932. 1

Yunzhi Zhang, Pieter Abbeel, and Lerrel Pinto. Automatic curriculum learning through value disagreement. *Advances in Neural Information Processing Systems*, 33:7648–7659, 2020. 2

A LOWER BOUNDING THE TEACHER OBJECTIVE

We are interested in showing that the lower bound used in Section 3.2 holds. As a reminder, the lower bound is:

$$\max_{\chi_\phi(C^{\text{curr}})} \mathbb{E}_{c^{\text{curr}}\sim\chi_\phi(\cdot)}\left[\sum_{r^{\text{curr}}\in R}\int_{\mathbb{R}^d} p(R^{\text{test}}=1|C^{\text{test}},\theta^{\text{test}})p(\theta^{\text{test}}|r^{\text{curr}},c^{\text{curr}},\theta^{\text{curr}})p(r^{\text{curr}}|c^{\text{curr}},\theta^{\text{curr}})d\theta^{\text{test}}\right]$$
$$\geq \max_{\chi_\phi(C^{\text{curr}})} \mathbb{E}_{c^{\text{curr}}\sim\chi_\phi(\cdot)}\left[\sum_{r^{\text{curr}}\in R}\int_{\theta^*\in\Theta^*} p(R^{\text{test}}=1|C^{\text{test}},\theta^*)p(\theta^*|r^{\text{curr}},c^{\text{curr}},\theta^{\text{curr}})p(r^{\text{curr}}|c^{\text{curr}},\theta^{\text{curr}})d\theta^*\right]$$

To see how this is true, we do the following:

$$\max_{\chi_\phi(C^{\text{curr}})} \mathbb{E}_{c^{\text{curr}}\sim\chi_\phi(\cdot)}\left[\sum_{r^{\text{curr}}\in R}\int_{\mathbb{R}^d} p(R^{\text{test}}=1|C^{\text{test}},\theta^{\text{test}})p(\theta^{\text{test}}|r^{\text{curr}},c^{\text{curr}},\theta^{\text{curr}})p(r^{\text{curr}}|c^{\text{curr}},\theta^{\text{curr}})d\theta^{\text{test}}\right] \tag{6}$$
$$= \max_{\chi_\phi(C^{\text{curr}})} \mathbb{E}_{c^{\text{curr}}\sim\chi_\phi(\cdot)}\left[\sum_{r^{\text{curr}}\in R}\int_{\mathbb{R}^d} p(R^{\text{test}}=1|C^{\text{test}},\theta^{\text{test}})p(\theta^{\text{test}}|r^{\text{curr}},c^{\text{curr}},\theta^{\text{curr}})\right. \tag{7}$$
$$\left.\mathbb{1}\left[\theta^{\text{test}}\in \operatorname{supp} p(\theta^{\text{test}}|r^{\text{curr}},c^{\text{curr}},\theta^{\text{curr}})\right]p(r^{\text{curr}}|c^{\text{curr}},\theta^{\text{curr}})d\theta^{\text{test}}\right] \tag{8}$$
$$= \max_{\chi_\phi(C^{\text{curr}})} \mathbb{E}_{c^{\text{curr}}\sim\chi_\phi(\cdot)}\left[\sum_{r^{\text{curr}}\in R}\int_{\mathbb{R}^d} p(R^{\text{test}}=1|C^{\text{test}},\theta^{\text{test}})p(\theta^{\text{test}}|r^{\text{curr}},c^{\text{curr}},\theta^{\text{curr}})\right. \tag{9}$$
$$\left(\mathbb{1}\left[\theta^{\text{test}}\in \max_{\theta \text{ s.t. } p(\theta^{\text{test}}|r^{\text{curr}},c^{\text{curr}},\theta^{\text{curr}})>0} p(R^{\text{test}}=1|C^{\text{test}},\theta^{\text{test}})\right]\right. \tag{10}$$
$$\left.\left.+\mathbb{1}\left[\theta^{\text{test}}\notin \max_{\theta \text{ s.t. } p(\theta^{\text{test}}|r^{\text{curr}},c^{\text{curr}},\theta^{\text{curr}})>0} p(R^{\text{test}}=1|C^{\text{test}},\theta^{\text{test}})\right]\right)p(r^{\text{curr}}|c^{\text{curr}},\theta^{\text{curr}})d\theta^{\text{test}}\right] \tag{11}$$
$$\geq \max_{\chi_\phi(C^{\text{curr}})} \mathbb{E}_{c^{\text{curr}}\sim\chi_\phi(\cdot)}\left[\sum_{r^{\text{curr}}\in R}\int_{\mathbb{R}^d} p(R^{\text{test}}=1|C^{\text{test}},\theta^{\text{test}})p(\theta^{\text{test}}|r^{\text{curr}},c^{\text{curr}},\theta^{\text{curr}})\right. \tag{12}$$
$$\left.\mathbb{1}\left[\theta^{\text{test}}\in \max_{\theta \text{ s.t. } p(\theta^{\text{test}}|r^{\text{curr}},c^{\text{curr}},\theta^{\text{curr}})>0} p(R^{\text{test}}=1|C^{\text{test}},\theta^{\text{test}})\right]p(r^{\text{curr}}|c^{\text{curr}},\theta^{\text{curr}})d\theta^{\text{test}}\right] \tag{13}$$
$$\geq \max_{\chi_\phi(C^{\text{curr}})} \mathbb{E}_{c^{\text{curr}}\sim\chi_\phi(\cdot)}\left[\sum_{r^{\text{curr}}\in R}\int_{\theta^*\in\Theta^*} p(R^{\text{test}}=1|C^{\text{test}},\theta^*)p(\theta^*|r^{\text{curr}},c^{\text{curr}},\theta^{\text{curr}})p(r^{\text{curr}}|c^{\text{curr}},\theta^{\text{curr}})\right] \tag{14}$$

We assume that Θ^* is the set of all model parameters that maximize $p(R^{\text{test}}=1|C^{\text{test}},\theta^*)$.

B ZONE ANALYSIS

B.1 PRELIMINARIES

For any arbitrary set $\mathcal{X}$, we use $\Delta(\mathcal{X})$ to denote the space of all probability distributions with support on $\mathcal{X}$. For any two arbitrary sets $\mathcal{X}$ and $\mathcal{Y}$, we denote the collection of all functions mapping between them as $\{\mathcal{X} \to \mathcal{Y}\} \triangleq \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$.

B.2 PROBLEM FORMULATION

A standard choice for representing a single sequential decision-making problem is the infinite-horizon, discounted Markov Decision Process (MDP) (Bellman, 1957; Puterman, 1994) defined by $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \mu, \gamma \rangle$. Here $\mathcal{S}$ denotes a set of states, $\mathcal{A}$ is a set of actions, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a deterministic reward function providing evaluative feedback signals (in the unit interval), $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a transition function prescribing distributions over next states, $\mu \in \Delta(\mathcal{S})$ is an initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. Building on this formalism, our work is concerned with decision-making agents that endeavor to learn optimal behaviors across multiple tasks or goals (Kaelbling, 1993; Schaul et al., 2015), a problem we formulate as a Contextual MDP (Brunskill & Li, 2013; Hallak et al., 2015; Modi et al., 2018) (CMDP) given by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{C}, \mathcal{R}, \mathcal{T}, \mu, \gamma, \chi \rangle$. Here, each task of interest is identified by an individual context contained in $\mathcal{C}$ while jointly sharing the same state-action space $\mathcal{S} \times \mathcal{A}$ and discount factor γ; meanwhile, variations in the context lead to different environment configurations that may differ in transition structure $\mathcal{T} : \mathcal{C} \times \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, reward structure $\mathcal{R} : \mathcal{C} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, and initial state distribution $\mu : \mathcal{C} \to \Delta(\mathcal{S})$.

At the beginning of an episode, a single random context is sampled $c \sim \chi(\cdot) \in \Delta(\mathcal{C})$ and held fixed as the agent contends with learning optimal behavior in the resulting MDP denoted by $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}_c, \mathcal{T}_c, \mu_c, \gamma \rangle$. At each discrete timestep $t \in \mathbb{N}$, beginning with an initial state $s_0 \sim \mu_c(\cdot)$, the agent observes the current state $s_t \in \mathcal{S}$, execute an action $a_t \in \mathcal{A}$, enjoys a reward $r_t = \mathcal{R}_c(s_t, a_t)$, and transitions to the next state $s_{t+1} \sim \mathcal{T}_c(\cdot \mid s_t, a_t)$.

Defining $\Pi \triangleq \{\mathcal{S} \times \mathcal{C} \to \Delta(\mathcal{A})\}$, a contextual policy $\pi \in \Pi$ encodes a pattern of behavior that maps the current context and state to a distribution over actions. For any fixed context c, the performance of an agent in the resulting MDP when starting in state $s \in \mathcal{A}$ and taking action $a \in \mathcal{A}$ is assessed by the associated action-value function $Q_c^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_c(s_t, a_t) \mid s_0 = s, a_0 = a\right]$, where the expectation integrates over randomness in the action selections $a_t \sim \pi(\cdot \mid s_t, c)$ and transition dynamics $s_{t+1} \sim \mathcal{T}_c(\cdot \mid s_t, a_t)$. With the corresponding value function defined as $V_c^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot \mid s_t, c)}\left[Q_c^\pi(s, a)\right]$, we slightly abuse notation and use $V_c^\pi(\mu) \triangleq \mathbb{E}_{s_0 \sim \mu(\cdot \mid c)}\left[V_c^\pi(s_0)\right]$ to integrate over the randomness in the initial state. The optimal CMDP policy $\pi^\star$ is defined as achieving supremal value $\sup_{\pi \in \Pi} \mathbb{E}_{c \sim \chi(\cdot)}\left[V_c^\pi(\mu)\right]$.

Our work operates in a general function-approximation setting where individual policies $\pi_\theta \in \Pi$ are parameterized by a vector $\theta \in \Theta \subset \mathbb{R}^d$ of arbitrary dimension d, for instance representing the weights of a neural network with fixed architecture. Consequently, the optimal policy $\pi_{\theta^\star}$ within this policy class $\Pi_\Theta \triangleq \{\pi_\theta \in \Pi \mid \theta \in \Theta\} \subseteq \Pi$ is defined as achieving $\sup_{\pi_\theta \in \Pi_\Theta} \mathbb{E}_{c \sim \chi(\cdot)}\left[V_c^{\pi_\theta}(\mu)\right]$. In this setting, the approximation error associated with a particular choice of policy parameterization is then given by $\mathbb{E}_{c \sim \chi(\cdot)}\left[V_c^\star(\mu) - V_c^{\pi_{\theta^\star}}(\mu)\right]$.

A priori, there is no reason to suspect that the context distribution χ an agent is charged with solving will be tailored in any sort of helpful manner to facilitate rapid or efficient learning. Intuitively, CMDPs that arise in application areas of interest will likely consist of a rich, expressive context space $\mathcal{C}$ alongside a distribution of challenging, complex tasks χ that can be easily specified by a domain expert. Consequently, the onerous burden of mastering a difficult collection of tasks with little to no scaffolding falls to the agent. This reality motivates the use of a teacher-student framework wherein the agent is viewed as a student who gradually faces tasks prescribed by a teacher. A successful teacher can incrementally synthesize a useful curriculum of tasks for the student to solve, building

competency that allows to student to ultimately generalize and succeed across the original collection of challenging tasks prescribed by the distribution χ.

In the next section, we provide a illustrative analysis that identifies a particular objective function for a teacher to maximize whose corresponding lower bound motivates the two key elements of the ZONE framework. Our proof techniques are inspired by the Natural Policy Gradient regret lemma of Agarwal et al. (2021) and the performance guarantee for the Policy Search by Dynamic Programming algorithm of Bagnell et al. (2003).

B.3 A Teacher Objective for Deriving the ZONE

Recall that a function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if

$$||\nabla f(x) - \nabla f(x')||_2 \leq \beta ||x - x'||_2 \qquad \forall x, x' \in \mathbb{R}^d.$$

A consequence of this is that $\nabla^2 f(x) \preceq \beta I, \forall x \in \mathbb{R}^d$ and so, by Taylor's Theorem,

$$|f(x') - f(x) - \nabla f(x)^\top (x' - x)| \leq \frac{\beta}{2} ||x' - x||_2^2 \qquad \forall x, x' \in \mathbb{R}^d.$$

Assumption 1. *(Policy Smoothness) We assume that* $\log(\pi_\theta(a \mid s, c))$ *is a* β*-smooth function of* $\theta \in \Theta$, $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$ *and* $c \in \mathcal{C}$.

Let $\tau = (s_0, a_0, s_1, a_1, \ldots)$ be a random trajectory sampled according to a current student policy π_θ under a fixed context $c \in \mathcal{C}$. Defining the advantage function as

$$A_c^\pi(s, a) = Q_c^\pi(s, a) - V_c^\pi(s),$$

we consider an abstract policy-gradient method that updates the student policy parameters based on the advantage function and a learning rate $\eta \in \mathbb{R}_{\geq 0}$ via

$$\theta' = \theta + \eta \nabla_\theta \log(\pi_\theta(a \mid s, c)) A_c^{\pi_\theta}(s, a).$$

In practice, one would choose a suitable estimator of the advantage function (Mnih et al., 2016; Schulman et al., 2016). Suppose that on-policy policy-gradient updates are performed sequentially on the state-action pairs of the sampled trajectory τ so that the student policy parameters at the beginning of the episode are $\theta^{(0)} = \theta$ and

$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta \log(\pi_{\theta^{(t)}}(a_t \mid s_t, c)) A_c^{\pi_\theta}(s_t, a_t).$$

By Assumption 1, we have that

$$\begin{aligned}
\log\left(\frac{\pi_{\theta^{(t+1)}}(a_t \mid s_t, c)}{\pi_{\theta^{(t)}}(a_t \mid s_t, c)}\right) &= \log(\pi_{\theta^{(t+1)}}(a_t \mid s_t, c)) - \log(\pi_{\theta^{(t)}}(a_t \mid s_t, c)) \\
&\geq \nabla_\theta \log(\pi_{\theta^{(t)}}(a_t \mid s_t, c))^\top \left(\theta^{(t+1)} - \theta^{(t)}\right) - \frac{\beta}{2} ||\theta^{(t+1)} - \theta^{(t)}||_2^2 \\
&= \eta A_c^{\pi_\theta}(s_t, a_t) ||\nabla_\theta \log(\pi_{\theta^{(t)}}(a_t \mid s_t, c))||_2^2 - \frac{\eta^2 \beta}{2} A_c^{\pi_\theta}(s_t, a_t)^2 ||\nabla_\theta \log(\pi_{\theta^{(t)}}(a_t \mid s_t, c))||_2^2 \\
&\geq \eta A_c^{\pi_\theta}(s_t, a_t) \left(1 - \frac{\eta \beta}{2(1-\gamma)}\right) ||\nabla_\theta \log(\pi_{\theta^{(t)}}(a_t \mid s_t, c))||_2^2,
\end{aligned}$$

where the final line leverages the fact that rewards are bounded in the unit interval implying value is upper bounded by $\frac{1}{(1-\gamma)}$. Note that this inequality only holds in this exact form for the (random) state-action pair (s_t, a_t) that led to the update from policy parameters $\theta^{(t)}$ to $\theta^{(t+1)}$. Let $\rho_c^{\pi_\theta}$ denote the distribution over trajectories induced by the policy π_θ under context $c \in \mathcal{C}$. Let π_θ denote the student policy at the start of the episode and $\pi_{\theta'}$ denote the updated student policy after the episode terminates.

For brevity, we omit the state and context arguments to each policy in the following. For any trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots)$, we introduce notation to denote a partial trajectory whose start and end are indexed by timesteps $i, j \in \mathbb{N}$ respectively: $\tau_i^j = (s_i, a_i, s_{i+1}, a_{i+1}, \ldots, s_{j-1}, a_{j-1}, s_j, a_j)$.

With a further abuse of notation, we still use ρ_c^π to denote the distribution over such partial trajectories sampled while executing policy π in the MDP induced under context $c \in \mathcal{C}$. An objective for the teacher $\Lambda \in \Delta(\mathcal{C})$ is

$$\max_{\Lambda \in \Delta(\mathcal{C})} \mathbb{E}_{c \sim \Lambda(\cdot)} \left[\sum_{t=0}^{\infty} \mathbb{E}_{\tau_0^{t-1} \sim \rho_c^{\pi_{\theta^\star}}(\cdot)} \left[\mathbb{E}_{s_t \sim \mathcal{T}_c(\cdot | s_{t-1}, a_{t-1})} \left[D_{\text{KL}}(\pi_{\theta^\star} \mid\mid \pi_{\theta^{(t)}}) - D_{\text{KL}}(\pi_{\theta^\star} \mid\mid \pi_{\theta^{(t+1)}}) \right] \right] \right].$$

At a high level, this says that a good teacher prescribes a distribution over environments for a fixed student such that the resulting policy updates bring the student closer to the optimal policy in Π_θ. In slightly more detail, this is achieved by examining rollouts of increasing lengths generated by the optimal policy $\pi_{\theta^\star}$ and assessing the reduction in KL-divergence between the student policy and the optimal policy before and after the policy-gradient update. For brevity, we continue onward assuming a fixed context $c \in \mathcal{C}$, allowing us to drop the outermost expectation.

Let $\mathcal{X}$ be an arbitrary set consider any two distributions $\nu, \nu' \in \Delta(\mathcal{X})$. Recall that the total variation distance is an integral probability metric (Müller, 1997; Sriperumbudur et al., 2009) defined as

$$D_{\text{TV}}(\nu \mid\mid \nu') = \sup_{f \in \mathcal{F}} |\mathbb{E}_{x \sim \nu(\cdot)} [f(x)] - \mathbb{E}_{x \sim \nu'(\cdot)} [f(x)]|, \qquad \mathcal{F} = \{f : \mathcal{X} \to \mathbb{R} \mid ||f||_\infty \leq 1\}.$$

Consequently, for any function $f : \mathcal{X} \to \mathbb{R}$ such that $||f||_\infty \leq C < \infty$, it follows that

$$|\mathbb{E}_{x \sim \nu(\cdot)} [f(x)] - \mathbb{E}_{x \sim \nu'(\cdot)} [f(x)]| \leq C \cdot D_{\text{TV}}(\nu \mid\mid \nu').$$

In order to apply this fact to induce a distribution shift, we make the following assumption which controls for the variability in log-likelihood ratio between two policies separated by a single policy-gradient update:

Assumption 2. *(Bounded log-likelihood ratio) Let θ be an initial set of policy parameters and θ' denote the policy parameters after a single policy-gradient update. For any fixed $c \in \mathcal{C}$, we assume that there exists a numerical constant $C < \infty$ such that $\log\left(\frac{\pi_{\theta'}(a|s,c)}{\pi_\theta(a|s,c)}\right) \leq C$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.*

Expanding from above, we have

$$\begin{aligned}
&\sum_{t=0}^{\infty} \mathbb{E}_{\tau_0^{t-1} \sim \rho_c^{\pi_{\theta^\star}}(\cdot)} \left[\mathbb{E}_{s_t \sim \mathcal{T}_c(\cdot | s_{t-1}, a_{t-1})} \left[D_{\text{KL}}(\pi_{\theta^\star} \mid\mid \pi_{\theta^{(t)}}) - D_{\text{KL}}(\pi_{\theta^\star} \mid\mid \pi_{\theta^{(t+1)}}) \right] \right] \\
&\quad = \sum_{t=0}^{\infty} \mathbb{E}_{\tau_0^{t-1} \sim \rho_c^{\pi_{\theta^\star}}(\cdot)} \left[\mathbb{E}_{s_t \sim \mathcal{T}_c(\cdot | s_{t-1}, a_{t-1})} \left[\mathbb{E}_{a_t \sim \pi_{\theta^\star}(\cdot | s_t, c)} \left[\log \left(\frac{\pi_{\theta^{(t+1)}}(a_t \mid s_t, c)}{\pi_{\theta^{(t)}}(a_t \mid s_t, c)} \right) \right] \right] \right] \\
&\quad = \sum_{t=0}^{\infty} \mathbb{E}_{\tau_0^{t} \sim \rho_c^{\pi_{\theta^\star}}(\cdot)} \left[\log \left(\frac{\pi_{\theta^{(t+1)}}(a_t \mid s_t, c)}{\pi_{\theta^{(t)}}(a_t \mid s_t, c)} \right) \right] \\
&\quad = \sum_{t=0}^{\infty} \mathbb{E}_{\tau \sim \rho_c^{\pi_{\theta^\star}}(\cdot)} \left[\log \left(\frac{\pi_{\theta^{(t+1)}}(a_t \mid s_t, c)}{\pi_{\theta^{(t)}}(a_t \mid s_t, c)} \right) \right] \\
&\quad = \mathbb{E}_{\tau \sim \rho_c^{\pi_{\theta^\star}}(\cdot)} \left[\sum_{t=0}^{\infty} \log \left(\frac{\pi_{\theta^{(t+1)}}(a_t \mid s_t, c)}{\pi_{\theta^{(t)}}(a_t \mid s_t, c)} \right) \right] \\
&\quad \geq \mathbb{E}_{\tau \sim \rho_c^{\pi_{\theta^\star}}(\cdot)} \left[\sum_{t=0}^{\infty} \gamma^t \log \left(\frac{\pi_{\theta^{(t+1)}}(a_t \mid s_t, c)}{\pi_{\theta^{(t)}}(a_t \mid s_t, c)} \right) \right] \\
&\quad \geq \mathbb{E}_{\tau \sim \rho_c^{\pi_\theta}(\cdot)} \left[\sum_{t=0}^{\infty} \gamma^t \log \left(\frac{\pi_{\theta^{(t+1)}}(a_t \mid s_t, c)}{\pi_{\theta^{(t)}}(a_t \mid s_t, c)} \right) \right] - \frac{C}{(1-\gamma)} \cdot D_{\text{TV}}(\rho_c^{\pi_{\theta^\star}} \mid\mid \rho_c^{\pi_\theta}) \\
&\quad \geq \eta \left(1 - \frac{\eta\beta}{2(1-\gamma)} \right) \mathbb{E}_{\tau \sim \rho_c^{\pi_\theta}(\cdot)} \left[\sum_{t=0}^{\infty} \gamma^t A_c^{\pi_\theta}(s_t, a_t) ||\nabla_\theta \log (\pi_{\theta^{(t)}}(a_t \mid s_t, c))||_2^2 \right] - \frac{C}{(1-\gamma)} \cdot D_{\text{TV}}(\rho_c^{\pi_{\theta^\star}} \mid\mid \rho_c^{\pi_\theta}) \\
&\quad \geq \eta \left(1 - \frac{\eta\beta}{2(1-\gamma)} \right) \mathbb{E}_{\tau \sim \rho_c^{\pi_\theta}(\cdot)} \left[\sum_{t=0}^{\infty} \gamma^t A_c^{\pi_\theta}(s_t, a_t) ||\nabla_\theta \log (\pi_{\theta^{(t)}}(a_t \mid s_t, c))||_2^2 \right] - \frac{C}{\sqrt{2}(1-\gamma)} \cdot \sqrt{D_{\text{KL}}(\rho_c^{\pi_{\theta^\star}} \mid\mid \rho_c^{\pi_\theta})},
\end{aligned}$$

where the first inequality follows as $\gamma \in [0, 1]$, the second inequality leverages Assumption 2 and the aforementioned fact to shift trajectory distributions, the penultimate inequality applies our earlier policy-gradient norm lower bound, and the final inequality follows as Pinsker's inequality (Pinsker, 1964).

Now accounting for the randomness in the contexts, we have the following lower bound to the teacher (maximization) objective:

$$\mathbb{E}_{c\sim\Lambda(\cdot)}\left[\sum_{t=0}^{\infty}\mathbb{E}_{\tau_0^{t-1}\sim\rho_c^{\pi_{\theta^\star}}(\cdot)}\left[\mathbb{E}_{s_t\sim\mathcal{T}_c(\cdot|s_{t-1},a_{t-1})}\left[D_{\mathrm{KL}}(\pi_{\theta^\star}\;||\;\pi_{\theta^{(t)}})-D_{\mathrm{KL}}(\pi_{\theta^\star}\;||\;\pi_{\theta^{(t+1)}})\right]\right]\right]$$

$$\geq \mathbb{E}_{c\sim\Lambda(\cdot)}\Bigg[\underbrace{\eta\left(1-\frac{\eta\beta}{2(1-\gamma)}\right)\mathbb{E}_{\tau\sim\rho_c^{\pi_\theta}}\left[\sum_{t=0}^{\infty}\gamma^t A_c^{\pi_\theta}(s_t,a_t)||\nabla_\theta\log\left(\pi_{\theta^{(t)}}(a_t\mid s_t,c)\right)||_2^2\right]}_{①}-\underbrace{\frac{C}{\sqrt{2}(1-\gamma)}\cdot\sqrt{D_{\mathrm{KL}}(\rho_c^{\pi_{\theta^\star}}\;||\;\rho_c^{\pi_\theta})}}_{②}\Bigg]$$

Observe that term ① captures a notion of learning potential for the current student under tasks sampled by the teacher. Intuitively, this is quantified by looking at the average, advantage-weighted policy-gradient norm across trajectories generated by the student policy π_θ in each sampled context $c \sim \Lambda(\cdot)$. Orthogonally, term ② encapsulates a notion of problem difficulty as measured by how much the trajectory distribution of the student differs from that of the optimal policy in each sampled context.

When this latter quantity ② is too large, suggesting an overwhelmingly difficult problem for the student where a large number of samples or environment interactions will be needed to improve performance, this term overpowers any learning potential captured in term ①. Conversely, tasks that are too easy for the student will lead to scenarios where both ② and the advantage terms encountered in ① will be small (or even zero), suggesting little opportunity for improving performance. Naturally, the "sweet spot" or ZPD suggested by this lower bound consists of a teacher selecting tasks with reasonably large policy gradient norms (signaling learning potential) while being within the students means (as measured by the divergence between the student's trajectory distribution from that of the optimal policy). Practical approaches to automated curriculum design use a notion of pseudo-regret in lieu of ②, accounting for a lack of knowledge about the optimal policy in advance (Florensa et al., 2018b; Dennis et al., 2020).

C Additional teacher analysis

Here we include additional analysis on the teacher.

C.1 PAIRED on MuJoCo environments

Figure 6 shows the rejection rate and student's gradient norms. Interestingly, REJECT tends to have higher rejection rates. In general, GRAD and PAIRED have similar rejection rates.

C.2 Goal GAN

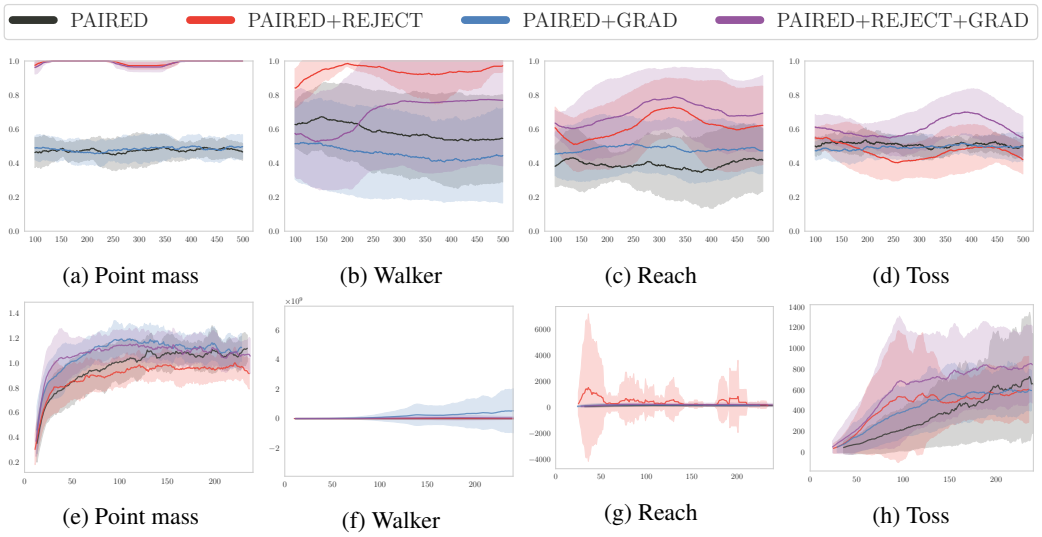
Figure 7 shows the rejection rates on the Goal GAN environments and the MuJoCo environments. In general, the rejection rates are similar across environments.

D Algorithm info

We provide information on the PAIRED and Goal GAN implementations.

D.1 PAIRED on MiniGrid environments

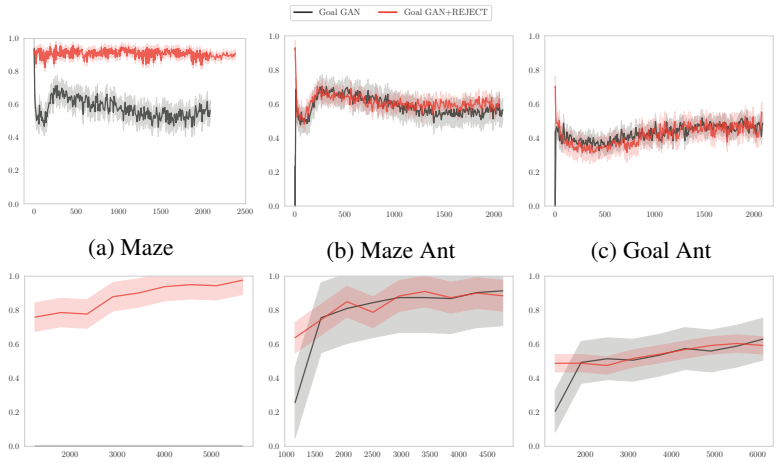
We use the implementation at `https://github.com/ucl-dark/paired` which is based on Dennis et al. (2020)'s implementation. We do not change any hyperparameters in their algorithm.



(a) Point mass (b) Walker (c) Reach (d) Toss

(e) Point mass (f) Walker (g) Reach (h) Toss

Figure 6: On the MuJoCo environments. The rejection rate is reported in the first row (a-d), and the gradient norms are reports in the second row (e-h).



(a) Maze (b) Maze Ant (c) Goal Ant

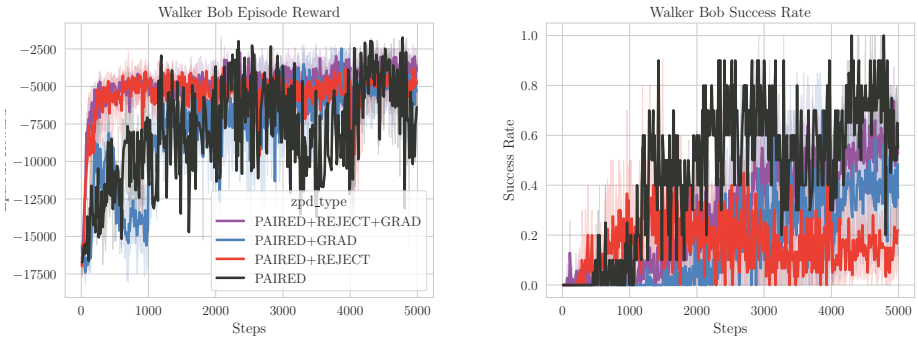
(d) Walker (e) Reach (f) Toss

Figure 7: Reported is the rejection rate. The rate for Goal GAN on the original Goal GAN environments is reported in the first row, and the rate on the MuJoCo environments is reported in the last row.

All hyperparameters are the same as those reported in Dennis et al. (2020). We refer to their paper for more details on PAIRED. We run all the variants of PAIRED with ZONE with 10 seeds.

D.2 Goal GAN on Goal GAN environments

We use the original implementation at https://github.com/florensacc/rllab-curriculum. We do not change any hyperparameters. We refer to their paper (Florensa et al., 2018b) for more details on Goal GAN. We run all the variants of Goal GAN with ZONE with 5 seeds.



(a) (b)

Figure 8: PAIRED on the Walker domain. The figures show the performance of the student on the out-of-domain goals. (a) shows the student's test return over the course of training. (b) shows the student's success rate over the course of training.

D.3 PAIRED AND GOAL GAN ON MUJOCO ENVIRONMENTS

We use Du et al. (2022)'s implementation of PAIRED and Goal GAN on their MuJoCo environments. Their MuJoCo setup is designed to test students on out-of-domain goals, that the students have not yet seen. We run all the variants of both algorithms with ZONE with 10 seeds. We do not change any of the hyperparameters. We use the default Goal GAN implementation in their work. Goal GAN stores 500 goals and the student is evaluated on the goals 3 times. The mean reward is taken and used to determine the label for the teacher. The label is 1 if the mean reward lies within the difficulty criterion ($r \in [0.1, 0.9]$), and 0 otherwise. Every 500 steps, the teacher trains on the labelled data.

PAIRED is implemented using the default parameters from Du et al. (2022)'s algorithms but with symmetrization turned off (ie. we remove the second teacher in their work).

E DO MEASURES OF DIFFICULTY MATTER?

A key component to ZONE is the choice of difficulty measure. Most prior work use reward to model difficulty: The lower of the reward, the more difficulty the problem. If the problem is more difficult, then the student is less likely to succeed on the problem.

However, we find that using *dense* rewards as a proxy measure for difficulty is misleading. Dense rewards are typically used for training students in the MuJoCo control setting, as done in Du et al. (2022). Running ZONE on PAIRED in this setting reveals an interesting discrepancy: ZONE can achieve higher episodic reward than the base algorithm (Figure 8a), however achieves low success (Figure 8b). For example, at 2000 steps, episode reward scores increasingly higher from PAIRED, PAIRED+GRAD, PAIRED+REJECT, to PAIRED+REJECT+GRAD. However, success scores increasingly higher from PAIRED+GRAD, PAIRED+REJECT, PAIRED+REJECT+GRADto PAIRED.

This discrepancy reveals that training the teacher based on a dense-reward difficulty measure can be misleading when the reward function does not correlate well with the student's success. ZONE is sensitive to this choice of difficulty measure that is not well correlated with success. Thus, for MuJoCo experiments, we choose to use the student's success as a measure of difficulty which is what PAIRED in Dennis et al. (2020) and Goal GAN assume. This gives us the results from the previous section in Figure 3 and Figure 4.