OFF-POLICY LEARNING TO REASON WORKS BECAUSE IT IS MORE PESSIMISTIC THAN YOU THINK

Anonymous authors
Paper under double-blind review

ABSTRACT

Large scale reinforcement learning has become a central tool for improving reasoning in large language models. At this scale, generation is often lagged or asynchronous, so updates are performed on data collected by older policies. This makes learning inherently off-policy. Most existing approaches nevertheless remain rooted in PPO-style trust-region objectives, treating training as approximately on-policy and using importance weights to correct distribution mismatch. These corrections can introduce high variance, destabilize optimization, and accelerate entropy collapse. Recent work suggests an alternative: rather than correcting the mismatch, one can embrace off-policy data and remove importance weights, often yielding stronger algorithms. In this paper, we provide an intuitive construction of off-policy objectives that include successful off-policy objectives and show that their effectiveness can be understood through implicit pessimism: they optimize toward target policies that are more conservative than their nominal objectives suggest. This perspective explains why some particular implementation choices improve stability: they implicitly control the effective target distribution. We then propose a principled modification that stabilize this induced distribution and improve off-policy learning.

1 INTRODUCTION

Reinforcement learning (RL) has emerged as a fundamental tool for improving the reasoning capabilities of large language models (LLMs) [1, 2, 3]. By optimizing models against task-specific or verifiable rewards, RL enables gains beyond supervised fine-tuning [1, 2], particularly on complex reasoning tasks such as mathematics, coding, and multi-step problem solving [4, 2, 5].

The reinforcement learning pipelines used for modern language models, however, differ significantly from the classical online setting. At large scale, generating rollouts for reasoning models is expensive. Completions are often produced asynchronously, by lagged inference workers, or by checkpoints that are older than the policy currently being updated. As a result, the data used for training are not sampled from the policy being optimized, but from an older behavior policy. Large-scale reinforcement learning for language model reasoning is therefore inherently off-policy [5, 6, 7].

Despite this, many existing methods are still motivated by on-policy or near-on-policy trust-region algorithms [8], such as PPO [9] and GRPO [10]. These methods attempt to correct the mismatch between the behavior policy and the learned policy using importance weighting, clipping , filtering, or related heuristics [9, 10, 3]. While these corrections are effective in some settings, they can introduce high-variance updates and become unstable when the policy lag is large. In practice, the mechanisms introduced to control these ratios often become central to training stability, rather than secondary implementation details.

Recent work suggests a different approach: instead of correcting off-policy data to make it look on-policy, one can train directly on data sampled from the behavior policy and remove importance weights entirely. $A^\star$-PO [11] and more recently OAPL [7] are a representative example of this direction. Both optimize a simple regression objective, and has been shown to work well in large-scale reasoning settings. However, the reason for its success is not fully understood. In particular, their nominal derivations suggest one target distribution, while practical implementations decouple the temperature used for advantage computation from the temperature used for regularization. These choices are empirically important, but their role is not explained by the nominal derivation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this paper, we study ratio-free off-policy objectives. Our goal is not to derive another trust-region surrogate, but to understand what target policy these objectives actually optimize toward when trained on data from an older policy. We show that the answer is different from the usual exponential target associated with KL-regularized policy optimization [12, 8]. Instead, these objectives induce a Lambert-tempered target policy [13]. This target can be more conservative than the nominal exponential objective suggests, especially on high-advantage samples and small regularizations.

This observation provides a simple explanation for why off-policy learning without importance weights can work well. The objective is not merely ignoring the distribution mismatch. Rather, under the right advantage normalization, the mismatch induces a form of implicit pessimism [14, 15, 16]: the learned policy increases the probability of good samples, but does not over-commit exponentially to the highest observed rewards. This is particularly valuable in language-model reasoning, where rollouts are stale and a small number of sampled completions can otherwise dominate the update.

Our analysis also reveals that this pessimism is not automatic. The target policy depends sensitively on how the advantage is normalized. We identify three qualitative regimes. In the positive regime, the target is conservative and tempers large advantages. At the boundary, the target recovers the usual exponential update, which is fragile because small changes in the advantage baseline can move the objective into a different regime. In the negative regime, the target becomes more aggressive than the exponential update, making it more prone to entropy collapse.

This perspective clarifies the behavior of OAPL. We show that the advantage estimator suggested by its nominal theory can induce an unstable target at the population level when the group size is finite, even though it is normalized on each sampled group. This explains why the implementation of $A^\star$-PO [11] and OAPL [7] deviates from the theory by using a larger temperature β for advantage computation than for regularization. Our analysis shows that this choice moves the advantage closer to a centered group advantage, which pushes the induced target back toward the pessimistic regime.

Motivated by this understanding, we propose a simpler, more principled modification. Instead of relying on temperature decoupling and hyperparameter tuning as an implementation heuristic, we directly choose the advantage baseline so that the induced target lies in the pessimistic Lambert regime. The resulting rule is a β-shifted mean advantage: start from the usual group-centered reward advantage and shift it upward by β. This choice is easy to implement, introduces no additional temperature hyperparameter, and guarantees that the induced Lambert target is in the desired conservative regime.

Empirically, this modification improves the robustness of off-policy learning. Compared with OAPL, the resulting method is less sensitive to the regularization strength and remains more stable under larger policy lags. In our experiments, it maintains entropy more reliably and avoids the collapse or instability observed in OAPL under aggressive regularization or stale rollouts.

Contributions.

- We characterize the target policies induced by a class of ratio-free off-policy objectives for language-model reasoning, including $A^\star$-PO/OAPL, as Lambert-tempered policies whose implicit pessimism is governed by advantage normalization.
- We explain why the nominal $A^\star$-PO/OAPL advantage can induce an unstable population target and why temperature decoupling improves stability in practice. We then propose a simple β-shifted-mean advantage that targets the pessimistic Lambert regime by construction, eliminating the need for this heuristic.
- We empirically validate the predicted target regimes and show that the proposed advantage improves training stability and robustness across regularization strengths and policy lags. Our method also achieves higher reward than near-on-policy GRPO while retaining more entropy, despite using substantially staler rollouts.

Outline. Section 2 introduces the reinforcement learning setting for language-model reasoning and reviews standard trust-region methods. Section 3 studies ratio-free off-policy objectives and derives their induced target policies. Section 4 characterizes the Lambert regimes, explains the behavior of OAPL, and introduces our shifted-mean advantage. Section 5 empirically validates our theoretical predictions and evaluates robustness across regularization strengths and policy lags, including comparisons with GRPO. Section 6 concludes.

2 PRELIMINARIES

We consider reinforcement learning for language-model reasoning. Let $x \sim \mathcal{D}$ denote a prompt and let $y = (y_1, \ldots, y_T)$ denote a sampled completion. A policy π_θ defines the autoregressive distribution

$$\pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid x, y_{<t}).$$

After generating a completion, the model receives a scalar, bounded reward $r(x, y) \in [0, 1]$, for example from a verifier, a unit test, or a reward model. For a group of $G \geq 2$ completions $y_{1:G}$ sampled for the same prompt, we write $r_i = r(x, y_i)$ and use $\widehat{A}_i$ to denote a group-relative score indicating whether completion y_i should be upweighted or downweighted relative to the group.

Trust-region optimization and GRPO. Trust-region optimization has established itself as a staple for large scale, deep RL [10, 5, 17]. The idea is to optimize a policy that improves π_{old} while staying in its vicinity, in the KL regularized sense. Group Relative Policy Optimization (GRPO) can be viewed as a PPO-style trust-region method adapted to the LLM setting [9, 10]. Given a behavior policy π_{old} used to generate completions, GRPO optimizes a clipped surrogate objective of the form

$$\mathbb{E}_{y_{1:G} \sim \pi_{\text{old}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{T_i} \sum_{t=1}^{T_i} \min \left(\rho_{i,t}(\theta) \widehat{A}_i,\ \text{clip}(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\, \widehat{A}_i \right) \right], \tag{1}$$

where

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\text{old}}(y_{i,t} \mid x, y_{i,<t})}, \qquad \widehat{A}_i = \frac{r_i - \bar{r}_G}{\sigma_G}.$$

The ratio $\rho_{i,t}(\theta)$ measures the change in token probability between the policy being optimized and the behavior policy that generated the data. The clipping operation restricts this ratio, preventing the new policy from moving too far from π_{old} on sampled tokens.

This objective is effective when the data are fresh and the optimized policy remains close to the behavior policy. However, large-scale LLM RL often violates this near-on-policy assumption [3, 6]. Indeed, rollouts are often generated by stale checkpoints. As optimization proceeds, the current policy π_θ can move far from the behavior policy π_{old}, making the data increasingly off-policy. In this regime, importance weighting becomes high-variance, and clipping or filtering becomes a central stabilizing heuristic rather than a minor implementation detail.

Embracing Off-Policy Generations. A natural way to avoid importance weighting is to define the desired policy directly as an improvement over the behavior policy π_{old}. The dominant formulation is the KL-regularized trust-region objective. For a fixed context x, it solves

$$\max_{\pi}\ \mathbb{E}_{y \sim \pi(\cdot|x)} \left[r(y, x) \right] - \beta\, \text{KL}\left(\pi(\cdot \mid x) \,\|\, \pi_{\text{old}}(\cdot \mid x) \right).$$

The solution is the exponentially tilted behavior policy

$$\pi^\star_\beta(y \mid x) = \frac{\exp(r(y, x)/\beta)}{\mathbb{E}_{y' \sim \pi_{\text{old}}(\cdot|x)} \left[\exp(r(y', x)/\beta) \right]} \pi_{\text{old}}(y \mid x).$$

Equivalently, the optimal sentence-level log-ratio satisfies

$$\beta \log \frac{\pi^\star_\beta(y \mid x)}{\pi_{\text{old}}(y \mid x)} = r(y, x) - \beta \log \mathbb{E}_{y' \sim \pi_{\text{old}}(\cdot|x)} \left[\exp(r(y', x)/\beta) \right].$$

Thus the KL-regularized target can be viewed as specifying an optimal advantage, or target log-ratio, relative to π_{old}.

In practice, the expectation above is unavailable and is estimated from a group of completions $y_{1:G} \sim \pi_{\text{old}}(\cdot \mid x)$, with $r_i = r(y_i, x)$. This gives the empirical advantage

$$\widehat{A}^\beta_i = r_i - \beta \log \left(\frac{1}{G} \sum_{j=1}^{G} \exp(r_j/\beta) \right).$$

This estimator is biased for finite G, due to the logarithm of the empirical average, but it is consistent.

Recent off-policy objectives use this target without trying to correct the behavior-policy samples into on-policy samples. Instead, they regress the current policy log-ratio toward the KL-regularized target:

$$\min_{\pi_\theta} \mathbb{E}_{y_{1:G}\sim\pi_{\text{old}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^{G} \left(\beta \log \frac{\pi_\theta(y_i \mid x)}{\pi_{\text{old}}(y_i \mid x)} - \widehat{A}_i^{\beta} \right)^2 \right].$$

This objective embraces off-policy data rather than correcting it with importance weights: the old policy provides the samples, and the loss directly constrains how the learned policy should move relative to π_{old} on those samples.

3 Regularized Off-Policy Optimization

Importance-weight-free learning can lead to behavior-tilted targets. Instead of correcting samples from π_{old} to mimic on-policy data from π_θ, we first consider the most direct ratio-free objective: weighted maximum likelihood on behavior-policy samples.

Let $G \geq 1$ and let $y_{1:G} \sim \pi_{\text{old}}(\cdot \mid x)$ denote a group of completions sampled from the behavior policy. For each completion y_i, let $r_i = r(x, y_i)$ be its reward and let $u_i = u_G(r_i; r_{1:G}) \geq 0$ be a nonnegative reward-dependent weight. A natural off-policy objective is

$$\mathbb{E}_{y_{1:G}\sim\pi_{\text{old}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^{G} u_i \log \pi_\theta(y_i \mid x) \right]. \tag{2}$$

This objective is attractive because it contains no importance-weighting ratio. The behavior policy simply provides samples, and high-reward samples receive larger weights.

A common choice in offline RL [18, 19] is an exponential reward weight, $u_i = \exp\left((r_i - \bar{r}_G)/\eta\right)$, which up-weights completions that outperform the group average. This objective is desirable as it makes the behavior policy the base measure of the target distribution. To see this, fix a prompt x and consider the population version of Eq. (2) over all sequence distributions $\pi(\cdot \mid x)$:

$$\max_{\pi\in\Delta} \mathbb{E}_{y\sim\pi_{\text{old}}(\cdot|x)} \left[u(x, y) \log \pi(y \mid x) \right].$$

The optimizer is $\pi^\star(y \mid x) \propto \pi_{\text{old}}(y \mid x) u(x, y)$. Therefore, when $u(x, y) = \exp(r(x, y)/\eta)$,

$$\pi^\star(y \mid x) \propto \pi_{\text{old}}(y \mid x) \exp\left(\frac{r(x, y)}{\eta} \right).$$

Thus, weighted maximum likelihood recovers a *behavior-tilted* distribution [12]. This is the same qualitative form as the exponentially tilted distributions arising from KL-regularized trust-region objectives: the behavior policy remains the base measure, and the reward only reweights it.

While the theoretical target is desirable, a major problem of this objective is that it can be overly aggressive in finite samples. The expectation under π_{old} is approximated by a small set of sampled completions, so weighted maximum likelihood can rapidly concentrate probability mass on the observed high-weight samples. With sharp exponential weights, this effect is amplified: the empirical target becomes highly peaked, and optimization can collapse toward the weighted empirical distribution. This leads to entropy collapse in practice and requires additional care to control [19].

Ideally, we want an update that preserves the simplicity of Eq. (2) and does not require importance sampling, while avoiding the aggressiveness of exponential behavior tilting. We can achieve this by adding to weighted likelihood a simple quadratic penalty in log-probability space.

Regularize toward π_{old}. Based on the previous observation, our off-policy objective implements the simple intuition of maximizing reward while staying close to π_{old}. We do so by adding a sentence-level regularizer to the weighted likelihood objective. Let $\beta > 0$. We focus on:

$$\mathbb{E}_{y_{1:G}\sim\pi_{\text{old}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^{G} \left(\widehat{A}_i \log \pi_\theta(y_i|x) - \frac{\beta}{2} \left(\log \frac{\pi_\theta(y_i|x)}{\pi_{\text{old}}(y_i|x)} \right)^2 \right) \right], \tag{3}$$

where $\widehat{A}_i$ is a group-wise advantage that will be specified later. The first term is an advantage-weighted log-likelihood term. The second term penalizes squared movement in sentence log-probability relative

to the behavior policy. Importantly, this penalty is not an importance weight: it does not multiply the gradient by a likelihood ratio, and it does not attempt to transform off-policy samples into on-policy samples. Instead, it directly regularizes how far the learned policy can move from the behavior policy on observed sentences.

This objective has a clean stationary target policy, characterized by the next proposition.

Proposition 3.1 (Lambert-tempered policy – sentence level)**.** *Fix a context x, and let $\mathcal{Y}(x)$ denote the set of possible responses. Let $A(x, y)$ be a sentence-level advantage for response $y \in \mathcal{Y}(x)$. Consider the sentence-level objective*

$$\mathcal{J}_s(\pi) = \sum_{y \in \mathcal{Y}(x)} \pi_{\text{old}}(y \mid x) \left[A(x, y) \log \pi(y \mid x) - \frac{\beta}{2} \left(\log \frac{\pi(y \mid x)}{\pi_{\text{old}}(y \mid x)} \right)^2 \right],$$

optimized over distributions $\pi(\cdot \mid x) \in \Delta(\mathcal{Y}(x))$, with $\beta > 0$. Let $\pi^\star(\cdot \mid x)$ be an interior stationary point of $\mathcal{J}_s$. If its normalization multiplier satisfies $\tau_s > 0$, then for every $y \in \mathcal{Y}(x)$,

$$\pi^\star(y \mid x) = \frac{\pi_{\text{old}}(y \mid x)}{\tau_s} W_0 \left(\tau_s \exp \left(\frac{A(x, y)}{\beta} \right) \right), \tag{4}$$

where W_0 is the principal branch of the Lambert W function, and τ_s is chosen so that

$$\sum_{y \in \mathcal{Y}(x)} \pi^\star(y \mid x) = 1.$$

In the limit $\tau_s \to 0$, Eq. (4) recovers

$$\pi^\star(y \mid x) \propto \pi_{\text{old}}(y \mid x) \exp \left(\frac{A(x, y)}{\beta} \right).$$

This proposition is the main lens through which we analyze ratio-free off-policy objectives. It shows that, once the advantage A is fixed, the objective does not generally target the usual exponential KL-regularized policy. Instead, it induces a Lambert-tempered target whose shape is controlled by the normalization multiplier τ_s. Before studying the regimes of this multiplier, we first show that this objective is closely connected to optimal advantage regression losses [11, 7].

Connection to $A^\star$-PO and OAPL. These losses can be viewed as a squared regression objective on the sentence-level log-ratio between the learned policy and the behavior policy. To see this, define

$$\ell_\theta(y_i \mid x) = \log \frac{\pi_\theta(y_i \mid x)}{\pi_{\text{old}}(y_i \mid x)}.$$

For a fixed sample y_i and a fixed advantage $\widehat{A}_i$, the corresponding term in our objective satisfies

$$\widehat{A}_i \log \pi_\theta(y_i \mid x) - \frac{\beta}{2} \left(\log \frac{\pi_\theta(y_i \mid x)}{\pi_{\text{old}}(y_i \mid x)} \right)^2 = -\frac{1}{2\beta} \left(\beta \ell_\theta(y_i \mid x) - \widehat{A}_i \right)^2 + C_i,$$

where C_i do not depend on θ. Therefore, maximizing our regularized weighted log-likelihood objective is equivalent, up to constants and a positive rescaling, to minimizing

$$\mathbb{E}_{y_{1:G} \sim \pi_{\text{old}}(\cdot \mid x)} \left[\frac{1}{G} \sum_{i=1}^{G} \left(\beta \log \frac{\pi_\theta(y_i \mid x)}{\pi_{\text{old}}(y_i \mid x)} - \widehat{A}_i \right)^2 \right].$$

This is precisely the $A^\star$-PO / OAPL form when $\widehat{A}_i$ is chosen as the log-sum-exp centered advantage

$$\widehat{A}_i^\beta = r_i - \beta \log \left(\frac{1}{G} \sum_{j=1}^{G} \exp(r_j / \beta) \right).$$

Thus, OAPL is not a separate objective from the one analyzed above; it is a particular instantiation obtained by a specific choice of advantage. Consequently, Proposition 3.1 characterizes the target policy induced by OAPL as well.

4 PRACTICAL OFF-POLICY OPTIMIZATION IS PESSIMISTIC

4.1 ANALYSIS OF THE OBJECTIVE

Our previous proposition shows that the Lambert target policy is governed by the density ratio $\rho_s^\star(y \mid x) = \pi^\star(y \mid x)/\pi_{\text{old}}(y \mid x)$. At the optimal solution, this ratio satisfies

$$\log \rho_s^\star(y \mid x) + \tau_s \rho_s^\star(y \mid x) = \frac{A(x,y)}{\beta}. \tag{5}$$

Thus τ_s is not merely a normalization constant: it determines the shape of the target policy. The sign of τ_s is controlled by the normalization of the exponential tilt. Define

$$Z_{\exp}(x) = \sum_{y \in \mathcal{Y}(x)} \pi_{\text{old}}(y \mid x) \exp\left(\frac{A(x,y)}{\beta}\right).$$

Then, on the principal branch,

$$\tau_s > 0 \iff Z_{\exp}(x) > 1, \qquad \tau_s = 0 \iff Z_{\exp}(x) = 1,$$

and, whenever a real stationary solution exists, $\tau_s < 0 \iff Z_{\exp}(x) < 1$.

This means that the way the advantage $A(x,y)$ is centered *directly determines* which regime the target policy lies in. In particular, adding a context-dependent shift $b(x)$ to the advantage rescales the exponential mass by $e^{-b(x)/\beta}$. Unlike in standard policy-gradient estimators, where such shifts can reduce variance without changing the objective, such shifts are not innocuous here: they change the target distribution optimized by the off-policy log-likelihood objective.

The pessimistic regime: $\tau_s > 0$. When $\tau_s > 0$, the Lambert target is more conservative than the exponential, KL-regularized target. Indeed, for every response y,

$$\rho_s^\star(y \mid x) = \frac{1}{\tau_s} W_0\left(\tau_s \exp\left(\frac{A(x,y)}{\beta}\right)\right) \le \exp\left(\frac{A(x,y)}{\beta}\right),$$

since $W_0(z) \le z$ for $z \ge 0$. Moreover, $W_0(z)$ is linear near the origin: $W_0(z) = z + O(z^2)$ as $z \to 0$, and grows only logarithmically for large z

$$W_0(z) = \log z - \log\log z + o(1) \quad \text{as } z \to \infty.$$

Therefore low-advantage responses are treated similarly to the exponential update, whereas very high-advantage responses are strongly tempered. This is precisely the desired pessimistic behavior: the policy does not over-commit exponentially to a small number of high-reward off-policy samples, which helps protect against over-optimization and entropy collapse [13].

The special case $\tau_s = 1$ gives the particularly simple target

$$\rho_s^\star(y \mid x) = W_0\left(\exp\left(\frac{A(x,y)}{\beta}\right)\right).$$

For large positive advantages, this ratio grows approximately linearly in $A(x,y)/\beta$, rather than exponentially. Thus $\tau_s = 1$ implements a simple form of pessimism [14]: it preserves the ordering induced by the advantage while substantially reducing the incentive to collapse onto the largest estimated rewards. This policy was demonstrated to be minimax for offline preference learning [13].

The exponential boundary: $\tau_s = 0$. The case $\tau_s = 0$ is the boundary between pessimistic and optimistic targets. Using Eq. (5), the stationary policy reduces to the usual exponential-tilting target associated with KL-regularized policy optimization,

$$\pi^\star(y \mid x) \propto \pi_{\text{old}}(y \mid x) \exp\left(\frac{A(x,y)}{\beta}\right).$$

However, in the present objective, reaching exactly this regime requires the exponential tilt to be correctly normalized at the population level:

$$\sum_y \pi_{\text{old}}(y \mid x) \exp\left(\frac{A(x,y)}{\beta}\right) = 1.$$

This makes the $\tau_s = 0$ target delicate. Small errors in the advantage baseline can move the population objective into either the pessimistic regime $\tau_s > 0$ or the unstable regime $\tau_s < 0$. Moreover, the exponential target is highly sensitive to large positive advantage estimates, since its density ratio grows as $\exp(A(x, y)/\beta)$. This sensitivity is one reason why directly targeting the KL-regularized update can be brittle in large-scale off-policy training.

The unstable regime: $\tau_s < 0$. When $\tau_s < 0$, write $\tau_s = -|\tau_s|$. The stationarity equation becomes

$$\log \rho_s^\star(y \mid x) + \tau_s \rho_s^\star(y \mid x) = \frac{A(x, y)}{\beta}. \tag{6}$$

Equivalently,

$$\rho_s^\star(y \mid x) = \exp\left(\frac{A(x, y)}{\beta}\right) \exp\left(|\tau_s| \rho_s^\star(y \mid x)\right).$$

Thus, whenever the principal-branch solution exists,

$$\rho_s^\star(y \mid x) \geq \exp\left(\frac{A(x, y)}{\beta}\right).$$

The target is therefore more aggressive than the exponential update: it amplifies high-advantage responses even more strongly than the KL-regularized target. This regime is also numerically fragile. From Equation (6),

$$\frac{\partial \rho_s^\star(y \mid x)}{\partial (A(x, y)/\beta)} = \frac{\rho_s^\star(y \mid x)}{1 + \tau_s \rho_s^\star(y \mid x)}.$$

For $\tau_s < 0$, the denominator can approach zero, causing the target ratio to become extremely sensitive to small errors in the advantage. This makes $\tau_s < 0$ a regime to avoid: it is more peaked than the exponential target, more sensitive to advantage noise, and more prone to rapid entropy collapse.

4.2 The nominal advantage regression target is in the unstable regime

As discussed above, $A^\star$-PO and OAPL instantiate the regularized off-policy objective with a specific choice of advantage. The previous analysis shows that advantage normalization strongly influences the induced target policy, and is therefore a target-design choice. OAPL uses the log-sum-exp centered advantage

$$\widehat{A}_i^\beta = r_i - \beta \log\left(\frac{1}{G}\sum_{j=1}^{G} \exp(r_j/\beta)\right), \quad A^\beta(x, y) = \mathbb{E}_{\pi_{\text{old}}}[\widehat{A}_i^\beta \mid y_i = y, x]$$

This gives by Jensen's inequality:

$$\mathbb{E}_{y \sim \pi_{\text{old}}(\cdot|x)}\left[\exp\left(\frac{A^\beta(x, y)}{\beta}\right)\right] \leq \mathbb{E}\left[\exp\left(\frac{\widehat{A}_i^\beta}{\beta}\right)\right] = 1 \implies \tau_s \leq 0.$$

In non-degenerate finite-G settings, the inequality is *strict*. Hence the target induced by this biased estimator lies in the unstable, negative-τ_s regime, even though the empirical group is normalized.

This explains why the implementation choice of OAPL deviates from the nominal theory. In the experimental setup, OAPL decouples the temperature β_1 used for regularization from the temperature β_2 used for advantage computation, choosing $\beta_2 \gg \beta_1$. This choice makes the advantage closer to a centered group advantage and pushes the induced target back toward the stable positive-τ_s regime. This is discussed more in Appendix B.1. This implicitly implements the pessimism needed to stabilize off-policy learning. Rather than relying on this heuristic and an additional hyperparameter, we propose to target the Lambert policy, which was demonstrated to be more stable in RL settings [13] and closer to a χ-square divergence, implementing the principle of pessimism in the face of uncertainty [14, 16].

4.3 LAMBERT POLICY AS A TARGET

The preceding analysis suggests a simple design principle: instead of trying to target the fragile exponential boundary $\tau_s = 0$, choose a positive Lambert multiplier by construction. Specifically, we want to be in the regime $\tau_s > 1$ to benefit from the induced pessimism. Recall that the Lambert target ratio is

$$\rho^\star(y \mid x) = \frac{\pi^\star(y \mid x)}{\pi_{\text{old}}(y \mid x)} = \frac{1}{\tau_s} W_0\left(\tau_s \exp\left(\frac{A(x,y)}{\beta}\right)\right),$$

where τ_s is chosen so that $\mathbb{E}_{y\sim\pi_{\text{old}}(\cdot|x)}\left[\rho^\star(y \mid x)\right] = 1$.

We give the simple sufficient condition on the advantage to obtain $\tau_s \geq 1$ in the following proposition.

Proposition 4.1 (Shifted mean baseline yields a pessimistic Lambert target)**.** *Fix a context x and let $\beta > 0$. For a sentence-level advantage $A(x, y)$, define the Lambert mass function*

$$M_A(\tau) = \mathbb{E}_{y\sim\pi_{\text{old}}(\cdot|x)}\left[\frac{1}{\tau} W_0\left(\tau \exp\left(\frac{A(x,y)}{\beta}\right)\right)\right].$$

Let τ_s be the positive normalization multiplier satisfying $M_A(\tau_s) = 1$. We have for any $\alpha \geq 0$:

$$\mathbb{E}_{y\sim\pi_{\text{old}}(\cdot|x)}[A(x,y)] \geq \beta \implies \tau_s \geq 1.$$

Consequently, the following choice of the advantage induces a Lambert multiplier satisfying $\tau_s \geq 1$.

$$\widehat{A}_i = r_i - \bar{r} + \beta,$$

5 EXPERIMENTS

Experimental Setup. We fine-tune Qwen3 models at different scales [17] on DeepScaleR [20] using AdamW [21]. We report evaluation reward as the average accuracy on a held-out mixture of mathematical reasoning benchmarks comprising Minerva [4], AIME2025, AMC23, and BRUMO [22]. We also track average token-level entropy to characterize policy concentration during training. Curves report means and standard errors across 3 independent random seeds. To isolate the role of advantage normalization, we compare nominal OAPL with our Lambert method using the same regularized, sentence-level log-ratio regression objective. The methods differ only in their advantage definitions:

$$\widehat{A}_i^{\text{OAPL}} = r_i - \beta \log\left(\frac{1}{G}\sum_{j=1}^{G} \exp(r_j/\beta)\right), \qquad \widehat{A}_i^{\text{Lambert}} = r_i - \bar{r} + \beta.$$

Here, nominal OAPL uses the same temperature for advantage computation and regularization, without temperature decoupling. The shifted-mean Lambert advantage instead targets the pessimistic regime by construction. Rollouts are generated by a lagged behavior policy π_{old}, while updates are applied to the current policy π_θ. The policy lag L measures the number of optimization steps separating these policies; larger lags therefore correspond to staler, more off-policy data. We additionally include GRPO with a lag of $L = 2$ as a near-on-policy reference, following [7].

Validation of the Target Regime. We first test the mechanism identified by our analysis. For a set of 20 prompts and different group sizes $G \in \{2, 4\}$, we average the group-based training advantages over sampled groups to estimate the corresponding population advantage, which we then use to estimate $\log Z_{\text{exp}}(x)$. This diagnostic concerns the population objective optimized by SGD, rather than the normalization within an individual sampled group: positive values identify the pessimistic regime, while negative values identify the unstable regime. Figure 1 shows a clear separation between the advantage definitions. OAPL produces negative estimates throughout, whereas Lambert advantage consistently places the target in the pessimistic regime. This separation holds for every evaluated prompt, not merely on average. This result supports the predicted mechanism: advantage normalization changes the population target induced by the objective.

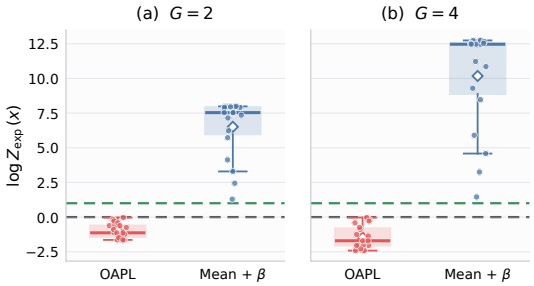


Figure 1: OAPL induces unstable targets, shifted-mean advantage induces pessimistic targets.

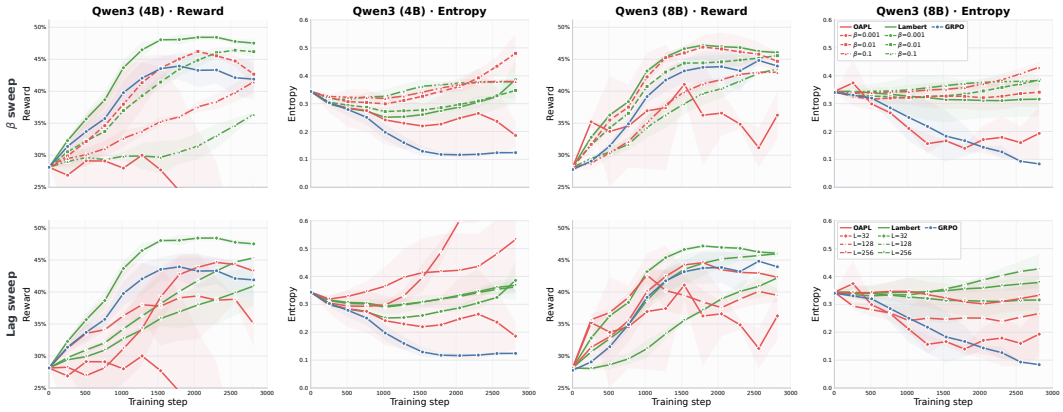


Figure 2: **Lambert improves performance and stability.** Evaluation reward and token-level entropy across model scales. Lambert remains stable under weak regularization (top) and stale rollouts (bottom), whereas OAPL exhibits unstable reward and entropy. GRPO rapidly loses entropy without matching Lambert's best reward.

Sensitivity to Regularization. We next examine whether the predicted difference in target regime translates into more stable learning. Figure 2 (top) compares training dynamics across regularization strengths at a fixed policy lag $L = 32$. Lambert remains stable across the tested settings, although stronger regularization can slow reward improvement. Under weaker regularization, it achieves high reward without rapid entropy loss, outperforming GRPO, which concentrates more quickly yet attains lower reward. In contrast, nominal OAPL exhibits erratic or deteriorating reward trajectories accompanied by pronounced entropy reduction. These observations are consistent with our analysis: by tempering high-advantage responses rather than amplifying them exponentially, the pessimistic Lambert target supports effective learning under weaker regularization without excessive policy concentration. Its benefit is therefore greater robustness in settings where nominal OAPL becomes unstable, rather than insensitivity to regularization strength.

Robustness to Stale Rollouts. Figure 2 (bottom) shows that Lambert continues to learn from increasingly stale rollouts while maintaining controlled entropy dynamics, although larger lags can slow progress. OAPL is substantially more sensitive, exhibiting unstable reward trajectories accompanied by either sharp entropy loss or excessive entropy growth, depending on the model and lag. These results highlight that avoiding entropy collapse alone is insufficient: uncontrolled entropy growth can also accompany poor optimization. Notably, even at $L = 128$, Lambert achieves higher final reward than near-on-policy GRPO while retaining substantially more entropy. Thus, learning with Lambert advantage, even from much staler rollouts can outperform a near-on-policy baseline, while avoiding the rapid policy concentration observed with GRPO.

Taken together, the diagnostic and training curves support our central claim: advantage normalization determines the induced population target, and targeting the pessimistic Lambert regime improves robustness to weak regularization and stale rollouts. Beyond stabilizing off-policy learning, Lambert can achieve higher reward than GRPO with substantially less entropy loss.

6 Conclusion

By characterizing regularized, ratio-free off-policy objectives as inducing Lambert-tempered policies, we showed that advantage normalization is not a benign baseline choice but a mechanism that determines whether the update is pessimistic, exponential, or unstable. This lens explains both the empirical strength and the fragility of optimal advantage regression methods: their nominal log-sum-exp advantage can induce an aggressive population target, while practical temperature decoupling implicitly restores a more conservative update. Building on this understanding, we proposed a shifted-mean advantage that targets the pessimistic Lambert regime by construction, improving robustness to regularization strength and policy lag in our experiments. This work sheds light on a mechanism underlying correction-free reinforcement learning for reasoning. We believe that a theoretical characterization of when retaining the correction is preferable to omitting it as an important direction for future research.

AI USE STATEMENT

In this work, we used generative AI tools to assist with scaling experiments, polishing the manuscript, and expanding details in proofs developed by the authors. We also used them to generate code for scientific figures and diagrams based on experimental results.

We did not use generative AI tools to generate synthetic datasets; develop theoretical models or conceptual frameworks; formulate mathematical claims or provide essential ingredients for their proofs; propose or refine hypotheses; translate text; clean or reformat datasets; analyze qualitative data; or interpret results.

All AI-assisted content was reviewed and verified by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies learning objectives for reinforcement learning in large language models. Its primary contribution is methodological: by characterizing and controlling the implicit pessimism induced by off-policy objectives, it aims to make RL training more stable and less prone to entropy collapse. More stable training may reduce computational waste and make post-training pipelines easier to reproduce and debug. At the same time, improvements in RL for reasoning can strengthen the capabilities of language models, which may have both beneficial and harmful downstream uses. Better reasoning systems can support education, scientific assistance, programming, and mathematical problem solving, but they may also improve the effectiveness of systems used for misinformation, cyber misuse, or other harmful applications. We therefore view the societal impact as indirect, mediated by how improved RL methods are incorporated into future language-model training pipelines. Responsible deployment should pair such optimization methods with standard safeguards, evaluation, and monitoring appropriate to the intended application.

REPRODUCIBILITY STATEMENT

We provide the complete source code implementing the proposed method and the experiments conducted in the supplementary materials for this submission. This is the simplest and most reliable way to reproduce the results. Detailed proofs of the theoretical statements presented in this article can be found in Appendices.

REFERENCES

[1] OpenAI. Openai o1 system card, 2026.

[2] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[3] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[4] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 3843–3857. Curran Associates, Inc., 2022.

[5] Kimi Team. Kimi k2: Open agentic intelligence, 2026.

[6] Michael Noukhovitch, Shengyi Huang, Sophie Xhonneux, Arian Hosseini, Rishabh Agarwal, and Aaron Courville. Faster, more efficient RLHF through off-policy asynchronous learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[7] Daniel Ritter, Owen Oertell, Bradley Guo, Jonathan Chang, Kianté Brantley, and Wen Sun. Llms can learn to reason from off-policy data, 2026.

[8] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1889–1897, Lille, France, 07–09 Jul 2015. PMLR.

[9] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[10] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[11] Kianté Brantley, Mingyu Chen, Zhaolin Gao, Jason D. Lee, Wen Sun, Wenhao Zhan, and Xuezhou Zhang. Accelerating RL for LLM reasoning with optimal advantage regression. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[12] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[13] Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D. Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J Foster. Correcting the mythos of KL-regularization: Direct alignment without overoptimization via chi-squared preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

[14] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[15] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[16] Otmane Sakhi, Imad Aouali, Pierre Alquier, and Nicolas Chopin. Logarithmic smoothing for pessimistic off-policy evaluation, selection and learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 80706–80755. Curran Associates, Inc., 2024.

[17] An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[18] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2021.

[19] Aleksei Arzhantsev, Otmane Sakhi, and Flavian Vasile. RoiRL: Efficient, self-supervised reasoning with offline iterative reinforcement learning. In *NeurIPS 2025 Workshop on Efficient Reasoning*, 2025.

[20] Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2, 2025. Notion Blog.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[22] Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvalddson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms. 2026.

A PROOFS

A.1 PROOF OF PROPOSITION 3.1

Proof. Fix x. For readability, we omit the conditioning on x when there is no ambiguity, and write

$$\rho(y) = \frac{\pi(y \mid x)}{\pi_{\text{old}}(y \mid x)}.$$

We assume $\pi_{\text{old}}(y \mid x) > 0$ on $\mathcal{Y}(x)$, or equivalently restrict $\mathcal{Y}(x)$ to the support of $\pi_{\text{old}}(\cdot \mid x)$. Since

$$\pi(y \mid x) = \pi_{\text{old}}(y \mid x)\rho(y),$$

the normalization constraint $\sum_y \pi(y \mid x) = 1$ becomes

$$\sum_{y \in \mathcal{Y}(x)} \pi_{\text{old}}(y \mid x)\rho(y) = 1.$$

Moreover,

$$\log \pi(y \mid x) = \log \pi_{\text{old}}(y \mid x) + \log \rho(y).$$

Therefore, up to a constant independent of ρ, the objective can be written as

$$\mathcal{J}_s(\rho) = \sum_{y \in \mathcal{Y}(x)} \pi_{\text{old}}(y \mid x) \left[A(x,y) \log \rho(y) - \frac{\beta}{2} (\log \rho(y))^2 \right].$$

Let $\pi^\star$ be an interior stationary point and let

$$\rho^\star(y) = \frac{\pi^\star(y \mid x)}{\pi_{\text{old}}(y \mid x)}.$$

Since the point is interior, $\rho^\star(y) > 0$ for all $y \in \mathcal{Y}(x)$. Introduce a Lagrange multiplier $\beta\tau_s$ for the normalization constraint and consider

$$\mathcal{L}(\rho, \tau_s) = \sum_y \pi_{\text{old}}(y \mid x) \left[A(x,y) \log \rho(y) - \frac{\beta}{2} (\log \rho(y))^2 \right] - \beta\tau_s \left(\sum_y \pi_{\text{old}}(y \mid x)\rho(y) - 1 \right).$$

The first-order condition with respect to $\rho(y)$ gives

$$\pi_{\text{old}}(y \mid x) \left[\frac{A(x,y)}{\rho(y)} - \frac{\beta \log \rho(y)}{\rho(y)} - \beta\tau_s \right] = 0.$$

Since $\pi_{\text{old}}(y \mid x) > 0$, this is equivalent to

$$A(x,y) - \beta \log \rho^\star(y) - \beta\tau_s\rho^\star(y) = 0.$$

Dividing by β, we obtain the stationarity equation

$$\log \rho^\star(y) + \tau_s\rho^\star(y) = \frac{A(x,y)}{\beta}.$$

Exponentiating both sides yields

$$\rho^\star(y) \exp\left(\tau_s\rho^\star(y)\right) = \exp\left(\frac{A(x,y)}{\beta}\right).$$

If $\tau_s > 0$, multiplying by τ_s gives

$$\tau_s\rho^\star(y) \exp\left(\tau_s\rho^\star(y)\right) = \tau_s \exp\left(\frac{A(x,y)}{\beta}\right).$$

Applying the principal branch W_0 of the Lambert W function, defined by $W_0(z)e^{W_0(z)} = z$, gives

$$\tau_s \rho^\star(y) = W_0\left(\tau_s \exp\left(\frac{A(x,y)}{\beta}\right)\right).$$

Thus

$$\rho^\star(y) = \frac{1}{\tau_s} W_0\left(\tau_s \exp\left(\frac{A(x,y)}{\beta}\right)\right).$$

Multiplying by $\pi_{\text{old}}(y \mid x)$, we obtain

$$\pi^\star(y \mid x) = \frac{\pi_{\text{old}}(y \mid x)}{\tau_s} W_0\left(\tau_s \exp\left(\frac{A(x,y)}{\beta}\right)\right).$$

The multiplier τ_s is chosen so that the distribution is normalized:

$$\sum_{y \in \mathcal{Y}(x)} \pi^\star(y \mid x) = 1.$$

Finally, consider the limit $\tau_s \to 0$. Since $W_0(z) = z + o(z)$ as $z \to 0$, we have

$$\frac{1}{\tau_s} W_0\left(\tau_s \exp\left(\frac{A(x,y)}{\beta}\right)\right) \longrightarrow \exp\left(\frac{A(x,y)}{\beta}\right).$$

Therefore the limiting density ratio is

$$\rho^\star(y) = \exp\left(\frac{A(x,y)}{\beta}\right),$$

and hence the limiting policy has the exponential-tilting form

$$\pi^\star(y \mid x) \propto \pi_{\text{old}}(y \mid x) \exp\left(\frac{A(x,y)}{\beta}\right).$$

This proves the claim. □

A.2 Proof of Proposition 4.1

Proof. Define

$$M(\tau) = \mathbb{E}_{y \sim \pi_{\text{old}}(\cdot|x)}\left[\frac{1}{\tau} W_0\left(\tau \exp\left(\frac{A(x,y)}{\beta}\right)\right)\right].$$

For $\tau > 0$, $M(\tau)$ is decreasing in τ. Therefore, to guarantee $\tau_s \geq 1$, it is enough to ensure

$$M(1) \geq 1,$$

that is,

$$\mathbb{E}_{y \sim \pi_{\text{old}}(\cdot|x)}\left[W_0\left(\exp\left(\frac{A(x,y)}{\beta}\right)\right)\right] \geq 1.$$

Let $r(y)$ be the raw reward and choose the shifted empirical mean baseline with G completions from π_{old}:

$$b(x) = \bar{r}_G - \beta.$$

Then

$$A(x,y) = r(y) - b(x) = r(y) - \mathbb{E}_{\pi_{\text{old}}}[r(y)] + \beta,$$

so

$$\mathbb{E}_{\pi_{\text{old}}}[A(x,y)] = \beta.$$

Now define

$$f(u) = W_0(e^u).$$

This function is convex, since if $w = f(u)$, then

$$f''(u) = \frac{w}{(1+w)^3} > 0.$$

Therefore, by Jensen's inequality,

$$\begin{aligned}\mathbb{E}_{\pi_{\text{old}}}\left[W_0\left(\exp\left(\frac{A(x,y)}{\beta}\right)\right)\right] &= \mathbb{E}_{\pi_{\text{old}}}\left[f\left(\frac{A(x,y)}{\beta}\right)\right] \\ &\geq f\left(\mathbb{E}_{\pi_{\text{old}}}\left[\frac{A(x,y)}{\beta}\right]\right) \\ &= f(1) = W_0(e) = 1.\end{aligned}$$

Hence $M(1) \geq 1$, and therefore

$$\tau_s \geq 1.$$

In practice, with a sampled group $r_1, \ldots, r_G$, we use

$$\bar{r} = \frac{1}{G}\sum_{i=1}^{G} r_i, \qquad \widehat{b}(x) = \bar{r} - \beta.$$

This gives the empirical advantage

$$\widehat{A}_i = r_i - \widehat{b}(x) = r_i - \bar{r} + \beta.$$

Since

$$\frac{1}{G}\sum_{i=1}^{G} \widehat{A}_i = \beta,$$

the same Jensen argument gives

$$\begin{aligned}\frac{1}{G}\sum_{i=1}^{G} W_0\left(\exp\left(\frac{\widehat{A}_i}{\beta}\right)\right) &\geq W_0\left(\exp\left(\frac{1}{G}\sum_{i=1}^{G}\frac{\widehat{A}_i}{\beta}\right)\right) \\ &= W_0(e) = 1.\end{aligned}$$

Thus the empirical target satisfies

$$\widehat{\tau}_s \geq 1.$$

For the population objective with the group baseline, the effective advantage is

$$A_G(x,y) = \mathbb{E}\left[r_i - \bar{r} + \beta \mid y_i = y, x\right] = \frac{G-1}{G}\left(r(y) - V_{\text{old}}(x)\right) + \beta,$$

where

$$V_{\text{old}}(x) = \mathbb{E}_{y\sim\pi_{\text{old}}(\cdot|x)}[r(y)].$$

This also satisfies

$$\mathbb{E}_{\pi_{\text{old}}}[A_G(x,y)] = \beta,$$

so the same Jensen argument guarantees that the induced population target satisfies

$$\tau_s \geq 1.$$

Therefore, the simple practical rule

$$\widehat{A}_i = r_i - \bar{r} + \beta$$

places the Lambert target in the desired pessimistic regime $\tau_s \geq 1$.

□

B Discussions

B.1 Large-β_2 limit of the OAPL advantage

In the experimental setup, OAPL decouples the temperature β_1 used for regularization from the temperature β_2 used for advantage computation, choosing $\beta_2 \gg \beta_1$. This choice makes the advantage closer to a centered group advantage and puts it in the nice regime. Indeed, letting

$$\bar{r} = \frac{1}{G}\sum_{j=1}^{G} r_j, \qquad \widehat{\text{Var}}_G(r) = \frac{1}{G}\sum_{j=1}^{G}(r_j - \bar{r})^2,$$

we have, for large β_2,

$$\begin{aligned}\widehat{A}_i^{\beta_2} &= r_i - \beta_2 \log\left(\frac{1}{G}\sum_{j=1}^{G}\exp(r_j/\beta_2)\right)\\ &= r_i - \left(\bar{r} + \frac{1}{2\beta_2}\widehat{\mathrm{Var}}_G(r) + O(\beta_2^{-2})\right)\\ &= r_i - \bar{r} - \frac{1}{2\beta_2}\widehat{\mathrm{Var}}_G(r) + O(\beta_2^{-2})\\ &\xrightarrow[\beta_2\to\infty]{} r_i - \bar{r}.\end{aligned}$$

Thus, as β_2 becomes large, the log-sum-exp centered advantage approaches the usual group-centered advantage.

At the population level, the limiting induced advantage is therefore

$$\begin{aligned}A_G^{\infty}(x,y) &= \mathbb{E}\left[r_i - \bar{r} \mid y_i = y, x\right]\\ &= r(y) - \frac{1}{G}\left(r(y) + (G-1)V_{\mathrm{old}}(x)\right)\\ &= \frac{G-1}{G}\left(r(y) - V_{\mathrm{old}}(x)\right),\end{aligned}$$

where

$$V_{\mathrm{old}}(x) = \mathbb{E}_{y\sim\pi_{\mathrm{old}}(\cdot|x)}[r(y)].$$

This induced advantage is centered under the behavior policy:

$$\begin{aligned}\mathbb{E}_{y\sim\pi_{\mathrm{old}}(\cdot|x)}\left[A_G^{\infty}(y \mid x)\right] &= \frac{G-1}{G}\mathbb{E}_{y\sim\pi_{\mathrm{old}}(\cdot|x)}\left[r(y) - V_{\mathrm{old}}(x)\right]\\ &= 0.\end{aligned}$$

Consequently, at the regularization temperature β_1,

$$\begin{aligned}Z_G^{\infty}(x) &:= \mathbb{E}_{y\sim\pi_{\mathrm{old}}(\cdot|x)}\left[\exp\left(\frac{A_G^{\infty}(x,y)}{\beta_1}\right)\right]\\ &\geq \exp\left(\frac{\mathbb{E}_{y\sim\pi_{\mathrm{old}}(\cdot|x)}[A_G^{\infty}(x,y)]}{\beta_1}\right)\\ &= 1,\end{aligned}$$

with strict inequality whenever the rewards are non-degenerate. Hence the large-β_2 limit lies in the positive-τ_s regime. By continuity, choosing $\beta_2 \gg \beta_1$ pushes the finite-β_2 target back toward this positive-τ_s regime. This implicitly implements the pessimism needed to stabilize off-policy learning.