UNDERSTANDING OFF- VS ON-POLICY DISTILLATION: A TALE OF DISTINCT TRAINING OBJECTIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) learns from teacher feedback on student-generated responses and has shown promise in reducing forgetting relative to supervised fine-tuning (SFT). However, its benefits and fragility remain incompletely understood. We study sequential distillation from multiple teachers, where the student minimizes its average divergence from the teachers. Forward Kullback–Leibler (KL) divergence yields a weighted arithmetic mixture, while reverse KL yields a normalized weighted geometric aggregate. We develop algorithms that learn these targets under off-policy and on-policy feedback, respectively, establishing logarithmic regret bounds in the tabular setting and extending the analysis to function approximation. By analyzing these aggregation targets, we identify mechanisms that help explain both the benefits and fragility of OPD. Relative to forward KL, reverse KL can better retain a confident expert's preferences under uninformative feedback, but is more sensitive to teachers that assign very low probabilities to correct responses. Its token-level conditionals also reveal a dependence on continuation distributions that can favor incorrect prefixes over long horizons.

1 INTRODUCTION

Knowledge distillation transfers capabilities from teacher models to a student language model. A conventional approach is supervised fine-tuning (SFT) on teacher-generated responses, where the student learns to predict tokens along trajectories supplied by the teacher. On-policy distillation (OPD) (Agarwal et al., 2024; Lu & Lab, 2025) instead samples responses from the current student and obtains teacher logits along the resulting trajectories. This allows the student to learn from teacher feedback on its own behavior, an approach that has been adopted in the post-training of large language models (Yang et al., 2025; Xiao et al., 2026; Zeng et al., 2026). Related studies show that on-policy reinforcement learning can reduce forgetting relative to SFT (Shenfeld et al., 2026b; Chen et al., 2025) and, in some settings, recover capabilities lost during SFT (Jin et al., 2025).

Despite these advantages, OPD can also be fragile. It may fail under teacher–student mismatch and suffer from training instability over long generation horizons (Li et al., 2026). However, a theoretical understanding of when OPD offers advantages over off-policy methods such as SFT and what limits those advantages remains to be established. Recent work has begun to clarify the conditions under which these advantages arise. For example, Sriraman et al. (2026) show that in a noisy-expert model, matching the clean expert's return can require exponentially many offline samples in the horizon, whereas an on-policy algorithm achieves polynomial dependence when the corruption is known. Viano et al. (2026) show that on-policy value-based imitation can efficiently match expert's return under realizability of its action-value function, without requiring policy realizability. Together, these works establish conditions under which on-policy interaction offers advantages over offline learning. However, they do not directly explain the fragility observed in practical OPD, leaving a gap in our understanding of when and why it succeeds or fails.

This naturally raises a question: *What are the pros and cons caused by the distinction between off-policy and off-policy feedback in knowledge distillation?*

In the first step to answering this question, we consider the setting where the student distills sequential data from multiple teachers, and the goal of the student is to minimize its average divergence from all teachers. Within this framework, we connect off- and on-policy distillation to forward- and reverse-Kullback–Leibler (KL) objectives, respectively. We then derive closed-form solutions for both objectives, offering a new perspective on the differences between SFT and OPD through the aggregation targets they induce. Our main contributions are listed as follows:

- We propose a sequential distillation framework in which the student minimizes its average divergence from all teachers, and characterize the resulting aggregation targets under forward and

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.

reverse KL divergence. Specifically, forward KL yields a weighted **arithmetic** mixture of the teacher distributions, while reverse KL yields a normalized weighted **geometric** mean. We also derive the token-level conditionals of both targets for autoregressive generation.

- We develop algorithms that learn the forward-KL target from off-policy data collection and the reverse-KL target from teacher logits on student-generated responses. In the tabular setting, we establish $\widetilde{O}(\log T)$ regret bounds for both protocols, with regret defined as the cumulative divergence between the student's policy at each round and the corresponding aggregation target. We further extend the logarithmic regret to the setting with general function approximation. As a result, these guarantees connect off-policy distillation to the forward-KL aggregation objective and on-policy distillation to the reverse-KL aggregation objective.
- We compare the two targets to examine the benefits and fragility of reverse-KL aggregation relative to forward-KL aggregation. First, we consider a setting with one confident expert teacher and several uninformative teachers and show that reverse-KL aggregation can retain a higher probability of the correct response than forward-KL aggregation. We also show that a misleading teacher can suppress correct responses under reverse KL, whereas forward KL preserves the expert's weighted contribution. Finally, by examining the token-level conditionals, we show that over long horizons, reverse-KL aggregation can favor an incorrect prefix even while preserving the expert's ranking of complete responses. These results offer possible explanations for the benefits and fragility of OPD observed in practice.

Notation. For a positive integer n, let $[n] := \{1, \dots, n\}$. For a finite set $\mathcal{S}$, let $\Delta(\mathcal{S})$ denote the set of probability distributions on $\mathcal{S}$. We write $\mathbb{1}(E)$ for the indicator of an event E, and use $\mathbb{E}$, $\mathbb{P}$, and Var for expectation, probability, and variance, respectively. For distributions $p, q \in \Delta(\mathcal{S})$, the Kullback–Leibler divergence is $\mathrm{KL}(p\|q) := \sum_{z\in\mathcal{S}} p(z)\log[p(z)/q(z)]$, where terms with $p(z) = 0$ are zero, and $\mathrm{KL}(p\|q) = +\infty$ if $p(z) > 0 = q(z)$ for some z. We use $O(\cdot)$ for asymptotic upper bounds and $\widetilde{O}(\cdot)$ to hide logarithmic factors except $\log T$. The notation $a \lesssim b$ means $a \le Cb$ for a universal constant $C > 0$.

2 PRELIMINARIES

We study sequential distillation, in which a single student learns from multiple teachers over successive training rounds with the goal of aggregating their knowledge.

Let $\mathcal{X}$ be a finite context space of size $S := |\mathcal{X}|$, $\mathcal{A}$ a finite token space of size $A := |\mathcal{A}|$, and H the generation horizon. We write $\mathcal{Y} := \mathcal{A}^H$ for the response space. Each response $y = (a_1, \dots, a_H) \in \mathcal{Y}$ is generated autoregressively, and we denote its length-$(h-1)$ prefix by $y_{<h} := (a_1, \dots, a_{h-1})$. We consider a sequential distillation setting with a finite collection of teacher models indexed by $\mathcal{I}$, where $|\mathcal{I}| = I$. For each $i \in \mathcal{I}$, let $\rho_i \in \Delta(\mathcal{X})$ be its context distribution and let p_i be an autoregressive teacher policy such that for any $h \in [H]$, the token conditionals are written as

$$p_i(\cdot \mid x, y_{<h}) \in \Delta(\mathcal{A}).$$

The induced sequence-level policy is

$$p_i(y \mid x) = \prod_{h=1}^{H} p_i(a_h \mid x, y_{<h}).$$

At round t, the student policy π_t is determined by the observations from previous rounds. A latent teacher index i_t is then drawn uniformly from $\mathcal{I}$, independently of previous rounds. Conditional on i_t, we draw m contexts independently from ρ_{i_t}. The entire batch shares this teacher: teacher responses and feedback are generated according to p_{i_t}. We consider two distillation protocols:

- (Off-policy Distillation) We sample m contexts $\{x_{t,j}\} \sim \rho_{i_t}$ and full teacher rollouts $y_{t,j} \sim p_{i_t}(\cdot|x_{t,j})$. If we further have access to the teacher logits, at each teacher-generated prefix $y_{t,j,<h}$, we additionally observe the full logit vector $Z_{t,j,h} \in \mathbb{R}^{\mathcal{A}}$. The corresponding token log-probabilities are obtained by applying log-softmax:

$$\log p_{i_t}(a \mid x_{t,j}, y_{t,j,<h}) = Z_{t,j,h}(a) - \log \sum_{b\in\mathcal{A}} e^{Z_{t,j,h}(b)}, \qquad a \in \mathcal{A}.$$

- (On-policy Distillation) We sample m contexts $\{x_{t,j}\} \sim \rho_{i_t}$, use the current student policy π_t to roll out with $y_{t,j} \sim \pi_t(\cdot|x_{t,j})$ token by token, and receive the per-token teacher log-probabilities

$$l_{t,j,h} = \log p_{i_t}(a_{t,j,h} \,|\, x_{t,j},\, y_{t,j,<h}), \qquad h \in [H].$$

Rather than matching any single teacher selected in the current round, our goal is to learn a single autoregressive student policy that aggregates the behavior of all teacher models. This objective is designed to avoid catastrophic forgetting: sequentially adapting the student to the current teacher may reduce its loss on the corresponding task while degrading its performance on teachers encountered previously. To formalize the desired aggregate policy, let

$$D : \Delta(\mathcal{Y}) \times \Delta(\mathcal{Y}) \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$$

be a divergence between distributions. Given the collection of teacher models $\{(\rho_i, p_i)\}_{i\in\mathcal{I}}$, we define the target policy by

$$\pi^* \in \operatorname*{argmin}_{\pi:\mathcal{X}\to\Delta(\mathcal{Y})} \sum_{i\in\mathcal{I}} \mathbb{E}_{x\sim\rho_i}\big[D\big(\pi(\cdot \mid x) \,\big\|\, p_i(\cdot \mid x)\big)\big]. \tag{2.1}$$

Here, $\pi(\cdot \mid x)$ and $p_i(\cdot \mid x)$ denote the induced distributions over complete responses in $\mathcal{Y} = \mathcal{A}^H$.

Remark 2.1. The initial student policy can also serve as a reference during training. In reinforcement learning from human feedback (RLHF), for example, the updated policy is often regularized toward its initial policy to limit how far it departs from that reference. To capture this role in our framework, we regard the reference policy as an additional teacher and include it in the teacher set as a modeling simplification. Under this interpretation, even distillation from a single external teacher can be viewed as learning from both that teacher and the initial student.

Specifically, when D is the reverse KL or forward KL divergence, the objective can be decomposed into a sum of token-level divergences using the chain rule. In this work, we will mainly focus on the cases of forward KL ($D(p\|q) = \mathrm{KL}(q\|p)$) and reverse KL ($D(p\|q) = \mathrm{KL}(p\|q)$).

Regret. Since teacher indices are sampled uniformly, the marginal context distribution is the uniform mixture of the teacher context distributions:

$$\bar{\rho}(x) := \frac{1}{I}\sum_{i\in\mathcal{I}} \rho_i(x).$$

For each choice of D, let π^* denote the corresponding minimizer of (2.1). We define regret as the cumulative divergence of the student policies $\{\pi_t\}_{t=1}^T$ from this target, averaged over the marginal context distribution:

$$\mathrm{Regret}(T) := \sum_{t=1}^{T} \mathbb{E}_{x\sim\bar{\rho}}\left[D\big(\pi_t(\cdot \mid x) \,\big\|\, \pi^*(\cdot \mid x)\big)\right]. \tag{2.2}$$

SFT has a well-known forward-KL interpretation, while reverse KL is commonly used in OPD. Motivated by these connections, we take $D(p\|q) = \mathrm{KL}(q\|p)$ for off-policy distillation and $D(p\|q) = \mathrm{KL}(p\|q)$ for on-policy distillation.

3 Off-policy Distillation & Forward KL

In this section, we study how off-policy feedback can be used to learn the forward-KL aggregate. We first characterize this target, then construct a tabular algorithm using teacher-generated responses, with or without full teacher logits, and establish logarithmic regret bounds. As the first step, the following theorem gives the target distribution and its token conditionals.

Theorem 3.1 (Forward KL)**.** If $D(p\|q) = \mathrm{KL}(q\|p)$, a minimizer of (2.1) is given at the sequence level, for each x with $\bar{\rho}(x) > 0$, by

$$\pi^*_{\mathrm{F}}(y|x) = \sum_{i\in\mathcal{I}} w_i(x) p_i(y|x), \tag{3.1}$$

where $w_i(x) = \rho_i(x)/[\sum_{j\in\mathcal{I}} \rho_j(x)]$. Moreover, for any $h \in [H]$, supposing $\sum_{j\in\mathcal{I}} \rho_j(x) p_j(y_{<h} \mid x) > 0$, the token-level distributions are given by

$$\pi^*_{\mathrm{F}}(a|x, y_{<h}) = \sum_{i\in\mathcal{I}} w_i(x, y_{<h})\, p_i(a|x, y_{<h}), \qquad w_i(x, y_{<h}) := \frac{\rho_i(x)\, p_i(y_{<h}|x)}{\sum_{j\in\mathcal{I}} \rho_j(x)\, p_j(y_{<h}|x)}. \tag{3.2}$$

Remark 3.2. The prefix-dependent weights admit a Bayesian interpretation. Under the off-policy sampling protocol, the teacher index is first sampled uniformly, the context is drawn from ρ_{i_t}, and

the response is generated by p_{i_t}. Therefore, for every prefix satisfying $\sum_{j\in\mathcal{I}} \rho_j(x)p_j(y_{<h} \mid x) > 0$, Bayes' rule gives

$$w_i(x, y_{<h}) = \mathbb{P}\big(i_t = i \big| x_{t,j} = x, y_{t,j,<h} = y_{<h}\big).$$

Thus, $w_i(x, y_{<h})$ is the posterior probability that teacher i generated the observed context and prefix. Consequently, teachers that better explain the observed prefix receive greater weight in the target policy's next-token distribution.

3.1 Algorithm Design

The posterior weights above also determine the next-token distribution in teacher-generated data. For a teacher-generated rollout $(x_{t,j}, y_{t,j})$, let $u \in \mathcal{A}^{h-1}$ denote a prefix. Conditioning on $(x_{t,j}, y_{t,j,<h}) = (x, u)$ gives teacher weights $w_i(x, u)$. Hence, the next-token distribution is

$$\mathbb{P}\big(a_{t,j,h} = a \mid x_{t,j} = x,\, y_{t,j,<h} = u\big) = \sum_{i\in\mathcal{I}} w_i(x, u)p_i(a|x, u) = \pi^*_{\mathrm{F}}(a|x, u).$$

Thus, at every visited prefix, the observed next token is a sample from the corresponding token conditional of the forward-KL target, which suggests estimating each target conditional from the token frequencies observed at the corresponding context–prefix pair. When full logits are available, each sampled token can instead be replaced by the teacher's conditional probability vector. To express both updates in a common form, define

$$q_{t,j,h}(a) := \begin{cases} \mathbb{1}(a_{t,j,h} = a) & \text{w/o logit,} \\ \frac{\exp(Z_{t,j,h}(a))}{\sum_{b\in\mathcal{A}} \exp(Z_{t,j,h}(b))}, & \text{w/ logit.} \end{cases}$$

We then define the per-round normalized counts as

$$c_t(x, u, a) := \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}(x_{t,j} = x,\, y_{t,j,<h} = u)\, q_{t,j,h}(a), \tag{3.3}$$

Here, the token position $h = |u| + 1$ is determined by the prefix u and is hidden from the notation $c_t(x, u, a)$. The algorithm accumulates these counts after each round according to

$$W_t(x, u, a) := \sum_{s=1}^{t} c_s(x, u, a),\; W_t(x, u) := \sum_{a\in\mathcal{A}} W_t(x, u, a),$$

and outputs the conditional policy

$$\widehat{\pi}_{t+1}(a|x, u) := \frac{W_t(x, u, a) + 1/2}{W_t(x, u) + A/2}. \tag{3.4}$$

Here we add a $(1/2)$-smoothing to each accumulative weight, following the spirit of the classical Krichevsky–Trofimov estimator (Krichevsky & Trofimov, 1981), which prevents zero predicted probabilities and assigns the uniform distribution to unvisited states.

Algorithm 1 Per-Prefix Forward KL

1: **Input:** Context set $\mathcal{X}$, token set $\mathcal{A}$, horizon H.
2: Initialize $W_0(x, u, a) = 0$; equivalently, $\widehat{\pi}_1(a|x, u) = 1/A$ at every active prefix.
3: **for** $t = 1, \ldots, T$ **do**
4: Observe the batch of teacher rollouts $\{(x_{t,j}, y_{t,j})\}_{j=1}^{m}$ and, when available, the teacher logits $\{Z_{t,j,h}\}_{j,h}$.
5: For every level-h state (x, u) and token a, set $c_t(x, u, a)$ as (3.3).
6: Update the weight $W_t(x, u, a) = W_{t-1}(x, u, a) + c_t(x, u, a)$ and $W_t(x, u) = \sum_a W_t(x, u, a)$.
7: Let $\widehat{\pi}_{t+1}(a|x, u)$ be defined in (3.4).
8: **end for**
9: Output $\{\widehat{\pi}_t\}_{t=1}^{T}$.

3.2 Theoretical Guarantees

We now bound the cumulative forward-KL divergence from the target π_{F}^* to the policies produced by Algorithm 1. The following theorem gives logarithmic regret bounds under both feedback protocols, with expectation and probability taken over the teachers and rollouts sampled during training.

Theorem 3.3. Let $T \geq 2$, and define the regret in (2.2) with $D(p\|q) = \mathrm{KL}(q\|p)$. Under the off-policy distillation protocol, both with and without access to full teacher logits, Algorithm 1 satisfies

$$\mathbb{E}\big[\,\mathrm{Regret}(T)\big] \leq 4SA^H \log T.$$

Moreover, for any $\delta \in (0,1)$, with probability at least $1-\delta$, Algorithm 1 satisfies

$$\mathrm{Regret}(T) \leq 8SA^H \log T + 4H\big(1+\log(2T+A)\big)\log\frac{1}{\delta}.$$

For fixed S, A, and H, Theorem 3.3 implies that the forward-KL divergence from π_{F}^* to $\widehat{\pi}_t$, averaged over contexts and training rounds, decays as $O(\log T/T)$ in expectation and with high probability. Thus, off-policy feedback allows the student to learn the forward-KL aggregate in this average-divergence sense. The factor SA^H reflects the size of the effective state–action space: at step h, each context–prefix pair $(x,u) \in \mathcal{X} \times \mathcal{A}^{h-1}$ acts as a distinct state, yielding SA^h state–action pairs. The exponential dependence on H comes from the growth of the effective state space itself and the bound scales linearly with the total number of state–action pairs across all steps.

4 On-policy Distillation & Reverse KL

In this section, we study how on-policy feedback can be used to learn the reverse-KL aggregate. With the same structure as the last section, we first characterize this target, then construct a tabular algorithm using teacher log-probabilities evaluated along student-generated responses. The following theorem gives the target distribution and its token conditionals.

Theorem 4.1 (Reverse KL)**.** Let $w_i(x) = \rho_i(x)/[\sum_{j\in\mathcal{I}}\rho_j(x)]$. Define the unnormalized token-level geometric mean

$$g_h(a|x,y_{<h}) := \prod_{i\in\mathcal{I}} p_i(a|x,y_{<h})^{w_i(x)}.$$

We define the value function iteratively backward as follows: for any $x \in \mathcal{X}$, $y \in \mathcal{A}^H$,

$$V_{H+1}(x,y) := 1, \qquad V_h(x,y_{<h}) := \sum_{a\in\mathcal{A}} g_h(a|x,y_{<h})\, V_{h+1}\big(x,(y_{<h},a)\big), \qquad h = H,\dots,1.$$

Then the sequence-level reverse-KL minimizer is unique and is given by

$$\pi_{\mathrm{R}}^*(y|x) = \prod_{i\in\mathcal{I}} p_i(y|x)^{w_i(x)}/V_1(x). \tag{4.1}$$

Moreover, we assume that at every prefix $y_{<h}$, $V_h(x,y_{<h}) > 0$. Then, the token-level distribution is unique, i.e.,

$$\pi_{\mathrm{R}}^*(a|x,y_{<h}) = g_h(a|x,y_{<h}) \cdot V_{h+1}\big(x,(y_{<h},a)\big)/V_h(x,y_{<h}). \tag{4.2}$$

Remark 4.2. In contrast with the last section, the teacher weights $w_i(x)$ depend only on the context. However, the token conditional is not obtained by normalizing g_h alone: each token score is also multiplied by V_{h+1}, which sums the products of subsequent scores over all possible continuations. To summarize, the target accounts for both the current token and the continuations that follow it.

4.1 Algorithm Design

To control the estimation error of teacher feedback, we assume that teacher token log-probabilities differ from those of a fixed reference policy by a bounded amount.

Assumption 4.3. Let $B > 0$. There exists a fixed full-support reference policy π_{ref} such that, for every $i \in \mathcal{I}$, $h \in [H]$, and $(x,u,a) \in \mathcal{X} \times \mathcal{A}^{h-1} \times \mathcal{A}$,

$$\big|\log[p_i(a|x,u)/\pi_{\mathrm{ref}}(a|x,u)]\big| \leq B.$$

No such assumption is required for our forward-KL analysis, since the feedback consists of token indicators or teacher probabilities, both bounded in $[0, 1]$. Here, we instead estimate teacher log-probabilities, which can be unbounded below. Assumption 4.3 imposes a two-sided bound on the token-level probability ratios between each teacher and the reference policy.

Let $\mathcal{F}_t$ denote the observations available before round t, so that π_t is $\mathcal{F}_t$-measurable. For a student-generated rollout $(x_{t,j}, y_{t,j})$, let $u \in \mathcal{A}^{h-1}$ denote a prefix. At any visited context–prefix–token triple (x, u, a), Bayes' rule gives

$$\mathbb{P}\big(i_t = i \mid \mathcal{F}_t,\, x_{t,j} = x,\, y_{t,j,<h} = u,\, a_{t,j,h} = a\big) = \frac{\rho_i(x)\pi_t(u|x)\pi_t(a|x,u)}{\sum_{k\in\mathcal{I}} \rho_k(x)\pi_t(u|x)\pi_t(a|x,u)} = w_i(x).$$

Given the context and past observations, the prefix and token are generated by the student policy, which is shared across teacher indices. Their probabilities therefore cancel, leaving the teacher weights unchanged. Recall that the observed feedback is $l_{t,j,h} = \log p_{i_t}(a|x,u)$. Its conditional mean satisfies

$$\mathbb{E}\big[l_{t,j,h} \,\big|\, \mathcal{F}_t,\, x_{t,j} = x,\, y_{t,j,<h} = u,\, a_{t,j,h} = a\big] = \sum_{i\in\mathcal{I}} w_i(x) \log p_i(a|x,u) = \log g_h(a|x,u).$$

Visits to (x, u, a) thus provide unbiased observations of the log-geometric score $\log g_h(a|x, u)$. Motivated by this, we estimate this score by averaging the feedback collected at the same triple. Define the cumulative visit count N_t, the cumulative feedback S_t, and the empirical mean $\bar{l}_t$ as

$$\begin{aligned}
N_t(x,u,a) &:= \sum_{s=1}^{t}\sum_{j=1}^{m} \mathbb{1}\big(x_{s,j} = x,\, y_{s,j,<h} = u,\, a_{s,j,h} = a\big),\\
S_t(x,u,a) &:= \sum_{s=1}^{t}\sum_{j=1}^{m} l_{s,j,h}\, \mathbb{1}\big(x_{s,j} = x,\, y_{s,j,<h} = u,\, a_{s,j,h} = a\big),\\
\bar{l}_t(x,u,a) &:= S_t(x,u,a)/\big(N_t(x,u,a) \vee 1\big).
\end{aligned} \tag{4.3}$$

To account for uncertainty in these empirical means, we construct optimistic estimates of the log-geometric scores. The following lemma bounds the estimation error at visited triples.

Lemma 4.4. Under Assumption 4.3, fix any $\delta \in (0, 1)$. With probability at least $1 - 2\delta$, the following inequality holds for every $t \in [T]$ and every (x, u, a) with $N_t(x, u, a) \geq 1$ simultaneously,

$$\big|\bar{l}_t(x,u,a) - \log g_h(a|x,u)\big| \leq \beta_t(x,u,a),$$

where $\beta_t(x,u,a) := \widetilde{O}\big(B\sqrt{mH/(N_t(x,u,a)\vee 1)} + BmH/(N_t(x,u,a)\vee 1)\big)$.

As a result, we can define the optimistic estimate

$$\widehat{l}_t(x,u,a) := \begin{cases} \bar{l}_t(x,u,a) + \min\big\{\beta_t(x,u,a), 2B\big\} & N_t(x,u,a) \geq 1,\\ \log \pi_{\text{ref}}(a|x,u) + B & N_t(x,u,a) = 0. \end{cases} \tag{4.4}$$

At visited triples, Lemma 4.4 ensures that $\widehat{l}_t(x, u, a) \geq \log g_h(a|x, u)$ on the confidence event. At unvisited triples, this inequality follows directly Assumption 4.3. We then substitute these optimistic scores into the backward recursion defining the reverse-KL target in (4.2):

$$\begin{aligned}
\widehat{g}_{t,h}(a|x,u) &:= \exp\big(\widehat{l}_t(x,u,a)\big),\ \widehat{V}_{t,H+1}(x,y) := 1,\\
\widehat{V}_{t,h}(x,u) &:= \sum_{a\in\mathcal{A}} \widehat{g}_{t,h}(a|x,u)\widehat{V}_{t,h+1}\big(x,(u,a)\big), \qquad h = H,\dots,1.
\end{aligned}$$

Therefore, for any h, any context x and prefix $y_{<h} = u$, the output policy for the next token is defined as

$$\widehat{\pi}_{t+1}(a|x,u) := \widehat{g}_{t,h}(a|x,u)\widehat{V}_{t,h+1}\big(x,(u,a)\big)/\widehat{V}_{t,h}(x,u). \tag{4.5}$$

The algorithm is outlined in Algorithm 2.

Algorithm 2 Per-Prefix Reverse KL

Input: Context set $\mathcal{X}$, token set $\mathcal{A}$, H, B, δ, and T.
Initialize $\widehat{\pi}_1(a|x,u) = 1/A$ for any (x,u,a).
for $t = 1, \dots, T$ **do**
  Observe the contexts $\{x_{t,j}\}_{j=1}^{m}$ generated from ρ_{i_t}.
  For each j, generate $y_{t,j} \sim \widehat{\pi}_t(\cdot|x_{t,j})$ and observe $\{l_{t,j,h}\}_{h=1}^{H}$.
  For each triple (x,u,a), update $N_t(x,u,a)$, $S_t(x,u,a)$, and $\widehat{l}_t(x,u,a)$ as (4.3).
  Construct the optimistic estimate $\widehat{l}_t$ according to (4.4).
  Define $\widehat{\pi}_{t+1}$ according to (4.5).
end for
Output $\{\widehat{\pi}_t\}_{t=1}^{T}$.

4.2 Theoretical Guarantees

We now bound the regret induced by Algorithm 2.

Theorem 4.5. Let $T \geq 2$, and define the regret in (2.2) with $D(p\|q) = \mathrm{KL}(p\|q)$. Under the on-policy distillation protocol and Assumption 4.3, for any $\delta \in (0, 1/3)$, with probability at least $1-3\delta$, Algorithm 2 satisfies

$$\mathrm{Regret}(T) \leq \widetilde{O}\big(B^2 H^2 S A^H \log T\big).$$

For fixed S, A, H, B, and δ, Theorem 4.5 implies that the reverse-KL divergence from $\widehat{\pi}_t$ to π^*_{R}, averaged over contexts and training rounds, decays as $\widetilde{O}(\log T/T)$ with high probability. Thus, on-policy feedback allows the student to learn the reverse-KL aggregate in this average-divergence sense. As in the off-policy setting, the factor SA^H reflects the size of the effective state–action space. For the detailed proof, see Appendix D.

5 Comparison of Distinct Aggregation Objectives

The preceding sections establish learning guarantees that connect off-policy distillation to forward-KL aggregation and on-policy distillation to reverse-KL aggregation. We now examine the differences between these two forms of aggregation and their implications for combining the capabilities of multiple teachers in a single policy.

5.1 Uninformative Teachers

Consider a context x that falls within teacher k's area of expertise but is unfamiliar to the remaining teachers. We represent this difference in coverage by assuming that $\rho_k(x)$ is large relative to $\sum_{i \neq k} \rho_i(x)$. The expert therefore receives a large aggregation weight $\alpha := w_k(x) \in (0.1)$.

Write $p(\cdot|x) := p_k(\cdot|x)$ and $N := |\mathcal{Y}| \geq 2$, and let $q := p(y^\star|x)$ denote the expert's probability of a desirable response $y^\star$. A value of q close to one means that the expert already produces the correct response with high probability. We ask whether the aggregate can maintain this high probability when the other teachers are uninformative. To represent the remaining teachers' lack of information at x, we model their response distributions as uniform: $p_i(y|x) = 1/N$ for every $i \neq k$ and $y \in \mathcal{Y}$. These teachers have total weight $1-\alpha$ and are equivalent, under either objective, to a single uniform teacher with this weight. The two aggregation targets are therefore

$$\pi^*_{\mathrm{F}}(y|x) = \alpha p(y|x) + \frac{1-\alpha}{N}, \quad \pi^*_{\mathrm{R}}(y|x) = \frac{p(y|x)^\alpha}{\sum_{z \in \mathcal{Y}} p(z|x)^\alpha}.$$

For a concrete comparison, suppose $y^\star$ is the only correct response and $q \in (1/N, 1)$, with the expert assigning uniform probability to each incorrect response. The following proposition shows that, for fixed teacher weights and response space, the comparison is determined by a confidence threshold.

Proposition 5.1. Fix $N \geq 2$ and $\alpha \in (0,1)$, and consider the teacher distributions above. If $N \geq 3$, there exists a unique $q_{\mathrm{c}} = q_{\mathrm{c}}(N,\alpha) \in (1/N, 1)$ such that, for every $q \in (1/N, 1)$,

$$\pi^*_{\mathrm{R}}(y^\star|x) > \pi^*_{\mathrm{F}}(y^\star|x) \quad \Longleftrightarrow \quad q > q_{\mathrm{c}}.$$

The proof is given in Appendix F. Figure 1(a) plots the two correct-response probabilities for $q > 0.5$, with $N = 10$ and $\alpha = 0.9$ fixed.

Remark 5.2. This setting illustrates how forgetting can arise when a student, after acquiring an expert teacher's capability, continues learning from other teachers that are uninformative on the

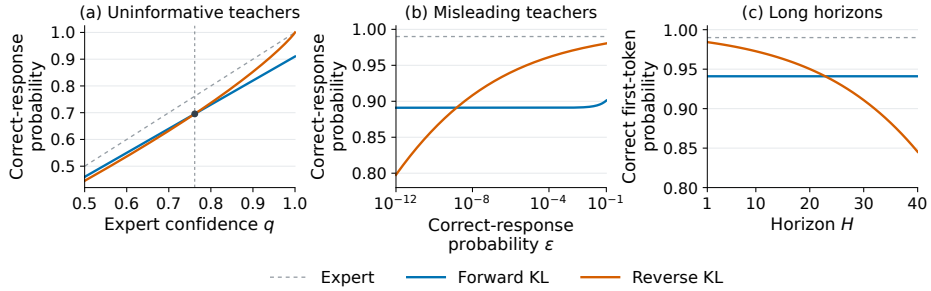

Figure 1: Forward- and reverse-KL aggregation with expert weight $\alpha = 0.9$. Panels (a,b) use $N = 10$ responses, with each teacher distributing its remaining probability uniformly over incorrect responses. (a) The other teachers are uniform, and expert confidence q varies above 0.5; the targets cross at $q_c \approx 0.7617$. (b) The expert's probability is fixed at $q = 0.99$, while the other teacher's correct-response probability ε varies on a logarithmic axis. (c) For binary responses, the horizon H varies while all token probabilities remain fixed. The expert chooses the correct first token a with probability 0.99; subsequent tokens have probabilities $(0.99, 0.01)$ after a and $(1/2, 1/2)$ after b. The other teacher is uniform at every prefix.

same context. Proposition 5.1 identifies when this reduction is smaller under reverse aggregation than under forward aggregation, offering a possible explanation for empirical findings that OPD is less susceptible to forgetting than SFT. For example, Shenfeld et al. (2026a) show in both single-task and sequential continual-learning experiments that OPD preserves general and previously acquired capabilities while learning new tasks, whereas SFT leads to substantial performance regression. Similar observations are reported by Hübotter et al. (2026) and Ma et al. (2026b), and have also been made in the field of vision language action model (Zhong et al., 2026).

5.2 Misleading Teachers

The preceding examples model teachers that are unfamiliar with a context as uniform. In this section, we consider another scenario where teachers may be misleading. As the small aggregation weights of some teachers reflect their limited coverage of the context, their predictions on such unfamiliar contexts may be highly inaccurate. Thus, it may assign extremely low probability to the correct response. We show that forward aggregation preserves a high probability of the correct response when the expert is highly confident and receives most of the weight. Under reverse aggregation, however, even a teacher with a small weight can drive this probability arbitrarily close to zero by assigning sufficiently low probability to the correct response.

Figure 1(b) illustrates this effect by varying only the misleading teacher's correct-response probability ε, while fixing the expert's probability at $q = 0.99$ and the teacher weights at 0.9 and 0.1. As ε decreases toward zero, the reverse target approaches zero, whereas the forward target remains bounded below by the expert's weighted contribution. The following proposition extends this comparison to a set of correct responses.

Proposition 5.3. Let $\varnothing \neq E \subsetneq \mathcal{Y}$ be a set of correct responses. For any teacher k, the forward target satisfies

$$\pi^*_{\mathrm{F}}(E|x) \geq w_k p_k(E|x).$$

In contrast, for any $\alpha, q, \eta \in (0, 1)$, there exist two teachers with weights $w_1 = \alpha$ and $w_2 = 1 - \alpha$, each assigning positive probability to every response, such that

$$p_1(E|x) = q, \qquad \pi^*_{\mathrm{R}}(E|x) < \eta.$$

Intuitively, forward aggregation is a weighted sum of nonnegative probabilities, so other teachers cannot cancel the expert's contribution to correct responses. Reverse aggregation instead averages log-probabilities, which can be arbitrarily negative. Consequently, a teacher that assigns sufficiently low probability to correct responses relative to incorrect ones can override the expert's preference, even with a small aggregation weight.

Remark 5.4. The sensitivity identified by Proposition 5.3 offers a possible explanation for some empirical failures of OPD under teacher–student mismatch. Such failures can occur even with a stronger teacher: Ma et al. (2026a) find that a stronger but distributionally different teacher produces predominantly negative feedback on student rollouts and degrades student performance. Related empirical remedies address either extreme feedback or the initial mismatch. Wang et al. (2026)

show that clipping or compressing extreme token-level log-ratios can improve training stability, while Li et al. (2026) find that an SFT cold start on teacher-generated responses reduces the initial pattern mismatch and improves subsequent OPD.

5.3 Long Horizons

The preceding comparisons focus on the probabilities of complete responses. During autoregressive generation, however, reverse aggregation can favor a prefix that the expert considers less likely, even while preserving the expert's ranking of complete responses. This can steer generation toward an incorrect trajectory.

Consider an expert teacher p_1 with weight $\alpha \in (0,1)$ and a uniform teacher p_2 with weight $\beta = 1-\alpha$. Since $p_2(y|x) = A^{-H}$ is the same for every complete response, its contribution to the geometric aggregate cancels upon normalization, leaving

$$\pi^*_{\mathrm{R}}(y|x) = p_1(y|x)^\alpha / \big[\textstyle\sum_{z\in\mathcal{Y}} p_1(z|x)^\alpha\big].$$

Raising probabilities to the power $\alpha < 1$ preserves the ranking of complete responses while increasing the relative weight of less likely responses. Token-level probabilities, however, also depend on a sum over possible continuations, as shown by (4.2). For a prefix u of length $h-1$ and a candidate token c,

$$\pi^*_{\mathrm{R}}(c|x,u) \propto p_1(c|x,u)^\alpha \sum_{v\in\mathcal{A}^{H-h}} p_1(v|x,u,c)^\alpha.$$

Since $\alpha < 1$, spreading probability mass more evenly across continuations increases this sum. The continuation factor can therefore favor tokens with more diffuse continuations over those whose continuations are concentrated on a few responses.

To further illustrate this effect, consider responses of length H over $\mathcal{A} = \{a, b\}$, where a is the correct first token. The expert p_1 has weight $\alpha = 0.9$ and selects a at the first step with probability $r = 0.99$. The other teacher p_2 has weight $\beta = 0.1$ and assigns probability $1/2$ to each token at every prefix.

At each of the remaining $H-1$ positions, the expert's token distribution depends only on the first token. After an initial a, it assigns probability 0.99 to a and 0.01 to b, regardless of the intervening tokens. After an initial b, it assigns probability $1/2$ to each token. Thus, the expert's continuations are concentrated after a and uniform after b, while all token probabilities remain strictly positive. Under forward aggregation, the probability of choosing a at the first step is the weighted average of the teachers' corresponding probabilities and is therefore independent of the continuation distributions and the horizon. Figure 1(c) plots the probability of choosing the correct first token a as H varies from 1 to 40, keeping the teacher weights and token-level probabilities fixed.

More generally, the following proposition shows that this reversal can occur even when the uniform teacher receives an arbitrarily small weight.

Proposition 5.5. Fix $\beta \in (0, 1/2)$ and $r \in (1/2, 1)$. For every sufficiently large horizon H, there exist two teacher policies p_1 and p_2 on $\mathcal{Y} = \{a, b\}^H$, each assigning positive probability to every response, such that $p_1(a|x) = r$ and p_2 is uniform at every prefix. With aggregation weights $w_1(x) = 1-\beta$ and $w_2(x) = \beta$, their forward and reverse targets satisfy

$$\pi^*_{\mathrm{F}}(a|x) > \pi^*_{\mathrm{F}}(b|x), \qquad \pi^*_{\mathrm{R}}(a|x) < \pi^*_{\mathrm{R}}(b|x).$$

Remark 5.6. The reversal in Proposition 5.5 illustrates a potential difficulty for OPD on long responses: increasing the horizon can change the reverse aggregation target's preferred prefix while the expert's preference remains fixed. Empirically, Li et al. (2026) find that longer rollouts can exhibit late-stage collapse, with instability emerging at later tokens and spreading toward earlier positions. Ziheng et al. (2026) find that teachers are less able to recover correct answers from longer student-generated prefixes and that stopping rollouts early can improve training stability.

6 Conclusion

We studied off- and on-policy distillation in a sequential multi-teacher framework. By characterizing the forward- and reverse-KL targets and establishing learning guarantees, we connected the two feedback protocols to distinct aggregation objectives. Our comparisons show that reverse-KL aggregation can better preserve a confident expert's preference in the presence of uniform teachers, but can suppress correct responses under misleading feedback or favor incorrect prefixes over long horizons. These results offer possible explanations for the benefits and fragility of OPD relative to SFT, providing a perspective of understanding the difference between SFT and OPD via the lens of distinct aggregation objectives.

AI USE STATEMENT

This work builds on the authors' research intuition and ideas. We used AI models, including GPT 5.6 Sol, GPT 6 Astra, Claude Fable 5, and Grok 4.6, to help articulate and summarize these ideas and to draft and polish the manuscript. These models also assisted in proving several key technical lemmas and developing proof strategies. The authors verified the proofs and reviewed the AI-assisted writing, and take full responsibility for the correctness and final content of the paper.

ETHICS STATEMENT

This work presents a theoretical analysis of knowledge distillation and does not involve human subjects or the collection of personal data. We do not identify specific ethical concerns arising directly from the analysis. Applications of knowledge distillation may nevertheless transfer biases or undesirable behaviors from teacher models to students; our theoretical guarantees concern distributional aggregation and learning, and do not establish the safety or fairness of the resulting models.

REPRODUCIBILITY STATEMENT

This work is theoretical. All assumptions and algorithms are stated explicitly, and complete proofs are provided in the appendices. All figures are deterministic evaluations of closed-form expressions, with parameter values given in the text and captions.

REFERENCES

Alekh Agarwal, Yujia Jin, and Tong Zhang. Vo q l: Towards optimal regret in model-free rl with nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 987–1063. PMLR, 2023.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Gholamali Aminian, Amir R Asadi, Idan Shenfeld, and Youssef Mroueh. Kl-regularized rlhf with multiple reference models: Exact solutions and sample complexity. *Advances in Neural Information Processing Systems*, 38:101117–101151, 2026.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Alireza Bakhtiari, Alex Ayoub, Samuel Robertson, David Janz, and Csaba Szepesvári. Eluder dimension: localise it! *arXiv preprint arXiv:2601.09825*, 2026.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. *arXiv preprint arXiv:2510.18874*, 2025.

Qiwei Di, Heyang Zhao, Jiafan He, and Quanquan Gu. Pessimistic nonlinear least-squares value iteration for offline reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4377–4410, 2024.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Dylan J Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. *Advances in Neural Information Processing Systems*, 37:120602–120666, 2024.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Haoyang Hong, Zichen Wang, Quanquan Gu, and Huazheng Wang. Online kl-regularized reinforcement learning with function approximation under misspecification. *arXiv preprint arXiv:2606.06053*, 2026.

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, et al. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe. *arXiv preprint arXiv:2605.03677*, 2026.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42217–42227, 2026.

Kaixuan Ji, Qiwei Di, Heyang Zhao, Qingyue Zhao, and Quanquan Gu. On the optimal sample complexity of offline multi-armed bandits with kl regularization. *arXiv preprint arXiv:2605.02141*, 2026a.

Kaixuan Ji, Qingyue Zhao, Heyang Zhao, Qiwei Di, and Quanquan Gu. Near-optimal regret for kl-regularized multi-armed bandits. *arXiv preprint arXiv:2603.02155*, 2026b.

Hangzhan Jin, Sitao Luan, Tianwei Ni, Sicheng Lyu, Guillaume Rabusseau, Reihaneh Rabbany, Doina Precup, and Mohammad Hamdaqa. Rl fine-tuning heals ood forgetting in sft. *arXiv preprint arXiv:2509.12235*, 2025.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. *arXiv preprint arXiv:2503.07067*, 2025.

Raphail Krichevsky and Victor Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Yifeng Liu, Shiyuan Zhang, Yifan Zhang, and Quanquan Gu. Self-distilled policy gradient. *arXiv preprint arXiv:2606.04036*, 2026.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. Demystifying opd: Length inflation and stabilization strategies for large language models. *arXiv preprint arXiv:2604.08527*, 2026.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, et al. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026a.

Yingzi Ma, Zichen Zhu, Ming Jiang, and Chaowei Xiao. One student, many teachers: Multi-task on-policy distillation via soft-prompt privileged context. *arXiv preprint arXiv:2607.18293*, 2026b.

Anupam Nayak, Tong Yang, Osman Yagan, Gauri Joshi, and Yuejie Chi. Achieving logarithmic regret in kl-regularized zero-sum markov games. *arXiv preprint arXiv:2510.13060*, 2025.

Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. Kl for a kl: On-policy distillation with control variate baseline. *arXiv preprint arXiv:2605.07865*, 2026.

Nived Rajaraman, Lin Yang, Jiantao Jiao, and Kannan Ramchandran. Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33:2914–2924, 2020.

Nived Rajaraman, Yanjun Han, Lin Yang, Jingbo Liu, Jiantao Jiao, and Kannan Ramchandran. On the value of interaction and function approximation in imitation learning. *Advances in Neural Information Processing Systems*, 34:1325–1336, 2021.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026a.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. Rl's razor: Why online reinforcement learning forgets less. In *International Conference on Learning Representations*, volume 2026, pp. 59839–59864, 2026b.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Ved Sriraman, Peihan Liu, Daniel Hsu, and Adam Block. Behavior cloning is not all you need: The optimality of on-policy distillation for noisy expert feedback. *arXiv preprint arXiv:2606.30923*, 2026.

Umar Syed and Robert Schapire. A reduction from apprenticeship learning to classification. *Advances in neural information processing systems*, 23, 2010.

Luca Viano, Antoine Moulin, Audrey Huang, Volkan Cevher, Philip Amortila, and Dylan J. Foster. When does on-policy interaction help? representational tradeoffs in value-based imitation learning. *arXiv preprint arXiv:2607.29617*, 2026.

Rui Wang, Hongru Wang, Yi Chen, Boyang Xue, Tianqing Fang, Wenhao Yu, and Kam-Fai Wong. Demystifying on-policy distillation: Roles, pathologies, and regulations. *arXiv preprint arXiv:2607.13399*, 2026.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020.

Di Wu, Chengshuai Shi, Jing Yang, and Cong Shen. Greedy sampling is provably efficient for rlhf. *Advances in Neural Information Processing Systems*, 38:108198–108232, 2026.

Yulian Wu, Rushil Thareja, Praneeth Vepakomma, and Francesco Orabona. Offline and online kl-regularized rlhf under differential privacy. *arXiv preprint arXiv:2510.13512*, 2025.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Zhengpeng Xie, Li Lyna Zhang, Zeke Xie, and Mao Yang. Trust region policy distillation. *arXiv preprint arXiv:2607.04751*, 2026.

Haoran Xin, Anhao Zhao, Ying Sun, Jin Li, Xiaoyu Shen, and Hui Xiong. Escaping the kl agreement trap in on-policy distillation. *arXiv preprint arXiv:2606.09471*, 2026.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, volume 2025, pp. 64616–64646, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D Lyng, Sanjit Singh Batra, and Robert E Tillman. Fast and effective on-policy distillation from reasoning prefixes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25553–25569, 2026.

Anhao Zhao, Junlong Tong, Yingqi Fan, Ping Nie, Wenjie Li, and Xiaoyu Shen. Poweropd: Stabilizing on-policy distillation with bounded power transformation. *arXiv preprint arXiv:2606.17199*, 2026a.

Heyang Zhao, Jiafan He, and Quanquan Gu. A nearly optimal and low-switching algorithm for reinforcement learning with general function approximation. *Advances in Neural Information Processing Systems*, 37:94684–94735, 2024.

Heyang Zhao, Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Logarithmic regret for online kl-regularized reinforcement learning. *arXiv preprint arXiv:2502.07460*, 2025.

Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for kl-regularized contextual bandits and rlhf. *Advances in Neural Information Processing Systems*, 38:107964–108002, 2026b.

Qingyue Zhao, Kaixuan Ji, Heyang Zhao, and Quanquan Gu. Fast rates for offline contextual bandits with forward-kl regularization under single-policy concentrability. *arXiv preprint arXiv:2605.09214*, 2026c.

Qingyue Zhao, Kaixuan Ji, Heyang Zhao, Tong Zhang, and Quanquan Gu. Towards a sharp analysis of offline policy learning for f-divergence-regularized contextual bandits. In *International Conference on Learning Representations*, volume 2026, pp. 117176–117210, 2026d.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026e.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. Scope: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting. *arXiv preprint arXiv:2604.10688*, 2026.

Zhide Zhong, Haodong Yan, Junfeng Li, Junjie He, Tianran Zhang, and Haoang Li. Vla-opd: Bridging offline sft and online rl for vision-language-action models via on-policy distillation. *arXiv preprint arXiv:2603.26666*, 2026.

Wenhong Zhu, Ruobing Xie, Rui Wang, and Pengfei Liu. Hybrid policy distillation for llms. *arXiv preprint arXiv:2604.20244*, 2026.

Zhou Ziheng, Jiaqi Li, Huacong Tang, Ying Nian Wu, and Demetri Terzopoulos. Less is more: Early stopping rollout for on-policy distillation. *arXiv preprint arXiv:2605.27028*, 2026.

A Related Work

Variants of on-policy distillation. On-policy distillation trains a language model by having it generate responses and using a teacher's next-token distributions to supervise those responses (Agarwal et al., 2024; Gu et al., 2024). Following the empirical success of OPD, subsequent work has explored improvements to its training objectives, the use of teacher feedback, and the construction of teachers. To improve the training objective, existing methods introduce skew KL or contrastive losses (Ko et al., 2024; 2025), combine forward and reverse KL using hybrid objectives or teacher–student agreement (Zhu et al., 2026; Jin et al., 2026; Xing et al., 2026), or construct intermediate distributions between the teacher and student as training targets (Jang et al., 2026; Xie et al., 2026). To make better use of teacher feedback, other methods use response outcomes to weight or calibrate supervision (Zheng et al., 2026; Hou et al., 2026), reduce variance or reshape distillation rewards (Oh et al., 2026; Zhao et al., 2026a), and stabilize training through policy constraints and mixed rollout sources (Luo et al., 2026). Data collection is also adapted through selective teacher intervention during student generation (Xu et al., 2025) and early termination of student rollouts to reduce generation costs and avoid unreliable teacher feedback (Ziheng et al., 2026; Xin et al., 2026; Zhang et al., 2026). Beyond these changes to objectives and training, the source of supervision has expanded from a separate teacher model to self-teachers conditioned on reasoning traces, demonstrations, additional context, or environmental feedback (Zhao et al., 2026e; Shenfeld et al., 2026a; Ye et al., 2026; Hübotter et al., 2026; Liu et al., 2026). Multiple teachers can also supervise a shared student, whether they are independently trained domain experts (Ma et al., 2026a) or constructed through task-specific soft prompts (Ma et al., 2026b). These extensions motivate studying how distillation combines knowledge from different teachers. We refer readers to Song & Zheng (2026) for a comprehensive survey of OPD methods.

Understanding on-policy distillation. Recent studies examine both the advantages and the failure modes of on-policy distillation. For capability retention during RL fine-tuning, Shenfeld et al. (2026b) relate reduced forgetting to an implicit bias toward solutions close to the initial policy in KL divergence, while Chen et al. (2025) emphasize on-policy data and mode-seeking behavior. Empirical studies also show that stronger teachers and denser supervision do not necessarily yield better distillation. Li et al. (2026) find that OPD gains depend on compatible reasoning patterns and the teacher offering capabilities beyond those already acquired by the student. They further observe that teacher guidance becomes less effective on longer student-generated prefixes, with training instability emerging at later tokens and spreading to earlier positions. Wang et al. (2026) identify settings where teacher–student mismatch causes incorrect responses to receive higher average token-level rewards than correct ones. Fu et al. (2026) identify three sources of instability in sampled-token OPD: predominantly negative token-level rewards that make learning sensitive to a small set of positively rewarded tokens; unreliable teacher feedback that can reinforce repetitive or meaningless continuations on student-generated prefixes; and tokenizer or special-token mismatches that penalize semantically valid outputs. The theoretically understanding of OPD is relatively limited. Sriraman et al. (2026) establish an exponential separation in horizon dependence between offline learning and an on-policy algorithm with known expert corruption, and Viano et al. (2026) obtain efficient imitation under expert action-value realizability without policy realizability. Our contribution is to characterize their distinct optimal targets when aggregating multiple teachers and connect those targets to learning guarantees. We then analyze how teacher confidence, misleading feedback, and autoregressive continuations affect the expert's preferences at these targets, providing possible mechanisms for both capability retention and fragility.

Relevant RL theory. To turn our perspective on distinct aggregation objectives into formal learning guarantees, we rely on the established RL theory, including supervised learning, KL-regularized bandits, and reinforcement learning with function approximation. Our analysis on the off-policy distillation protocol builds on the classical theory of behavior cloning, which aims to learn a policy from expert demonstrations. Classical analyses characterize how prediction errors compound under distribution shift (Syed & Schapire, 2010; Ross & Bagnell, 2010). DAgger addresses this issue by collecting expert feedback on learner-visited states, reducing imitation learning to no-regret online learning (Ross et al., 2011). Subsequent work further characterizes the statistical limits of imitation learning and the benefits of interaction (Rajaraman et al., 2020; 2021). More recently, Foster et al. (2024) establish guarantees for behavior cloning with logarithmic loss under realizability, highlighting the importance of the loss function and policy-class complexity. Our forward-KL guarantees draw on log-loss prediction methods, combining smoothed probability estimation in the

tabular setting and exponentially weighted policy aggregation under function approximation with a new one-sided bounded martingale concentration inequality.

For on-policy distillation, our analysis draws on prior work establishing logarithmic regret for KL-regularized bandits and reinforcement learning. Zhao et al. (2026b) establish $\mathcal{O}(1/\epsilon)$ sample complexity in a sufficiently small-error regime under reference-policy coverage, and later Zhao et al. (2025) extend the results to obtain logarithmic regret in online contextual bandits and RL. For multi-armed bandits, Ji et al. (2026b) sharpen the dependence on the number of arms and regularization strength through nearly matching regret upper and lower bounds. Similar fast-rate guarantees have also been established in the offline setting for multi-armed bandit (Ji et al., 2026a), or with sufficient coverage conditions under KL and more general divergences. (Zhao et al., 2026d;c). This line of work has since been extended to multiple reference models (Aminian et al., 2026), general preference models (Wu et al., 2026), game-theoretic settings (Nayak et al., 2025), differential privacy (Wu et al., 2025), and misspecified models (Hong et al., 2026). We adapt these techniques to sequential distillation by treating teacher log-probability ratios as rewards and accounting for dependence among rollouts from the same sampled teacher. This yields logarithmic regret guarantees for learning the reverse-KL aggregation target from on-policy teacher feedback.

While we have established the connection between off- and on-policy distillation and their respective aggregation targets in the tabular setting, we seek to extend this connection to more general settings, drawing inspiration from prior work on general function approximation. In contextual bandits, Russo & Van Roy (2013) introduce eluder dimension to quantify how past observations constrain predictions at unobserved actions. Subsequent work extends eluder-based analyses to model estimation through value-targeted regression (Ayoub et al., 2020) and general value-function approximation (Wang et al., 2020). We use a generalized notion of eluder dimension developed in later work (Agarwal et al., 2023; Zhao et al., 2024; Di et al., 2024) and adapt it to the autoregressive setting.

B Proof of the KL Aggregation Theorems

In this section, we prove Theorems 3.1 and 4.1. We first reduce the objective to a separate optimization problem at each context. Recall that

$$\bar{\rho}(x) := \frac{1}{I}\sum_{i\in\mathcal{I}} \rho_i(x), \qquad w_i(x) := \frac{\rho_i(x)}{\sum_{j\in\mathcal{I}} \rho_j(x)}$$

for every x with $\bar{\rho}(x) > 0$. Define

$$\mathcal{I}_x^+ := \{i \in \mathcal{I} : w_i(x) > 0\},\ L_{D,x}(\pi) := \sum_{i\in\mathcal{I}_x^+} w_i(x) D\big(\pi(\cdot \mid x) \,\big\|\, p_i(\cdot \mid x)\big).$$

Using $\rho_i(x) = I\bar{\rho}(x)w_i(x)$, we obtain

$$\sum_{i\in\mathcal{I}} \mathbb{E}_{x\sim\rho_i}\big[D\big(\pi(\cdot \mid x) \,\big\|\, p_i(\cdot \mid x)\big)\big] = I \sum_{x:\bar{\rho}(x)>0} \bar{\rho}(x) L_{D,x}(\pi). \tag{B.1}$$

The optimization domain in (2.1) allows a separate distribution in $\Delta(\mathcal{Y})$ at every context. Therefore, it suffices to minimize $L_{D,x}$ at each context with $\bar{\rho}(x) > 0$. Contexts with $\bar{\rho}(x) = 0$ do not affect the objective, and their policies may be chosen arbitrarily. Throughout the proofs, zero-weight teachers are omitted from geometric products. We use the conventions $0\log(0/q) = 0$ for $q \geq 0$ and $p\log(p/0) = +\infty$ for $p > 0$.

B.1 Proof of Theorem 3.1

Proof of Theorem 3.1. Fix a context x with $\bar{\rho}(x) > 0$. For forward KL, the contextwise objective in (B.1) is

$$L_{\mathrm{F},x}(\pi) := \sum_{i\in\mathcal{I}_x^+} w_i(x)\mathrm{KL}\big(p_i(\cdot \mid x) \,\big\|\, \pi(\cdot \mid x)\big).$$

We first identify its minimizing sequence distribution, and then derive the corresponding token conditionals.

Sequence-level minimizer. Define the candidate distribution

$$q_{\mathrm{F}}(y \mid x) := \sum_{i\in\mathcal{I}_x^+} w_i(x) p_i(y \mid x), \qquad y \in \mathcal{Y}.$$

Since the weights are nonnegative and sum to one, $q_{\mathrm{F}}(\cdot \mid x) \in \Delta(\mathcal{Y})$. Let

$$\mathcal{S}_{\mathrm{F}}(x) := \{y \in \mathcal{Y} : q_{\mathrm{F}}(y \mid x) > 0\},$$
$$C_{\mathrm{F}}(x) := \sum_{i \in \mathcal{I}_x^+} w_i(x) \mathrm{KL}\big(p_i(\cdot \mid x) \,\big\|\, q_{\mathrm{F}}(\cdot \mid x)\big).$$

For every $i \in \mathcal{I}_x^+$, we have $q_{\mathrm{F}}(y \mid x) \geq w_i(x) p_i(y \mid x)$. Consequently, $C_{\mathrm{F}}(x)$ is finite and does not depend on π. For a policy that is positive on $\mathcal{S}_{\mathrm{F}}(x)$, expanding the KL divergences gives

$$\begin{aligned} L_{\mathrm{F},x}(\pi) - C_{\mathrm{F}}(x) &= \sum_{i \in \mathcal{I}_x^+} w_i(x) \sum_{y \in \mathcal{S}_{\mathrm{F}}(x)} p_i(y \mid x) \log \frac{q_{\mathrm{F}}(y \mid x)}{\pi(y \mid x)} \\ &= \sum_{y \in \mathcal{S}_{\mathrm{F}}(x)} q_{\mathrm{F}}(y \mid x) \log \frac{q_{\mathrm{F}}(y \mid x)}{\pi(y \mid x)} \\ &= \mathrm{KL}\big(q_{\mathrm{F}}(\cdot \mid x) \,\big\|\, \pi(\cdot \mid x)\big), \end{aligned} \tag{B.2}$$

where the second equality uses the definition of q_{F}. If π assigns zero probability to some $y \in \mathcal{S}_{\mathrm{F}}(x)$, at least one positive-weight teacher assigns positive probability to that response. In that case, both sides of (B.2) are $+\infty$, so the identity remains valid in the extended-real sense.

By nonnegativity of KL divergence, the right-hand side of (B.2) is minimized at zero, with equality if and only if $\pi(\cdot \mid x) = q_{\mathrm{F}}(\cdot \mid x)$. Thus, the minimizing sequence distribution is unique at this context and satisfies

$$\pi_{\mathrm{F}}^*(y \mid x) = q_{\mathrm{F}}(y \mid x) = \sum_{i \in \mathcal{I}} w_i(x) p_i(y \mid x),$$

which proves (3.1).

Token-level conditionals. Fix $h \in [H]$ and $u \in \mathcal{A}^{h-1}$. Marginalizing the sequence distribution over all continuations gives

$$\begin{aligned} \pi_{\mathrm{F}}^*(u \mid x) &= \sum_{i \in \mathcal{I}} w_i(x) p_i(u \mid x), \\ \pi_{\mathrm{F}}^*((u,a) \mid x) &= \sum_{i \in \mathcal{I}} w_i(x) p_i(u \mid x) p_i(a \mid x, u). \end{aligned} \tag{B.3}$$

The first equality follows from linearity of marginalization; the second additionally uses the autoregressive factorization of each teacher. Suppose that $\sum_i \rho_i(x) p_i(u \mid x) > 0$. Then $\pi_{\mathrm{F}}^*(u \mid x) > 0$, and taking the ratio of the two prefix probabilities yields

$$\begin{aligned} \pi_{\mathrm{F}}^*(a \mid x, u) &= \frac{\pi_{\mathrm{F}}^*((u,a) \mid x)}{\pi_{\mathrm{F}}^*(u \mid x)} \\ &= \frac{\sum_{i \in \mathcal{I}} w_i(x) p_i(u \mid x) p_i(a \mid x, u)}{\sum_{j \in \mathcal{I}} w_j(x) p_j(u \mid x)} \\ &= \sum_{i \in \mathcal{I}} \frac{\rho_i(x) p_i(u \mid x)}{\sum_{j \in \mathcal{I}} \rho_j(x) p_j(u \mid x)} p_i(a \mid x, u), \end{aligned} \tag{B.4}$$

where the last equality substitutes the definition of $w_i(x)$ and cancels its common normalization factor. This is exactly (3.2). At a prefix with $\pi_{\mathrm{F}}^*(u \mid x) = 0$, the conditional may be chosen arbitrarily without changing the sequence distribution. Since x was arbitrary, combining the contextwise minimizers using (B.1) completes the proof. □

B.2 Proof of Theorem 4.1

Proof of Theorem 4.1. Fix a context x with $\bar{\rho}(x) > 0$. The contextwise reverse-KL objective is

$$L_{\mathrm{R},x}(\pi) := \sum_{i \in \mathcal{I}_x^+} w_i(x) \mathrm{KL}\big(\pi(\cdot \mid x) \,\big\|\, p_i(\cdot \mid x)\big).$$

Define the unnormalized sequence score and its total mass by

$$G_x(y) := \prod_{i \in \mathcal{I}_x^+} p_i(y \mid x)^{w_i(x)}, \qquad Z_x := \sum_{y \in \mathcal{Y}} G_x(y).$$

We first show that $Z_x = V_1(x)$, then identify the minimizing sequence distribution and its token conditionals.

Backward recursion and normalization. By the autoregressive factorization of the teachers,

$$\begin{aligned} G_x(y) &= \prod_{i \in \mathcal{I}_x^+} \left(\prod_{h=1}^{H} p_i(a_h \mid x, y_{<h}) \right)^{w_i(x)} \\ &= \prod_{h=1}^{H} \prod_{i \in \mathcal{I}_x^+} p_i(a_h \mid x, y_{<h})^{w_i(x)} \\ &= \prod_{h=1}^{H} g_h(a_h \mid x, y_{<h}). \end{aligned} \tag{B.5}$$

For every $h \in [H+1]$ and $u \in \mathcal{A}^{h-1}$, we claim that

$$V_h(x, u) = \sum_{a_h, \ldots, a_H \in \mathcal{A}} \prod_{k=h}^{H} g_k(a_k \mid x, y_{<k}), \qquad y = (u, a_h, \ldots, a_H). \tag{B.6}$$

At $h = H+1$, the sum contains one empty continuation and its empty product is one, so the identity agrees with $V_{H+1}(x, y) = 1$. Suppose that it holds at level $h+1$. Using the defining recursion for V_h, we obtain

$$\begin{aligned} V_h(x, u) &= \sum_{a_h \in \mathcal{A}} g_h(a_h \mid x, u) V_{h+1}(x, (u, a_h)) \\ &= \sum_{a_h \in \mathcal{A}} g_h(a_h \mid x, u) \sum_{a_{h+1}, \ldots, a_H \in \mathcal{A}} \prod_{k=h+1}^{H} g_k(a_k \mid x, y_{<k}) \\ &= \sum_{a_h, \ldots, a_H \in \mathcal{A}} \prod_{k=h}^{H} g_k(a_k \mid x, y_{<k}), \end{aligned}$$

where the second equality uses the induction hypothesis. This proves (B.6) by backward induction. At the empty prefix, it gives

$$V_1(x) = \sum_{y \in \mathcal{Y}} G_x(y) = Z_x. \tag{B.7}$$

In particular, the positivity condition $V_1(x) > 0$ implies $Z_x > 0$.

Sequence-level minimizer. Define

$$q_{\mathrm{R}}(y \mid x) := \frac{G_x(y)}{Z_x}, \qquad \mathcal{S}_{\mathrm{R}}(x) := \{y \in \mathcal{Y} : G_x(y) > 0\}.$$

By (B.7), $q_{\mathrm{R}}(\cdot \mid x)$ is a probability distribution. The set $\mathcal{S}_{\mathrm{R}}(x)$ is the common support of all positive-weight teachers. Any policy assigning positive mass outside this set has infinite reverse KL to at least one such teacher, and therefore incurs infinite objective value. For a policy supported on $\mathcal{S}_{\mathrm{R}}(x)$, we have

$$\begin{aligned} L_{\mathrm{R},x}(\pi) &= \sum_{y \in \mathcal{S}_{\mathrm{R}}(x)} \pi(y \mid x) \left[\log \pi(y \mid x) - \sum_{i \in \mathcal{I}_x^+} w_i(x) \log p_i(y \mid x) \right] \\ &= \sum_{y \in \mathcal{S}_{\mathrm{R}}(x)} \pi(y \mid x) \log \frac{\pi(y \mid x)}{G_x(y)} \\ &= \sum_{y \in \mathcal{S}_{\mathrm{R}}(x)} \pi(y \mid x) \log \frac{\pi(y \mid x)}{q_{\mathrm{R}}(y \mid x)} - \log Z_x \\ &= \mathrm{KL}\big(\pi(\cdot \mid x) \,\big\|\, q_{\mathrm{R}}(\cdot \mid x)\big) - \log V_1(x). \end{aligned} \tag{B.8}$$

Here, the second equality uses the definition of G_x, the third uses $G_x = Z_x q_{\mathrm{R}}$ and $\sum_y \pi(y \mid x) = 1$, and the last uses (B.7). The same identity holds in the extended-real sense when π puts positive mass outside $\mathcal{S}_{\mathrm{R}}(x)$, since both sides are then infinite.

Since $-\log V_1(x)$ does not depend on π, nonnegativity of KL divergence shows that the unique minimizing sequence distribution is

$$\pi^*_{\mathrm{R}}(y \mid x) = q_{\mathrm{R}}(y \mid x) = \frac{\prod_{i \in \mathcal{I}_x^+} p_i(y \mid x)^{w_i(x)}}{V_1(x)},$$

which proves (4.1).

Token-level conditionals. Fix $h \in [H]$ and a prefix $u = (u_1, \ldots, u_{h-1})$, and let

$$G_{<h}(x, u) := \prod_{k=1}^{h-1} g_k(u_k \mid x, u_{<k}),$$

where the empty product is one. Combining (B.5) and (B.6), and summing over all continuations, gives

$$\begin{aligned} \pi^*_{\mathrm{R}}(u \mid x) &= \frac{G_{<h}(x,u) V_h(x,u)}{V_1(x)}, \\ \pi^*_{\mathrm{R}}((u,a) \mid x) &= \frac{G_{<h}(x,u) g_h(a \mid x, u) V_{h+1}(x,(u,a))}{V_1(x)}. \end{aligned} \tag{B.9}$$

If $\pi^*_{\mathrm{R}}(u \mid x) > 0$, then both $G_{<h}(x,u)$ and $V_h(x,u)$ are positive. Taking the ratio of the two prefix probabilities yields

$$\begin{aligned} \pi^*_{\mathrm{R}}(a \mid x, u) &= \frac{\pi^*_{\mathrm{R}}((u,a) \mid x)}{\pi^*_{\mathrm{R}}(u \mid x)} \\ &= \frac{g_h(a \mid x, u) V_{h+1}(x,(u,a))}{V_h(x,u)}, \end{aligned} \tag{B.10}$$

which is (4.2). These conditionals are uniquely determined at every positive-probability prefix.

At a null prefix with $V_h(x,u) > 0$, the right-hand side of (B.10) still defines a valid conditional, because its numerator sums to $V_h(x,u)$. When $V_h(x,u) = 0$, choose an arbitrary conditional instead. To verify that these choices induce the minimizing sequence law, take any $y \in \mathcal{S}_{\mathrm{R}}(x)$. All factors and continuation values along this response are positive, so the selected conditionals telescope:

$$\begin{aligned} \prod_{h=1}^{H} \frac{g_h(a_h \mid x, y_{<h}) V_{h+1}(x, y_{\le h})}{V_h(x, y_{<h})} &= \frac{\prod_{h=1}^{H} g_h(a_h \mid x, y_{<h})}{V_1(x)} \\ &= \frac{G_x(y)}{Z_x} = \pi^*_{\mathrm{R}}(y \mid x), \end{aligned}$$

where $y_{\le h} := (a_1, \ldots, a_h)$ and the first equality uses $V_{H+1}(x,y) = 1$. These probabilities already sum to one over $\mathcal{S}_{\mathrm{R}}(x)$. Since every selected token conditional is normalized, the induced autoregressive law assigns zero mass to all other responses and equals $\pi^*_{\mathrm{R}}(\cdot \mid x)$. Thus, uniqueness concerns the sequence distribution and its conditionals at positive-probability prefixes, not arbitrary conditionals at null prefixes. Applying (B.1) across contexts completes the proof. □

C Proof of Theorem 3.3

Proof. In this section, we prove Theorem 3.3. Let $\mathcal{F}_t := \sigma\big(\{x_{s,j},\, y_{s,j},\, q_{s,j,h}(a) : 1 \le s < t,\, j \in [m],\, h \in [H],\, a \in \mathcal{A}\}\big)$. Then, $\widehat{\pi}_t$ is $\mathcal{F}_t$-measurable.

Fix a level h, a state $(x,u) \in \mathcal{X} \times \mathcal{A}^{h-1}$, and a token $a \in \mathcal{A}$. Under either feedback model,

$$\mathbb{E}\left[q_{t,j,h}(a) \,|\, \mathcal{F}_t,\, i_t = i,\, x_{t,j} = x,\, y_{t,j,<h} = u\right] = p_i(a|x,u). \tag{C.1}$$

Indeed, without logits, $q_{t,j,h}(a) = \mathbb{1}(a_{t,j,h} = a)$, whose conditional mean is $p_i(a|x,u)$. With logits, $q_{t,j,h}(a) = p_i(a|x,u)$ is observed directly.

Conditioned on $i_t = i$, using the independence of samples between each round, we have

$$\mathbb{P}\big(x_{t,j} = x,\, y_{t,j,<h} = u | i_t = i, \mathcal{F}_t\big) = \rho_i(x) p_i(u|x). \tag{C.2}$$

Combining (C.1) and (C.2), and using that i_t is uniform over $\mathcal{I}$, we have

$$\begin{aligned}
&\mathbb{E}\big[\,\mathbb{1}(x_{t,j}=x,\,y_{t,j,<h}=u)q_{t,j,h}(a)\,\big|\,\mathcal{F}_t\big]\\
&\quad=\sum_{i\in\mathcal{I}}\mathbb{P}(i_t=i\mid\mathcal{F}_t)\,\mathbb{E}\left[\mathbb{1}(x_{t,j}=x,\,y_{t,j,<h}=u)q_{t,j,h}(a)\,|\,\mathcal{F}_t,i_t=i\right]\\
&\quad=\frac{1}{I}\sum_{i\in\mathcal{I}}\mathbb{P}\big(x_{t,j}=x,\,y_{t,j,<h}=u\big|\mathcal{F}_t,i_t=i\big)\\
&\qquad\cdot\mathbb{E}\big[q_{t,j,h}(a)\big|\mathcal{F}_t,i_t=i,\,x_{t,j}=x,\,y_{t,j,<h}=u\big]\\
&\quad=\frac{1}{I}\sum_{i\in\mathcal{I}}\rho_i(x)p_i(u|x)p_i(a|x,u).
\end{aligned}\tag{C.3}$$

The right-hand side is precisely the joint prefix–token probability under the forward target (3.2). More explicitly,

$$\frac{1}{I}\sum_{i\in\mathcal{I}}\rho_i(x)p_i(u|x)p_i(a|x,u)=\bar{\rho}(x)\pi^*_{\mathrm{F}}(u|x)\pi^*_{\mathrm{F}}(a|x,u).\tag{C.4}$$

To see this identity, first note that

$$\begin{aligned}
\bar{\rho}(x)\pi^*_{\mathrm{F}}(u|x)&=\left(\frac{1}{I}\sum_{k\in\mathcal{I}}\rho_k(x)\right)\left(\sum_{i\in\mathcal{I}}\frac{\rho_i(x)}{\sum_{k\in\mathcal{I}}\rho_k(x)}p_i(u|x)\right)\\
&=\frac{1}{I}\sum_{i\in\mathcal{I}}\rho_i(x)p_i(u|x).
\end{aligned}\tag{C.5}$$

Moreover, conditioned on the prefix (x,u), the posterior weight of teacher i is

$$w_i(x,u)=\frac{\rho_i(x)p_i(u|x)}{\sum_{k\in\mathcal{I}}\rho_k(x)p_k(u|x)}.$$

Therefore,

$$\begin{aligned}
\pi^*_{\mathrm{F}}(a|x,u)&=\sum_{i\in\mathcal{I}}w_i(x,u)p_i(a|x,u)\\
&=\frac{\sum_{i\in\mathcal{I}}\rho_i(x)p_i(u|x)p_i(a|x,u)}{\sum_{k\in\mathcal{I}}\rho_k(x)p_k(u|x)}.
\end{aligned}\tag{C.6}$$

Multiplying (C.5) and (C.6), the prefix normalizing factor cancels, giving

$$\bar{\rho}(x)\pi^*_{\mathrm{F}}(u|x)\pi^*_{\mathrm{F}}(a|x,u)=\frac{1}{I}\sum_{i\in\mathcal{I}}\rho_i(x)p_i(u|x)p_i(a|x,u).$$

Recall that

$$c_t(x,u,a):=\frac{1}{m}\sum_{j=1}^{m}\mathbb{1}(x_{t,j}=x,\,y_{t,j,<h}=u)\,q_{t,j,h}(a).$$

Averaging (C.3) over the m rollouts therefore yields

$$\mathbb{E}\left[c_t(x,u,a)\,|\,\mathcal{F}_t\right]=\bar{\rho}(x)\pi^*_{\mathrm{F}}(u|x)\pi^*_{\mathrm{F}}(a|x,u).\tag{C.7}$$

Only linearity of expectation is used here. In particular, the m rollouts need not be independent after marginalizing over their shared teacher index i_t.

For any autoregressive policy π, define its round-t empirical loss by

$$\ell_t(\pi):=-\sum_{h=1}^{H}\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{A}^{h-1}}\sum_{a\in\mathcal{A}}c_t(x,u,a)\log\pi(a|x,u).\tag{C.8}$$

Define $L(\pi) := \mathbb{E}_{x\sim\bar{\rho},y\sim\pi^*_{\mathrm{F}}(\cdot|x)}\big[-\log\pi(y|x)\big]$. Using the autoregressive representation of π, we have

$$-\log\pi(y|x) = -\sum_{h=1}^{H}\log\pi(a_h|x,y_{<h}).$$

Consequently, using the linearity of expectation, $L(\pi)$ can be expressed as

$$L(\pi) = -\sum_{h=1}^{H}\sum_{x,u,a}\bar{\rho}(x)\pi^*_{\mathrm{F}}(u|x)\pi^*_{\mathrm{F}}(a|x,u)\log\pi(a|x,u).$$

Substituting (C.7) into (C.8) now gives, for every $\mathcal{F}_t$-measurable policy π,

$$\mathbb{E}\left[\ell_t(\pi)\,|\,\mathcal{F}_t\right] = L(\pi). \tag{C.9}$$

Moreover, we have

$$\begin{aligned} L(\pi) - L(\pi^*_{\mathrm{F}}) &= \mathbb{E}_{x\sim\bar{\rho},y\sim\pi^*_{\mathrm{F}}(\cdot|x)}\left[\log\frac{\pi^*_{\mathrm{F}}(y|x)}{\pi(y|x)}\right] \\ &= \mathbb{E}_{x\sim\bar{\rho}}\mathrm{KL}\big[\pi^*_{\mathrm{F}}(\cdot|x)\|\pi(\cdot|x)\big]. \end{aligned} \tag{C.10}$$

Applying (C.9) and (C.10) to $\widehat{\pi}_t$ and π^*_{F} and the tower property, we have

$$\begin{aligned} \mathbb{E}\big[\,\mathrm{Regret}(T)\big] &= \mathbb{E}\left[\sum_{t=1}^{T}\big[L(\widehat{\pi}_t) - L(\pi^*_{\mathrm{F}})\big]\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\big[\ell_t(\widehat{\pi}_t) - \ell_t(\pi^*_{\mathrm{F}})\big]\right]. \end{aligned} \tag{C.11}$$

Fix $h\in[H]$, define

$$\ell_{t,h}(\pi) := -\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{A}^{h-1}}\sum_{a\in\mathcal{A}} c_t(x,u,a)\log\pi(a|x,u).$$

Then, $\ell_t(\pi) = \sum_h \ell_{t,h}$. We consider $\sum_{t=1}^{T}\big[\ell_{t,h}(\widehat{\pi}_t) - \ell_{t,h}(\pi^*_{\mathrm{F}})\big]$. Recall that $W_t(x,u,a) = \sum_{i=1}^{t} c_i(x,u,a)$, $W_t(x,u) = \sum_a W_t(x,u,a)$. For a fixed state (x,u), we omit (x,u) and use the shorthand notation $W_t(a)$ and W_t when it will not cause any confusion. Similarly, we write $c_t(a) = c_t(x,u,a)$ and define $c_t := \sum_a c_t(a)$. Thus, $W_t = \sum_{i=1}^{t} c_i$.
Using the definition of $c_t(x,u,a)$ in (3.3), we have

$$c_t = \frac{1}{m}\sum_{j=1}^{m}\mathbb{1}(x_{t,j}=x,\, y_{t,j,<h}=u)\sum_a q_{t,j,h}(a).$$

Note that in both cases, $\sum_a q_{t,j,h}(a) = 1$. Thus, we have

$$c_t = \frac{1}{m}\sum_{j=1}^{m}\mathbb{1}(x_{t,j}=x,\, y_{t,j,<h}=u) \le 1.$$

In Algorithm 1, the policy at context (x,u) is defined as

$$\widehat{\pi}_t(a|x,u) = \frac{W_{t-1}(a)+1/2}{W_{t-1}+A/2}.$$

For simplicity, we write the action set as $\mathcal{A} = \{1,2,\ldots,A\}$. Let $\mathbf{Q}_t$ be a random A-dim probability vector with Dirichlet distribution

$$\mathbf{Q}_t \sim \mathrm{Dir}\left(W_{t-1}(1)+\frac{1}{2},\ldots,W_{t-1}(A)+\frac{1}{2}\right).$$

The goal of this construction is the following fact: the mean of $\mathbf{Q}_t$ is exactly the output policy:

$$\mathbb{E}_{q\sim\mathbf{Q}_t}[q(a)] = \frac{W_{t-1}(a)+1/2}{W_{t-1}+A/2} = \widehat{\pi}_t(a|x,u).$$

For a positive vector $v = (v_1, \dots, v_A)$, define the multivariate beta function by

$$\mathrm{B}(v) := \frac{\prod_{a=1}^{A} \Gamma(v_a)}{\Gamma(\sum_{a=1}^{A} v_a)}, \tag{C.12}$$

where Γ is the gamma function. The beta function is the normalizing constant of the Dirichlet distribution. Indeed, for every positive vector v, the density of the Dirichlet distribution $\mathbf{Q}_v$ on the simplex $\Delta_A := \{q \in \mathbb{R}_+^A : \sum_a q(a) = 1\}$ is

$$f_v(q) = \frac{1}{\mathrm{B}(v)} \prod_{a=1}^{A} q(a)^{v_a - 1}, \qquad q \in \Delta_A.$$

Equivalently,

$$\int_{\Delta_A} \prod_{a=1}^{A} q(a)^{v_a - 1} \, \mathrm{d}q = \mathrm{B}(v).$$

More generally, for every nonnegative vector $c \in \mathbb{R}_+^A$,

$$\begin{aligned} \mathbb{E}_{q \sim \mathbf{Q}_v}\left[\prod_{a=1}^{A} q(a)^{c_a} \right] &= \frac{1}{B(v)} \int_{\Delta_A} \prod_{a=1}^{A} q(a)^{c_a + v_a - 1} \mathrm{d}q \\ &= \frac{\mathrm{B}(v + c)}{\mathrm{B}(v)}. \end{aligned}$$

Therefore, we have

$$\begin{aligned} \mathbb{E}_{q \sim \mathbf{Q}_t}\left[\prod_{a \in \mathcal{A}} q(a)^{c_t(a)} \right] &= \frac{\mathrm{B}\big(W_{t-1}(1) + c_t(1) + \frac{1}{2}, \dots, W_{t-1}(A) + c_t(A) + \frac{1}{2}\big)}{\mathrm{B}\big(W_{t-1}(1) + \frac{1}{2}, \dots, W_{t-1}(A) + \frac{1}{2}\big)} \\ &= \frac{\mathrm{B}\big(W_t(1) + \frac{1}{2}, \dots, W_t(A) + \frac{1}{2}\big)}{\mathrm{B}\big(W_{t-1}(1) + \frac{1}{2}, \dots, W_{t-1}(A) + \frac{1}{2}\big),} \end{aligned} \tag{C.13}$$

where we use $W_t(a) = W_{t-1}(a) + c_t(a)$. Let

$$\alpha_t := \sum_{a \in \mathcal{A}} c_t(a) \le 1, \qquad \mathcal{A}_t^+ := \{a \in \mathcal{A} : c_t(a) > 0\}.$$

For every $a \in \mathcal{A}_t^+$, define the Hölder exponent $r_a := 1/c_t(a)$. If $\alpha_t < 1$, introduce one additional exponent $r_0 := 1/1 - \alpha_t$. These exponents satisfy

$$\frac{1}{r_0} + \sum_{a \in \mathcal{A}_t^+} \frac{1}{r_a} = (1 - \alpha_t) + \sum_{a \in \mathcal{A}_t^+} c_t(a) = 1.$$

Using the generalized Hölder's inequality, we have

$$\begin{aligned} \mathbb{E}_{q \sim \mathbf{Q}_t}\left[\prod_{a \in \mathcal{A}} q(a)^{c_t(a)} \right] &\le \mathbb{E}[1^{r_0}]^{\frac{1}{r_0}} \prod_{a \in \mathcal{A}_t^+} \mathbb{E}_{q \sim \mathbf{Q}_t}\left[\left[q(a)^{c_t(a)} \right]^{r_a} \right]^{\frac{1}{r_a}} \\ &= \prod_{a \in \mathcal{A}_t^+} \mathbb{E}_{q \sim \mathbf{Q}_t}\big[q(a)\big]^{c_t(a)} = \prod_{a \in \mathcal{A}} \widehat{\pi}_t(a|x,u)^{c_t(a)}. \end{aligned} \tag{C.14}$$

where we used $\mathbb{E}_{q \sim \mathbf{Q}_t}[q(a)] = \widehat{\pi}_t(a|x,u)$. If $\alpha_t = 1$, we can skip r_0 and use the same generalized Hölder's inequality.

Combining (C.13) and (C.14), multiplying over t, and using the telescoping argument, we have

$$\begin{aligned} \prod_{t=1}^{T} \prod_{a \in \mathcal{A}} \widehat{\pi}_t(a|z)^{c_t(a)} &\ge \prod_{t=1}^{T} \frac{\mathrm{B}\big(W_t(\cdot) + \frac{1}{2}\mathbf{1}\big)}{\mathrm{B}\big(W_{t-1}(\cdot) + \frac{1}{2}\mathbf{1}\big)} \\ &= \frac{\mathrm{B}\big(W_T(\cdot) + \frac{1}{2}\mathbf{1}\big)}{\mathrm{B}\big(\frac{1}{2}\mathbf{1}\big)}. \end{aligned}$$

Taking the negative logarithm, we have

$$-\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} c_t(a)\log\widehat{\pi}_t(a|z) \le \log\frac{\mathrm{B}(\frac{1}{2}\mathbf{1})}{\mathrm{B}\big(W_T(\cdot)+\frac{1}{2}\mathbf{1}\big)}. \tag{C.15}$$

Next, we consider an arbitrary policy $\pi(\cdot|x,u)$. First suppose $W_T > 0$. By the nonnegativity of KL divergence,

$$\begin{aligned}\sum_{a\in\mathcal{A}} W_T(a)\log\pi(a|x,u) &= \sum_{a\in\mathcal{A}} W_T(a)\log\frac{W_T(a)}{W_T} - W_T\sum_{a\in\mathcal{A}}\frac{W_T(a)}{W_T}\log\frac{W_T(a)}{W_T\cdot\pi(a|x,u)}\\ &= \sum_{a\in\mathcal{A}} W_T(a)\log\frac{W_T(a)}{W_T} - W_T\mathrm{KL}\bigg(\frac{W_T(\cdot)}{W_T}\bigg\|\pi(\cdot|x,u)\bigg)\\ &\le \sum_{a\in\mathcal{A}} W_T(a)\log\frac{W_T(a)}{W_T},\end{aligned} \tag{C.16}$$

where we set $0\log 0 := 0$. Summing (C.15) and (C.16), we have

$$\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} c_t(a)\log\frac{\pi(a|z)}{\widehat{\pi}_t(a|z)} \le \log\frac{\mathrm{B}(\frac{1}{2}\mathbf{1})}{\mathrm{B}(W_T(\cdot)+\frac{1}{2}\mathbf{1})} + \sum_{a\in\mathcal{A}} W_T(a)\log\frac{W_T(a)}{W_T} =: \mathcal{R}\big(W_T(\cdot)\big).$$

When $W_T = 0$, let $\mathcal{R}\big(W_T(\cdot)\big) = 0$. The following lemma bounds $\mathcal{R}\big(W_T(\cdot)\big)$ using Stirling's formula.

Lemma C.1. Let $W_T(a) \ge 0$ for every $a \in \mathcal{A}$, and let

$$W_T := \sum_{a\in\mathcal{A}} W_T(a).$$

Then, $\mathcal{R}\big(W_T(\cdot)\big)$ satisfies

$$\mathcal{R}\big(W_T(\cdot)\big) \le A\log(W_T+1) + 3A.$$

We defer the proof of Lemma C.1 to Appendix E. Choose $\pi = \pi^*_{\mathrm{F}}$. For any $(x,u) \in \mathcal{X}\times\mathcal{A}^{h-1}$, we have

$$\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} c_t(x,u,a)\log\frac{\pi^*_{\mathrm{F}}(a|x,u)}{\widehat{\pi}_t(a|x,u)} \le A\log\big(W_T(x,u)+1\big) + 3A.$$

Therefore, we have

$$\begin{aligned}\sum_{t=1}^{T}\big[\ell_{t,h}(\widehat{\pi}_t) - \ell_{t,h}(\pi^*_{\mathrm{F}})\big] &= \sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{A}^{h-1}}\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} c_t(x,u,a)\log\frac{\pi(a|x,u)}{\pi^*_{\mathrm{F}}(a|x,u)}\\ &\le \sum_{(x,u)\in\mathcal{X}\times\mathcal{A}^{h-1}}\Big[A\log\big(W_T(x,u)+1\big)+3A\Big].\end{aligned}$$

For every fixed level h and round t, easy to check

$$\sum_{(x,u)\in\mathcal{X}\times\mathcal{A}^{h-1}}\sum_{a\in\mathcal{A}} c_t(x,u,a) = 1.$$

Consequently,

$$\sum_{(x,u)\in\mathcal{X}\times\mathcal{A}^{h-1}} W_T(x,u) = T,$$

and hence $W_T(x,u) \le T$ for any (x,u). This leads to

$$\sum_{t=1}^{T}\big[\ell_{t,h}(\widehat{\pi}_t) - \ell_{t,h}(\pi^*_{\mathrm{F}})\big] \le 2SA^h\log T. \tag{C.17}$$

Finally, summing (C.17) over $h \in [H]$, we have

$$\begin{aligned}\sum_{t=1}^{T}\big[\ell_t(\widehat{\pi}_t) - \ell_t(\pi_{\mathrm{F}}^*)\big] &\le 2S\log T\sum_{h=1}^{H}A^h \\ &= 2S\cdot\log T\frac{A(A^H-1)}{A-1} \\ &\le 4SA^H\log T. \end{aligned} \tag{C.18}$$

Using (C.11), we obtain

$$\mathbb{E}\big[\operatorname{Regret}(T)\big] \le 4SA^H\log T.$$

We next consider the high-probability regret bound. Define

$$\begin{aligned} X_t &:= \ell_t(\pi_{\mathrm{F}}^*) - \ell_t(\widehat{\pi}_t), \\ r_t &:= -\mathbb{E}[X_t \mid \mathcal{F}_t] = \mathbb{E}_{x\sim\bar{\rho}}\mathrm{KL}\big(\pi_{\mathrm{F}}^*(\cdot|x)\|\widehat{\pi}_t(\cdot|x)\big). \end{aligned}$$

Thus, $\operatorname{Regret}(T) = \sum_{t=1}^{T} r_t$. Define $M_0 = 0$, and

$$M_t := \sum_{s=1}^{t}(X_s + r_s), \qquad t \in [T].$$

Since $r_t = -\mathbb{E}[X_t \mid \mathcal{F}_t]$, we have $\mathbb{E}[M_t - M_{t-1}|\mathcal{F}_t] = 0$. Thus, $\{M_t\}_{t=0}^{T}$ is a martingale with respect to $\{\mathcal{F}_{t+1}\}_{t=0}^{T}$.
Our goal is to control M_T. Using the definition of ℓ_t,

$$X_t = \sum_{h=1}^{H}\sum_{(x,u,a)} c_t(x,u,a)\log\frac{\widehat{\pi}_t(a|x,u)}{\pi_{\mathrm{F}}^*(a|x,u)}.$$

First note that in Algorithm 1,

$$\begin{aligned}\widehat{\pi}_t(a|x,u) &= \frac{W_{t-1}(x,u,a)+1/2}{W_{t-1}(x,u)+A/2} \\ &\ge \frac{1}{2(t-1)+A} \ge \frac{1}{2T+A}.\end{aligned}$$

Since $\pi_{\mathrm{F}}^*(a|x,u) \le 1$, it follows that

$$\frac{X_t}{H} \ge -L_T, \qquad L_T := \log(2T+A). \tag{C.19}$$

Here we use the following equation

$$\sum_{(x,u,a)} c_t(x,u,a) = 1.$$

Moreover, we can see

$$\sum_{h=1}^{H}\sum_{(x,u,a)}\frac{c_t(x,u,a)}{H} = 1.$$

Using the AM-GM inequality, we have

$$\begin{aligned}\exp\left(\frac{X_t}{H}\right) &= \prod_{h=1}^{H}\prod_{(x,u,a)}\left(\frac{\widehat{\pi}_t(a|x,u)}{\pi_{\mathrm{F}}^*(a|x,u)}\right)^{c_t(x,u,a)/H} \\ &\le \frac{1}{H}\sum_{h=1}^{H}\sum_{(x,u,a)} c_t(x,u,a)\frac{\widehat{\pi}_t(a|x,u)}{\pi_{\mathrm{F}}^*(a|x,u)}.\end{aligned}$$

Taking conditional expectation over $\mathcal{F}_t$ and using

$$\mathbb{E}[c_t(x,u,a) \mid \mathcal{F}_t] = \bar{\rho}(x)\pi_{\mathrm{F}}^*(u|x)\pi_{\mathrm{F}}^*(a|x,u),$$

we obtain

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{X_t}{H}\right)\middle|\mathcal{F}_t\right] &\le \frac{1}{H}\sum_{h=1}^{H}\sum_{(x,u,a)}\bar{\rho}(x)\pi_{\mathrm{F}}^*(u|x)\pi_{\mathrm{F}}^*(a|x,u)\frac{\widehat{\pi}_t(a|x,u)}{\pi_{\mathrm{F}}^*(a|x,u)} \\
&= \frac{1}{H}\sum_{h=1}^{H}\sum_{(x,u)}\bar{\rho}(x)\pi_{\mathrm{F}}^*(u|x)\sum_{a}\widehat{\pi}_t(a|x,u) \\
&\le \frac{1}{H}\sum_{h=1}^{H}\sum_{x\in\mathcal{X}}\bar{\rho}(x)\sum_{u\in\mathcal{A}^{h-1}}\pi_{\mathrm{F}}^*(u|x) \\
&= 1. \qquad \text{(C.20)}
\end{aligned}$$

The following lemma shows that a one-sided lower bound and an exponential central condition provide exponential control of deviations from the conditional mean.

Lemma C.2. Let $\mathcal{F}$ be a σ-algebra, and let Y be a random variable satisfying, almost surely,

$$Y \ge -L, \qquad \mathbb{E}[e^Y \mid \mathcal{F}] \le 1,$$

for some $L \ge 0$. Define $\mu := -\mathbb{E}[Y \mid \mathcal{F}]$. Then $\mu \in [0, L]$, and, for every $\theta \in [0,1]$,

$$\log \mathbb{E}\left[e^{\theta(Y+\mu)}\middle|\mathcal{F}\right] \le (1+L)\theta^2\mu.$$

Applying Lemma C.2 to $Y = X_t/H$, $L = L_T$, we have for every $\lambda \in [0, 1/H]$

$$\log \mathbb{E}\left[e^{\lambda(X_t+r_t)}\,\middle|\,\mathcal{F}_t\right] \le H(1+L_T)\lambda^2 r_t. \qquad \text{(C.21)}$$

Define

$$\mathcal{E}_t(\lambda) := \exp\left(\lambda\sum_{s=1}^{t}(X_s+r_s) - H(1+L_T)\lambda^2\sum_{s=1}^{t}r_s\right).$$

Then, (C.21) implies that $\{\mathcal{E}_s(\lambda)\}_{s=0}^T$ is a nonnegative supermartingale with $\mathcal{E}_0(\lambda) = 1$. As a result, $\mathbb{E}[\mathcal{E}_T(\lambda)] \le \mathbb{E}[\mathcal{E}_0(\lambda)] = 1$. Moreover, we have

$$\mathcal{E}_T(\lambda) = \exp\Big(\lambda M_T - H(1+L_T)\lambda^2\,\mathrm{Regret}(T)\Big).$$

Therefore, for any $C > 0$, by Markov's inequality, we have

$$\mathbb{P}\Big[\lambda M_T - H(1+L_T)\lambda^2\,\mathrm{Regret}(T) \ge C\Big] \le \frac{\mathbb{E}[\mathcal{E}_T(\lambda)]}{e^C} \le \frac{1}{e^C}.$$

Set $C = \log(1/\delta)$. Now we obtain with probability at least $1-\delta$,

$$M_T = \sum_{t=1}^{T}(X_t + r_t) \le H(1+L_T)\lambda\,\mathrm{Regret}(T) + \frac{\log(1/\delta)}{\lambda}. \qquad \text{(C.22)}$$

On the other hand, (C.18) gives

$$-\sum_{t=1}^{T}X_t = \sum_{t=1}^{T}[\ell_t(\widehat{\pi}_t) - \ell_t(\pi_{\mathrm{F}}^*)] \le 4SA^H\log T.$$

Moreover, we have

$$\mathrm{Regret}(T) = -\sum_{t=1}^{T}X_t + \sum_{t=1}^{T}(X_t + r_t),$$

Therefore, with probability at least $1-\delta$, the following inequality holds:

$$\mathrm{Regret}(T) \le 4SA^H \log T + H(1+L_T)\lambda\,\mathrm{Regret}(T) + \frac{\log(1/\delta)}{\lambda}. \tag{C.23}$$

Choose

$$\lambda = \frac{1}{2H(1+\log(2T+A))} \le \frac{1}{H}.$$

We have $H(1+L_T)\lambda = 1/2$. Rearranging (C.23) yields

$$\mathrm{Regret}(T) \le 8SA^H \log T + 4H\big(1+\log(2T+A)\big)\log\frac{1}{\delta}.$$

□

D Proof of Theorem 4.5

Proof of Theorem 4.5. We first control the error of the optimistic token-score estimates. Let $\mathcal{F}_t$ denotes the history before round t. In this way, both $\widehat{l}_{t-1}$ and $\widehat{\pi}_t$ are $\mathcal{F}_t$-measurable.
Let $\mathcal{E}$ be the event in Lemma 4.4, which satisfies $\mathbb{P}(\mathcal{E}) \ge 1-2\delta$. On this event, simultaneously for every round s and every visited triple (x,u,a),

$$\big|\bar{l}_s(x,u,a) - \log g_h(a|x,u)\big| \le \beta_s(x,u,a). \tag{D.1}$$

Moreover, Assumption 4.3 and the definition of g_h imply that for every teacher i,

$$\big|\log p_i(a|x,u) - \log g_h(a|x,u)\big| = \left|\log\frac{p_i(a|x,u)}{\pi_{\text{ref}}(a|x,u)} - \sum_{k\in\mathcal{I}} w_k(x)\log\frac{p_k(a|x,u)}{\pi_{\text{ref}}(a|x,u)}\right| \le 2B.$$

Since $\bar{l}_s(x,u,a)$ is an average of the observed teacher log-probabilities at this triple, it follows that

$$\big|\bar{l}_s(x,u,a) - \log g_h(a|x,u)\big| \le 2B. \tag{D.2}$$

Combining (D.1) and (D.2), we obtain

$$\big|\bar{l}_s(x,u,a) - \log g_h(a|x,u)\big| \le \min\big\{\beta_s(x,u,a), 2B\big\}.$$

Consequently, on the event $\mathcal{E}$, the optimistic estimate $\widehat{l}_s = \bar{l}_s + \min\{\beta_s, 2B\}$ satisfies

$$0 \le \widehat{l}_s(x,u,a) - \log g_h(a|x,u) \le \min\big\{2\beta_s(x,u,a), 4B\big\}.$$

For an unvisited triple, the algorithm sets

$$\widehat{l}_s(x,u,a) = \log \pi_{\text{ref}}(a|x,u) + B.$$

Assumption 4.3 implies

$$-B \le \log g_h(a|x,u) - \log \pi_{\text{ref}}(a|x,u) \le B.$$

Therefore,

$$0 \le \widehat{l}_s(x,u,a) - \log g_h(a|x,u) \le 2B.$$

We next consider the output policy. Multiplying the token conditionals defined by the backward recursion gives

$$\begin{aligned}\widehat{\pi}_t(y|x) &= \prod_{h=1}^{H} \frac{\exp\big(\widehat{l}_{t-1}(x,y_{<h},a_h)\big)\widehat{V}_{t-1,h+1}(x,y_{\le h})}{\widehat{V}_{t-1,h}(x,y_{<h})} \\ &= \frac{\exp\Big(\sum_{h=1}^{H}\widehat{l}_{t-1}(x,y_{<h},a_h)\Big)}{\widehat{V}_{t-1,1}(x)},\end{aligned} \tag{D.3}$$

where the second equation holds by canceling out the telescope terms and $\widehat{V}_{t-1,H+1} = 1$. Comparing this expression with the reverse target in (4.1), we obtain

$$\log\frac{\widehat{\pi}_t(y|x)}{\pi^*_{\mathrm{R}}(y|x)} = \sum_{h=1}^{H}\Big[\widehat{l}_{t-1}(x,y_{<h},a_h) - \log g_h(a_h|x,y_{<h})\Big] + \log\frac{V_1(x)}{\widehat{V}_{t-1,1}(x)}. \tag{D.4}$$

Moreover, the ratio of the normalizing constants can be expressed as an expectation under $\widehat{\pi}_t$. To see this, first expand the expectation:

$$\mathbb{E}_{y\sim\widehat{\pi}_t(\cdot|x)} \exp\Big(-\sum_{h=1}^{H} \Big[\widehat{l}_{t-1}(x, y_{<h}, a_h) - \log g_h(a_h|x, y_{<h})\Big]\Big)$$
$$= \sum_{y\in\mathcal{Y}} \widehat{\pi}_t(y|x) \exp\Big(-\sum_{h=1}^{H} \widehat{l}_{t-1}(x, y_{<h}, a_h) + \sum_{h=1}^{H} \log g_h(a_h|x, y_{<h})\Big). \tag{D.5}$$

Substituting (D.3) into (D.5), the right-hand side becomes

$$\sum_{y\in\mathcal{Y}} \frac{\exp\big(\sum_{h=1}^{H} \widehat{l}_{t-1}(x, y_{<h}, a_h)\big)}{\widehat{V}_{t-1,1}(x)} \exp\Big(-\sum_{h=1}^{H} \widehat{l}_{t-1}(x, y_{<h}, a_h) + \sum_{h=1}^{H} \log g_h(a_h|x, y_{<h})\Big)$$
$$= \frac{1}{\widehat{V}_{t-1,1}(x)} \sum_{y\in\mathcal{Y}} \exp\Big(\sum_{h=1}^{H} \log g_h(a_h|x, y_{<h})\Big)$$
$$= \frac{1}{\widehat{V}_{t-1,1}(x)} \sum_{y\in\mathcal{Y}} \prod_{h=1}^{H} g_h(a_h|x, y_{<h}).$$

Finally, recall that in Theorem 4.1, we have seen

$$V_1(x) = \sum_{y\in\mathcal{Y}} \prod_{h=1}^{H} g_h(a_h|x, y_{<h}).$$

Consequently,

$$\mathbb{E}_{y\sim\widehat{\pi}_t(\cdot|x)} \exp\Big(-\sum_{h=1}^{H} \Big[\widehat{l}_{t-1}(x, y_{<h}, a_h) - \log g_h(a_h|x, y_{<h})\Big]\Big) = \frac{V_1(x)}{\widehat{V}_{t-1,1}(x)}. \tag{D.6}$$

Taking expectation over $\widehat{\pi}_t(\cdot|x)$ and substituting (D.6) into (D.4) yields

$$\mathrm{KL}\big(\widehat{\pi}_t(\cdot|x) \,\big\|\, \pi^*_{\mathrm{R}}(\cdot|x)\big) = \mathbb{E}_{y\sim\widehat{\pi}_t(\cdot|x)}\Big[\sum_{h=1}^{H} \Big[\widehat{l}_{t-1}(x, y_{<h}, a_h) - \log g_h(a_h|x, y_{<h})\Big]\Big]$$
$$+ \log \mathbb{E}_{y\sim\widehat{\pi}_t(\cdot|x)} \exp\Big(-\sum_{h=1}^{H} \Big[\widehat{l}_{t-1}(x, y_{<h}, a_h) - \log g_h(a_h|x, y_{<h})\Big]\Big). \tag{D.7}$$

On the event $\mathcal{E}$, $\widehat{l}_{t-1}(x, y_{<h}, a_h) \geq \log g_h(a_h|x, y_{<h})$ for any t, h. We can therefore apply the basic inequality $e^{-v} \leq 1 - v + v^2/2$ for $v \geq 0$. Let $V = \sum_{h=1}^{H} \big[\widehat{l}_{t-1}(x, y_{<h}, a_h) - \log g_h(a_h|x, y_{<h})\big]$. The right-hand side can be bounded as

$$\begin{aligned}\mathbb{E}[V] + \log \mathbb{E}\exp(-V) &\leq \mathbb{E}[V] + \log\big(1 - \mathbb{E}[V] + \mathbb{E}[V^2]/2\big)\\ &\leq \mathbb{E}[V] - \mathbb{E}[V] + \frac{1}{2}\mathbb{E}[V^2]\\ &= \frac{1}{2}\mathbb{E}[V^2],\end{aligned}$$

where we use $\log x \leq x - 1$ for any $x > 0$. As a result, we have

$$\mathrm{KL}\big(\widehat{\pi}_t(\cdot|x) \,\big\|\, \pi^*_{\mathrm{R}}(\cdot|x)\big) \leq \frac{1}{2}\mathbb{E}_{y\sim\widehat{\pi}_t(\cdot|x)}\Big[\Big(\sum_{h=1}^{H} \Big[\widehat{l}_{t-1}(x, y_{<h}, a_h) - \log g_h(a_h|x, y_{<h})\Big]\Big)^2\Big]$$
$$\leq \frac{H}{2}\mathbb{E}_{y\sim\widehat{\pi}_t(\cdot|x)}\Big[\sum_{h=1}^{H} \Big[\widehat{l}_{t-1}(x, y_{<h}, a_h) - \log g_h(a_h|x, y_{<h})\Big]^2\Big], \tag{D.8}$$

where the last inequality holds due to the Cauchy–Schwarz inequality.

We next relate the expected squared errors in (D.8) to those evaluated on the observed rollouts. Define

$$X_t := \frac{H}{2m} \sum_{j=1}^{m} \sum_{h=1}^{H} \Big[\widehat{l}_{t-1}(x_{t,j}, y_{t,j,<h}, a_{t,j,h}) - \log g_h(a_{t,j,h} | x_{t,j}, y_{t,j,<h}) \Big]^2.$$

Conditionally on $\mathcal{F}_t$, each rollout has joint distribution

$$\begin{aligned} \mathbb{P}(x_{t,j} = x, y_{t,j} = y \mid \mathcal{F}_t) &= \frac{1}{I} \sum_{i \in \mathcal{I}} \rho_i(x) \widehat{\pi}_t(y|x) \\ &= \bar{\rho}(x) \widehat{\pi}_t(y|x). \end{aligned}$$

Since $\widehat{l}_{t-1}$ and $\widehat{\pi}_t$ are $\mathcal{F}_t$-measurable, linearity of conditional expectation therefore gives

$$\begin{aligned} \mathbb{E}[X_t \mid \mathcal{F}_t] &= \frac{H}{2m} \sum_{j=1}^{m} \mathbb{E}_{x \sim \bar{\rho}} \mathbb{E}_{y \sim \widehat{\pi}_t(\cdot|x)} \left[\sum_{h=1}^{H} \Big[\widehat{l}_{t-1}(x, y_{<h}, a_h) - \log g_h(a_h | x, y_{<h}) \Big]^2 \right] \\ &= \frac{H}{2} \mathbb{E}_{x \sim \bar{\rho}} \mathbb{E}_{y \sim \widehat{\pi}_t(\cdot|x)} \left[\sum_{h=1}^{H} \Big[\widehat{l}_{t-1}(x, y_{<h}, a_h) - \log g_h(a_h | x, y_{<h}) \Big]^2 \right]. \end{aligned} \tag{D.9}$$

Note that this does not require independence among the rollouts within a round. On the event $\mathcal{E}$, combining (D.9) with (D.8), we have

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} \mathbb{E}_{x \sim \bar{\rho}} \mathrm{KL}\big(\widehat{\pi}_t(\cdot|x) \,\big\|\, \pi_{\mathrm{R}}^*(\cdot|x) \big) \le \sum_{t=1}^{T} \mathbb{E}[X_t \mid \mathcal{F}_t]. \tag{D.10}$$

To apply Lemma J.6, we need a bound on X_t. At every previously visited triple, (D.2) and the definition of $\widehat{l}_{t-1}$ give

$$\begin{aligned} \big| \widehat{l}_{t-1}(x,u,a) - \log g_h(a|x,u) \big| &\le \big| \bar{l}_{t-1}(x,u,a) - \log g_h(a|x,u) \big| + \min\{\beta_{t-1}(x,u,a), 2B\} \\ &\le 4B. \end{aligned}$$

At an unvisited triple, $\widehat{l}_{t-1}(x,u,a) = \log \pi_{\text{ref}}(a|x,u) + B$ indicates $\big| \widehat{l}_{t-1}(x,u,a) - \log g_h(a|x,u) \big| \in [0, 2B]$. Consequently, in both cases,

$$0 \le X_t \le \frac{H}{2m} \cdot mH \cdot (4B)^2 = 8H^2B^2.$$

Moreover, X_t is $\mathcal{F}_{t+1}$-measurable. Applying Lemma J.6, with probability at least $1 - \delta$, the following inequality holds

$$\sum_{t=1}^{T} \mathbb{E}[X_t \mid \mathcal{F}_t] \le 2 \sum_{t=1}^{T} X_t + 64H^2B^2 \log \frac{2}{\delta}. \tag{D.11}$$

Substituting (D.11) into (D.10) and taking a union bound, we conclude that, with probability at least $1 - 3\delta$,

$$\begin{aligned} \mathrm{Regret}(T) \le \frac{H}{m} \underbrace{\sum_{t=1}^{T} \sum_{j=1}^{m} \sum_{h=1}^{H} \Big[\widehat{l}_{t-1}(x_{t,j}, y_{t,j,<h}, a_{t,j,h}) - \log g_h(a_{t,j,h} | x_{t,j}, y_{t,j,<h}) \Big]^2}_{I} \\ + 64H^2B^2 \log \frac{2}{\delta}. \end{aligned} \tag{D.12}$$

It remains to bound I for every data trajectory on $\mathcal{E}$. Regrouping the observations by their corresponding triples gives

$$I = \sum_{h=1}^{H} \sum_{(x,u,a)} \sum_{t=1}^{T} \big[N_t(x,u,a) - N_{t-1}(x,u,a) \big] \Big[\widehat{l}_{t-1}(x,u,a) - \log g_h(a|x,u) \Big]^2.$$

Here $N_t(x,u,a) - N_{t-1}(x,u,a)$ counts the visits to this triple in round t, each of which contributes the same squared error. Since there are m samples per round,

$$0 \le N_t(x,u,a) - N_{t-1}(x,u,a) \le m, \qquad N_T(x,u,a) \le mT.$$

For each fixed triple (x,u,a), we split the inner sum according to whether $N_{t-1}(x,u,a) = 0$ or $N_{t-1}(x,u,a) \ge 1$. In the former case, only the first round that visits this triple can contribute, with at most m visits. Hence,

$$\sum_{\substack{t\in[T]:\\ N_{t-1}(x,u,a)=0}} \big[N_t(x,u,a) - N_{t-1}(x,u,a)\big]\Big[\widehat{l}_{t-1}(x,u,a) - \log g_h(a|x,u)\Big]^2 \le 4mB^2.$$

Summing this bound over all $K := S\sum_{h=1}^{H} A^h$ triples and applying Lemma 4.4 gives

$$\begin{aligned} I &\le 4mB^2K \\ &\quad + \sum_{h=1}^{H} \sum_{(x,u,a)} \sum_{\substack{t\in[T]:\\ N_{t-1}(x,u,a)\ge 1}} \big[N_t(x,u,a) - N_{t-1}(x,u,a)\big] \cdot \min\Big\{16B^2,\, 4\beta_{t-1}(x,u,a)^2\Big\}. \end{aligned} \tag{D.13}$$

Recall that

$$\beta_t(x,u,a) = \widetilde{O}\bigg(B\sqrt{\frac{mH}{N_t(x,u,a)}} + \frac{BmH}{N_t(x,u,a)}\bigg).$$

For $N_{t-1}(x,u,a) \ge 1$, squaring the confidence radius gives

$$4\beta_{t-1}(x,u,a)^2 \le \widetilde{O}\left(\frac{B^2mH}{N_{t-1}(x,u,a)} + \frac{B^2m^2H^2}{N_{t-1}(x,u,a)^2}\right),$$

where we used $(v+w)^2 \le 2v^2 + 2w^2$. Combining this with the constant bound $16B^2$, we obtain

$$\begin{aligned} \min\Big\{16B^2,\, 4\beta_{t-1}(x,u,a)^2\Big\} &\le \widetilde{O}\bigg(B^2 \min\bigg\{1,\, \frac{mH}{N_{t-1}(x,u,a)} + \frac{m^2H^2}{N_{t-1}(x,u,a)^2}\bigg\}\bigg) \\ &\le \widetilde{O}\bigg(B^2 \min\bigg\{1, \frac{mH}{N_{t-1}(x,u,a)}\bigg\}\bigg), \end{aligned}$$

where we use $\min\{1, v+v^2\} \le 2\min\{1,v\}$ since $v^2 \le v$ when $0 \le v \le 1$.
Fix a triple (x,u,a) with $N_T(x,u,a) \ge 1$. Order its visits chronologically, breaking ties within each round arbitrarily. If its k-th visit occurs in round t, then

$$k \le N_{t-1}(x,u,a) + m \le N_{t-1}(x,u,a) + mH,$$

because at most m visits occur in that round and $H \ge 1$. For each visit included in the remaining sum, $N_{t-1}(x,u,a) \ge 1$, and hence

$$\begin{aligned} \min\bigg\{1, \frac{mH}{N_{t-1}(x,u,a)}\bigg\} &\le \frac{2mH}{N_{t-1}(x,u,a) + mH} \\ &\le \frac{2mH}{k}. \end{aligned}$$

The first inequality uses $\min\{1,v\} \le 2v/(1+v)$ for $v \ge 0$. Since $N_t(x,u,a) - N_{t-1}(x,u,a)$ counts the visits in round t, the following summation counts all the visits to (x,u,a) and thus can be indexed by k, leading to

$$\begin{aligned} &\sum_{\substack{t\in[T]:\\ N_{t-1}(x,u,a)\ge 1}} \big[N_t(x,u,a) - N_{t-1}(x,u,a)\big] \min\bigg\{1, \frac{mH}{N_{t-1}(x,u,a)}\bigg\} \\ &\quad \le 2mH \sum_{k=1}^{N_T(x,u,a)} \frac{1}{k} \end{aligned}$$

$$\leq 2mH\big[1+\log N_T(x,u,a)\big]$$
$$\leq 2mH\log(emT). \tag{D.14}$$

where we use $\sum_{k=1}^{T} 1/k \leq 1+\log T$ and $N_T(x,u,a) \leq mT$.
Substituting (D.14) into (D.13) and summing over all (x,u,a) pairs, we have

$$I \leq 4mB^2K + \widetilde{O}\big(B^2mHK\log(emT)\big) = \widetilde{O}\big(B^2mHK\log T\big) = \widetilde{O}\big(B^2mSHA^H\log T\big).$$

In conclusion, (D.12) gives that with probability at least $1-3\delta$, the following inequality holds:

$$\begin{aligned}\mathrm{Regret}(T) &\leq \frac{H}{m}I + \widetilde{O}(B^2H^2)\\ &\leq \widetilde{O}\big(B^2H^2SA^H\log T\big).\end{aligned}$$

This completes the proof of Theorem 4.5. □

E Proof of Lemmas Used in Appendices C and D

E.1 Proof of Lemma C.1

Proof of Lemma C.1. When $W_T = 0$, the inequality holds naturally. We therefore assume throughout the rest of the proof that $W_T > 0$. If $A = 1$, then $\mathrm{B}(v) = \frac{\Gamma(v)}{\Gamma(v)} = 1$ for every $v > 0$, and

$$W_T(1)\log\frac{W_T(1)}{W_T} = 0.$$

Thus, $\mathcal{R}(W_T(\cdot)) = 0$. It remains to consider $A \geq 2$.
By the definition of the beta function (C.12), we have

$$-\log\mathrm{B}\Big(W_T(\cdot)+\frac{1}{2}\mathbf{1}\Big) = \log\Gamma\Big(W_T+\frac{A}{2}\Big) - \sum_{a\in\mathcal{A}}\log\Gamma\Big(W_T(a)+\frac{1}{2}\Big). \tag{E.1}$$

Using Lemma J.8 with $s = W_T + A/2$, we have

$$\begin{aligned}\log\Gamma\Big(W_T+\frac{A}{2}\Big) \leq \Big(W_T+\frac{A-1}{2}\Big)\log\Big(W_T+\frac{A}{2}\Big) - \Big(W_T+\frac{A}{2}\Big)\\ + \frac{1}{2}\log(2\pi) + \frac{1}{12(W_T+A/2)}.\end{aligned} \tag{E.2}$$

For every $a \in \mathcal{A}$, applying Lemma J.8 with $s = W_T(a) + 1/2$, we have

$$\log\Gamma\Big(W_T(a)+\frac{1}{2}\Big) \geq W_T(a)\log\Big(W_T(a)+\frac{1}{2}\Big) - \Big(W_T(a)+\frac{1}{2}\Big) + \frac{1}{2}\log(2\pi). \tag{E.3}$$

Substituting (E.2) and (E.3) into (E.1), we obtain

$$\begin{aligned}-\log\mathrm{B}\Big(W_T(\cdot)+\frac{1}{2}\mathbf{1}\Big) \leq -\sum_{a\in\mathcal{A}} W_T(a)\log\Big(W_T(a)+\frac{1}{2}\Big)\\ + \Big(W_T+\frac{A-1}{2}\Big)\log\Big(W_T+\frac{A}{2}\Big) - \frac{A-1}{2}\log(2\pi) + \frac{1}{12(W_T+A/2)}.\end{aligned}$$

where we use $\sum_{a\in\mathcal{A}} W_T(a) = W_T$. Recall that

$$\mathcal{R}\big(W_T(\cdot)\big) = \log\mathrm{B}\Big(\frac{1}{2}\mathbf{1}\Big) - \log\mathrm{B}\Big(W_T(\cdot)+\frac{1}{2}\mathbf{1}\Big) + \sum_{a\in\mathcal{A}} W_T(a)\log\frac{W_T(a)}{W_T}.$$

Therefore, we have

$$\mathcal{R}\big(W_T(\cdot)\big) \leq \log\mathrm{B}\Big(\frac{1}{2}\mathbf{1}\Big) + \underbrace{\sum_{a\in\mathcal{A}} W_T(a)\log\frac{W_T(a)}{W_T(a)+1/2}}_{I_1}$$

$$+\underbrace{W_T \log \frac{W_T + A/2}{W_T}}_{I_2} + \underbrace{\frac{A-1}{2} \log\Big(W_T + \frac{A}{2}\Big)}_{I_3} - \frac{A-1}{2}\log(2\pi) + \frac{1}{12(W_T + A/2)}. \quad \text{(E.4)}$$

Since $0 \le W_T(a)/[W_T(a) + 1/2] \le 1$, $I_1 \le 0$. For I_2, using $\log(1+s) \le s$ for every $s \ge 0$, we have

$$I_2 = W_T \log\left(1 + \frac{A}{2W_T}\right) \le \frac{A}{2}.$$

For I_3, we have $W_T + \frac{A}{2} \le (W_T + 1)\big(1 + A/2\big)$. Thus,

$$W_T + \frac{A}{2} \le (W_T + 1)\Big(1 + \frac{A}{2}\Big),$$

and hence $I_3 \le (A-1)\log(W_T+1)/2 + (A-1)\log(1+A/2)/2$. As a result, (E.4) indicates

$$\mathcal{R}\big(W_T(\cdot)\big) \le \frac{A-1}{2}\log(W_T+1) + \log \mathrm{B}\Big(\frac{1}{2}\mathbf{1}\Big) + \frac{A}{2} + \frac{A-1}{2}\log\Big(1+\frac{A}{2}\Big) + \frac{1}{12}. \quad \text{(E.5)}$$

Finally, by definition,

$$\log \mathrm{B}\Big(\frac{1}{2}\mathbf{1}\Big) = \frac{A}{2}\log\pi - \log\Gamma\Big(\frac{A}{2}\Big).$$

Applying Lemma J.8 with $s = A/2$, we have

$$\log\Gamma\Big(\frac{A}{2}\Big) \ge \frac{A-1}{2}\log\frac{A}{2} - \frac{A}{2} + \frac{1}{2}\log(2\pi).$$

Since $A \ge 2$, we have

$$\log \mathrm{B}\Big(\frac{1}{2}\mathbf{1}\Big) \le \frac{A}{2}(\log\pi + 1) - \frac{A-1}{2}\log\frac{A}{2}. \quad \text{(E.6)}$$

Substituting (E.6) into (E.5), we have

$$\begin{aligned}\mathcal{R}\big(W_T(\cdot)\big) &\le \frac{A-1}{2}\log(W_T+1) + \frac{A}{2} + \frac{A-1}{2}\log\Big(1+\frac{A}{2}\Big) + \frac{1}{12} \\ &\quad + \frac{A}{2}(\log\pi + 1) - \frac{A-1}{2}\log\frac{A}{2} \\ &\le A\log(W_T+1) + 3A,\end{aligned}$$

where we use $\log(1 + A/2) - \log(A/2) \le 2/A$. This completes the proof of Lemma C.1. □

E.2 Proof of Lemma C.2

Proof of Lemma C.2. By conditional Jensen's inequality,

$$\exp\big(\mathbb{E}[Y \mid \mathcal{F}]\big) \le \mathbb{E}[e^Y \mid \mathcal{F}] \le 1.$$

Therefore, $\mathbb{E}[Y \mid \mathcal{F}] \le 0$ and hence $\mu \ge 0$. Moreover, $Y \ge -L$ implies $\mathbb{E}[Y \mid \mathcal{F}] \ge -L$, so $\mu \le L$. We first show that, for every $y \ge -L$ and $\theta \in [0,1]$,

$$e^{\theta y} - 1 - \theta y \le (1+L)\theta^2(e^y - 1 - y). \quad \text{(E.7)}$$

First, suppose that $y \ge 0$. Expanding the exponential using Taylor's series gives

$$\begin{aligned}e^{\theta y} - 1 - \theta y &= \sum_{k=2}^{\infty} \frac{\theta^k y^k}{k!} \\ &\le \theta^2 \sum_{k=2}^{\infty} \frac{y^k}{k!} \\ &= \theta^2(e^y - 1 - y),\end{aligned}$$

where we used $\theta^k \le \theta^2$ for every $k \ge 2$.

It remains to show (E.7) when $-L \le y \le 0$. Let $y = -s$ for some $s \in [0, L]$. Then, on the one hand, the left-hand side of (E.7) satisfies

$$e^{-\theta s} - 1 + \theta s \le \frac{\theta^2 s^2}{2}.$$

On the other hand, the right-hand side of (E.7) satisfies:

$$\begin{aligned}(1+L)\theta^2(e^{-s} - 1 + s) &= (1+L)\theta^2 \int_0^s (1 - e^{-v})\,\mathrm{d}v \\ &\ge (1+L)\theta^2 \int_0^s \frac{v}{1+v}\,\mathrm{d}v \\ &\ge (1+L)\theta^2 \int_0^s \frac{v}{1+L}\,\mathrm{d}v \\ &= \frac{\theta^2 s^2}{2},\end{aligned}$$

where the first inequality holds due to $e^x \ge 1 + x$, and thus $e^{-x} \le 1/(1+x)$ whenever $x \ge 0$. The second inequality holds due to $v \le s \le L$. Therefore, we have proved (E.7) for $y \in [-L, 0]$.
Using (E.7), we obtain

$$\begin{aligned}\mathbb{E}[e^{\theta Y} \mid \mathcal{F}] &= 1 + \theta\mathbb{E}[Y \mid \mathcal{F}] + \mathbb{E}[e^{\theta Y} - 1 - \theta Y \mid \mathcal{F}] \\ &\le 1 - \theta\mu + (1+L)\theta^2 \mathbb{E}[e^Y - 1 - Y \mid \mathcal{F}].\end{aligned} \tag{E.8}$$

Using $\mathbb{E}[e^Y \mid \mathcal{F}] \le 1$, we have

$$\begin{aligned}\mathbb{E}[e^Y - 1 - Y \mid \mathcal{F}] &= \mathbb{E}[e^Y \mid \mathcal{F}] - 1 - \mathbb{E}[Y \mid \mathcal{F}] \\ &\le -\mathbb{E}[Y \mid \mathcal{F}] \\ &= \mu.\end{aligned}$$

Therefore, (E.8) gives

$$\mathbb{E}[e^{\theta Y} \mid \mathcal{F}] \le 1 - \theta\mu + (1+L)\theta^2\mu.$$

Consequently,

$$\begin{aligned}\log \mathbb{E}\left[e^{\theta(Y+\mu)} \,\middle|\, \mathcal{F}\right] &= \theta\mu + \log \mathbb{E}[e^{\theta Y} \mid \mathcal{F}] \\ &\le \theta\mu + \log\left(1 - \theta\mu + (1+L)\theta^2\mu\right) \\ &\le \theta\mu - \theta\mu + (1+L)\theta^2\mu \\ &= (1+L)\theta^2\mu,\end{aligned}$$

where the last inequality uses $\log(1+s) \le s$. This completes the proof of Lemma C.2. □

E.3 Proof of Lemma 4.4

Proof. Fix a level h and a triple $z = (x, u, a)$. Let $\mathcal{F}_t$ denote the history before round t. Define

$$Y_t(z) := \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\big(x_{t,j} = x,\, y_{t,j,<h} = u,\, a_{t,j,h} = a\big)\big[l_{t,j,h} - \log g_h(a|x,u)\big].$$

By the definitions of S_t and N_t, we have

$$\sum_{s=1}^{t} Y_s(z) = \frac{S_t(x,u,a) - N_t(x,u,a)\log g_h(a|x,u)}{m}. \tag{E.9}$$

Under the on-policy protocol, conditionally on $\mathcal{F}_t$, the teacher index is uniform over $\mathcal{I}$, the context follows ρ_{i_t}, and the response is generated by $\widehat{\pi}_t$. Therefore,

$$\mathbb{E}[Y_t(z) \mid \mathcal{F}_t] = \frac{1}{m}\sum_{j=1}^{m} \frac{1}{I}\sum_{i\in\mathcal{I}} \rho_i(x)\widehat{\pi}_t(u|x)\widehat{\pi}_t(a|x,u)\big[\log p_i(a|x,u) - \log g_h(a|x,u)\big]$$

$$= \bar{\rho}(x)\widehat{\pi}_t(u|x)\widehat{\pi}_t(a|x,u)\sum_{i\in\mathcal{I}} w_i(x)\big[\log p_i(a|x,u) - \log g_h(a|x,u)\big]$$
$$= 0,$$

where the second equality uses $\rho_i(x)/I = \bar{\rho}(x)w_i(x)$. The last equality holds due to $\sum_i w_i(x) = 1$ and $\log g_h(a|x,u) = \sum_i w_i(x)\log p_i(a|x,u)$. Using Assumption 4.3, for every teacher i,

$$\big|\log p_i(a|x,u) - \log g_h(a|x,u)\big| = \left|\log\frac{p_i(a|x,u)}{\pi_{\text{ref}}(a|x,u)} - \sum_{k\in\mathcal{I}} w_k(x)\log\frac{p_k(a|x,u)}{\pi_{\text{ref}}(a|x,u)}\right|$$
$$\le B + \sum_{k\in\mathcal{I}} w_k(x)B = 2B.$$

When $\mathbb{1}\big(x_{t,j} = x,\, y_{t,j,<h} = u,\, a_{t,j,h} = a\big) \neq 0$, $l_{t,j,h} = \log p_i(a|x,u)$ for some $i \in \mathcal{I}$. Thus,

$$\big|l_{t,j,h} - \log g_h(a|x,u)\big| \le 2B.$$

Since in the definition of $Y_t(z)$, at most m summands can be nonzero. Thus, $|Y_t(z)| \le 2Bm/m = 2B$. Moreover, Jensen's inequality gives

$$Y_t(z)^2 \le \frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\big(x_{t,j} = x,\, y_{t,j,<h} = u,\, a_{t,j,h} = a\big)\big[l_{t,j,h} - \log g_h(a|x,u)\big]^2$$
$$\le \frac{4B^2}{m}\sum_{j=1}^{m}\mathbb{1}\big(x_{t,j} = x,\, y_{t,j,<h} = u,\, a_{t,j,h} = a\big).$$

Taking the conditional expectation, we obtain

$$\mathrm{Var}(Y_t(z) \mid \mathcal{F}_t) = \mathbb{E}[Y_t(z)^2 \mid \mathcal{F}_t]$$
$$\le \frac{4B^2}{m}\mathbb{E}\left[\sum_{j=1}^{m}\mathbb{1}\big(x_{t,j} = x,\, y_{t,j,<h} = u,\, a_{t,j,h} = a\big)\,\middle|\,\mathcal{F}_t\right]. \tag{E.10}$$

Define $V_t(z) := \sum_{s=1}^{t}\mathrm{Var}(Y_s(z) \mid \mathcal{F}_s)$. Since $\mathbb{E}[Y_s(z) \mid \mathcal{F}_s] = 0$ and $|Y_s(z)| \le 2B$, Lemma J.4 implies that both

$$\left(\sum_{s=1}^{t} Y_s(z),\, V_t(z)\right) \quad \text{and} \quad \left(-\sum_{s=1}^{t} Y_s(z),\, V_t(z)\right)$$

are sub-gamma processes with parameter $2B/3$. Let $K := S\sum_{h=1}^{H} A^h$. Then, applying Lemma J.5 with $\rho = 2B$, with probability at least $1 - \delta/(2K)$, we have

$$\sum_{s=1}^{t} Y_s(z) \le 4\sqrt{V_t(z)\log(2H_tK/\delta)} + \frac{88}{3}B\log(2H_tK/\delta), \tag{E.11}$$

where we can choose $H_t = e + \log(1+T)$ since $V_t(z) \le 4B^2T$. Applying the same inequality to $-Y_s(z)$, and taking the union bound, we obtain that with probability at least $1-\delta$, the following inequality holds for any h and any $z = (x,u,a)$:

$$\left|\sum_{s=1}^{t} Y_s(z)\right| \le 4\sqrt{V_t(z)\log(2H_tK/\delta)} + \frac{88}{3}B\log(2H_tK/\delta).$$

Since

$$\frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\big(x_{s,j} = x,\, y_{s,j,<h} = u,\, a_{s,j,h} = a\big) \in [0,1],$$

we can apply Lemma J.6 and obtain that for any h and $z = (x,u,a)$, with probability at least $1 - \delta/K$, the following inequality holds simultaneously for all $t \in [T]$:

$$\sum_{s=1}^{t}\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\big(x_{s,j} = x,\, y_{s,j,<h} = u,\, a_{s,j,h} = a\big)\,\middle|\,\mathcal{F}_s\right]$$

$$\leq 2\sum_{s=1}^{t}\frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\big(x_{s,j}=x,\, y_{s,j,<h}=u,\, a_{s,j,h}=a\big)+8\log(2K/\delta). \tag{E.12}$$

Taking a union bound over h and (x,u,a), the inequality holds for all h and (x,u,a) simultaneously. Combining (E.10) and (E.12), we have

$$\begin{aligned}V_t(z)&\leq\frac{8B^2}{m}\sum_{s=1}^{t}\sum_{j=1}^{m}\mathbb{1}\big(x_{s,j}=x,\, y_{s,j,<h}=u,\, a_{s,j,h}=a\big)+32B^2\log(2K/\delta)\\&=\frac{8B^2N_t(x,u,a)}{m}+32B^2\log(2K/\delta).\end{aligned} \tag{E.13}$$

Substituting (E.13) into (E.11), we have, with probability at least $1-2\delta$, simultaneously for every $t\in[T]$, h, and $z=(x,u,a)$, the following inequality holds

$$\begin{aligned}\left|\sum_{s=1}^{t}Y_s(z)\right|&\leq 4\sqrt{\Big(\frac{8B^2N_t(x,u,a)}{m}+32B^2\log\frac{2K}{\delta}\Big)\log\frac{2H_tK}{\delta}}+\frac{88}{3}B\log\frac{2H_tK}{\delta}\\&\lesssim B\sqrt{\frac{N_t(x,u,a)}{m}\log\frac{2H_tK}{\delta}}+B\log\frac{2H_tK}{\delta},\end{aligned}$$

where we use $\sqrt{v+w}\leq\sqrt{v}+\sqrt{w}$ and $\log(2K/\delta)\leq\log(2H_tK/\delta)$.
For every triple with $N_t(x,u,a)\geq 1$, (E.9) gives

$$\begin{aligned}\big|\bar{l}_t(x,u,a)-\log g_h(a|x,u)\big|&=\frac{m}{N_t(x,u,a)}\left|\sum_{s=1}^{t}Y_s(z)\right|\\&\lesssim B\sqrt{\frac{m}{N_t(x,u,a)}\log\frac{2H_tK}{\delta}}+\frac{Bm}{N_t(x,u,a)}\log\frac{2H_tK}{\delta}.\end{aligned}$$

Finally, since $K\leq SHA^H$ and $H_t=e+\log(1+T)$,

$$\log\frac{2H_tK}{\delta}\leq H\log A+\log\left[\frac{2SH\big(e+\log(1+T)\big)}{\delta}\right].$$

Hiding the logarithmic factors, we can define

$$\beta_t(x,u,a)=\widetilde{O}\bigg(B\sqrt{\frac{mH}{N_t(x,u,a)}}+\frac{BmH}{N_t(x,u,a)}\bigg).$$

and conclude that

$$\big|\bar{l}_t(x,u,a)-\log g_h(a|x,u)\big|\leq\beta_t(x,u,a).$$

This completes the proof of Lemma 4.4. □

F Missing Proof in Section 5

Proof of Proposition 5.1. Fix $N\geq 2$ and $\alpha\in(0,1)$, and let $b:=1/N$. For $q\in(b,1)$, write the correct-response probabilities as

$$\begin{aligned}F(q)&:=\pi^*_{\mathrm{F}}(y^\star|x)=\alpha q+(1-\alpha)b,\\R(q)&:=\pi^*_{\mathrm{R}}(y^\star|x)=\frac{q^\alpha}{q^\alpha+(N-1)^{1-\alpha}(1-q)^\alpha}.\end{aligned}$$

Both probabilities lie in $(0,1)$. Since the log-odds function is strictly increasing, $F(q)-R(q)$ has the same sign as

$$\begin{aligned}G(q)&:=\log\frac{F(q)}{1-F(q)}-\log\frac{R(q)}{1-R(q)}\\&=\log\frac{F(q)}{1-F(q)}-\alpha\log\frac{q}{1-q}+(1-\alpha)\log(N-1).\end{aligned}$$

The function G extends continuously to $q = b$, with $G(b) = 0$. Differentiating gives

$$G'(q) = \frac{\alpha}{F(q)(1-F(q))} - \frac{\alpha}{q(1-q)} = \frac{\alpha(q-F(q))(1-q-F(q))}{F(q)(1-F(q))q(1-q)}.$$

For $q > b$, we have $q - F(q) = (1-\alpha)(q-b) > 0$. Thus, the sign of $G'(q)$ is the sign of

$$1 - q - F(q) = 1 - (1+\alpha)q - (1-\alpha)b,$$

which vanishes at $q_{\mathrm{m}} := [1-(1-\alpha)b]/(1+\alpha)$. If $N \geq 3$, then $b < q_{\mathrm{m}} < 1$, so G is strictly increasing on (b, q_{m}) and strictly decreasing on $(q_{\mathrm{m}}, 1)$. In particular, $G(q_{\mathrm{m}}) > 0$. Moreover, $F(q) \to \alpha + (1-\alpha)b \in (0,1)$ as $q \uparrow 1$, and hence $G(q) \to -\infty$. The intermediate value theorem and strict decrease on $(q_{\mathrm{m}}, 1)$ imply that there is exactly one zero $q_{\mathrm{c}} \in (q_{\mathrm{m}}, 1)$. We have $G(q) > 0$ for $b < q < q_{\mathrm{c}}$ and $G(q) < 0$ for $q_{\mathrm{c}} < q < 1$, giving the claimed ordering.
If $N = 2$, then $q_{\mathrm{m}} = b = 1/2$, so $G'(q) < 0$ throughout $(1/2, 1)$. Since $G(b) = 0$, this gives $G(q) < 0$ and therefore $R(q) > F(q)$ on that interval. Finally, both targets are below the expert probability for $q > b$: we have $F(q) < q$, and

$$\frac{R(q)}{1-R(q)} = \frac{q}{1-q}\left(\frac{1-q}{(N-1)q}\right)^{1-\alpha} < \frac{q}{1-q},$$

which implies $R(q) < q$. □

Proof of Proposition 5.3. Fix a context x and suppress its dependence in the notation. Since the forward target is the weighted arithmetic mixture of the teachers, summing over E gives

$$\pi^*_{\mathrm{F}}(E) = \sum_i w_i p_i(E) \geq w_k p_k(E),$$

where the inequality follows from the nonnegativity of every summand.
For the reverse target, fix $\alpha, q, \eta \in (0,1)$ and write $m := |E|$ and $n := |\mathcal{Y} \setminus E|$. Both m and n are positive. For $\varepsilon \in (0,1)$, define two teachers by

$$p_1(y) = \begin{cases} q/m, & y \in E, \\ (1-q)/n, & y \notin E, \end{cases} \qquad p_2(y) = \begin{cases} \varepsilon/m, & y \in E, \\ (1-\varepsilon)/n, & y \notin E. \end{cases}$$

These distributions have full support, and $p_1(E) = q$. With weights α and $1-\alpha$, the unnormalized geometric masses on the two sets are

$$\sum_{y \in E} p_1(y)^\alpha p_2(y)^{1-\alpha} = q^\alpha \varepsilon^{1-\alpha},$$
$$\sum_{y \notin E} p_1(y)^\alpha p_2(y)^{1-\alpha} = (1-q)^\alpha (1-\varepsilon)^{1-\alpha}.$$

Consequently,

$$\pi^*_{\mathrm{R}}(E) = \frac{q^\alpha \varepsilon^{1-\alpha}}{q^\alpha \varepsilon^{1-\alpha} + (1-q)^\alpha (1-\varepsilon)^{1-\alpha}}.$$

As $\varepsilon \downarrow 0$, the numerator tends to zero and the denominator tends to $(1-q)^\alpha > 0$. Hence $\pi^*_{\mathrm{R}}(E) \to 0$, and a sufficiently small positive ε ensures $\pi^*_{\mathrm{R}}(E) < \eta$. This choice retains full support for both teachers and proves the claim. □

Proof of Proposition 5.5. Fix $\beta \in (0, 1/2)$ and $r \in (1/2, 1)$, and set $\alpha := 1 - \beta$. We construct the expert using token probabilities that do not depend on the horizon. Fix $\delta \in (0, 1/2)$, let $p_1(a|x) = r$, and, at every subsequent position, set

$$p_1(a|x,u) = \begin{cases} 1-\delta, & \text{if the first token of } u \text{ is } a, \\ 1/2, & \text{if the first token of } u \text{ is } b. \end{cases}$$

The probability of b is the complementary probability at every prefix. Let p_2 assign probability $1/2$ to each token at every prefix. Both teachers assign positive probability to every response in $\{a, b\}^H$ for every finite H.

Marginalizing the forward mixture over the last $H-1$ tokens yields

$$\pi^*_{\mathrm{F}}(a|x) = \alpha r + \frac{\beta}{2} = \frac{1}{2} + \alpha\left(r - \frac{1}{2}\right) > \frac{1}{2}.$$

Thus $\pi^*_{\mathrm{F}}(a|x) > \pi^*_{\mathrm{F}}(b|x)$ for every H.

For the reverse target, the uniform teacher contributes the same factor $2^{-\beta H}$ to every complete response, so

$$\pi^*_{\mathrm{R}}(y|x) = \frac{p_1(y|x)^\alpha}{\sum_{z\in\{a,b\}^H} p_1(z|x)^\alpha}.$$

Define $C_\alpha := (1-\delta)^\alpha + \delta^\alpha$ and $D_\alpha := 2^{1-\alpha}$. Summing the powered expert probabilities over continuations gives

$$\sum_{v\in\{a,b\}^{H-1}} p_1(av|x)^\alpha = r^\alpha C_\alpha^{H-1},$$
$$\sum_{v\in\{a,b\}^{H-1}} p_1(bv|x)^\alpha = (1-r)^\alpha D_\alpha^{H-1}.$$

Since $0<\alpha<1$, the function $t\mapsto t^\alpha$ is strictly concave. As $\delta \neq 1/2$, strict concavity implies

$$C_\alpha = (1-\delta)^\alpha + \delta^\alpha < 2\left(\frac{1}{2}\right)^\alpha = D_\alpha.$$

The reverse target's first-token odds therefore satisfy

$$\frac{\pi^*_{\mathrm{R}}(a|x)}{\pi^*_{\mathrm{R}}(b|x)} = \left(\frac{r}{1-r}\right)^\alpha \left(\frac{C_\alpha}{D_\alpha}\right)^{H-1}.$$

In particular, these odds are strictly less than one whenever

$$H > 1 + \frac{\alpha \log(r/(1-r))}{\log(D_\alpha/C_\alpha)}.$$

The denominator is strictly positive, so this bound is finite for every fixed β and r in the stated ranges. The same construction thus satisfies $\pi^*_{\mathrm{R}}(a|x) < \pi^*_{\mathrm{R}}(b|x)$ for every sufficiently large integer H, completing the proof. □

G Parameter Sensitivity of Aggregation Targets

We extend the examples in Section 5 by varying the teacher weights, expert confidence, response-space size, and continuation distributions. All curves are deterministic evaluations of the closed-form targets, rather than results of training a student. Each comparison holds the other parameters fixed. The vertical axes show the full probability range, and dashed lines indicate the expert's probability.

Uninformative teachers. The expert has weight α, assigns probability q to the correct response, and distributes its remaining mass uniformly over the $N-1$ incorrect responses. The other teachers are uniform and have total weight $1-\alpha$. Writing $F(q) := \pi^*_{\mathrm{F}}(y^\star|x)$ and $R(q) := \pi^*_{\mathrm{R}}(y^\star|x)$, we evaluate

$$F(q) = \alpha q + \frac{1-\alpha}{N}, \qquad R(q) = \frac{q^\alpha}{q^\alpha + (N-1)^{1-\alpha}(1-q)^\alpha}.$$

Here N counts responses; splitting the uniform teachers' total weight among more identical teachers leaves both targets unchanged.

Figure 2 shows that the region in which reverse aggregation retains more correct-response probability depends on both parameters. For the weights shown at $N = 10$, increasing α lowers the confidence threshold, although the targets become close as both approach the expert. At $\alpha = 0.9$, increasing N from 10 to 1000 raises the threshold, narrowing the range of q with a reverse advantage. The binary case differs: reverse aggregation is larger throughout $q \in (1/2, 1)$, with no interior crossing.

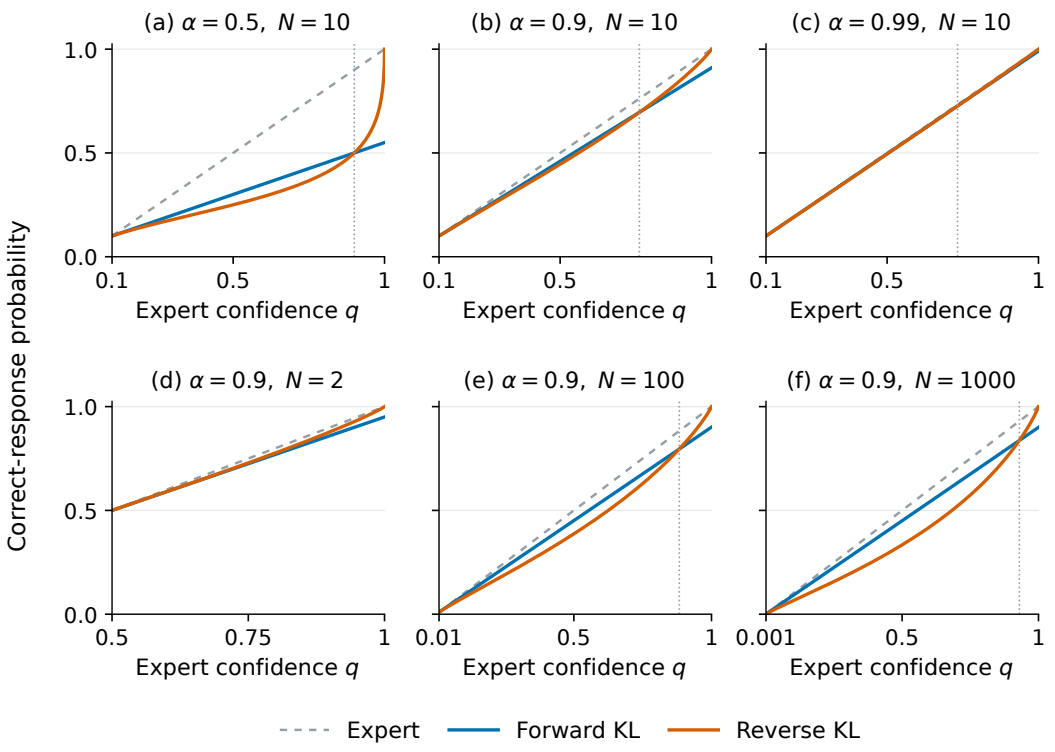

Figure 2: Sensitivity to expert weight and response-space size with uninformative teachers. Panels (a–c) vary $\alpha \in \{0.5, 0.9, 0.99\}$ at $N = 10$. Panels (d–f), together with (b), compare $N \in \{2, 10, 100, 1000\}$ at $\alpha = 0.9$. Each curve varies q from $1/N$ toward 1. Dotted vertical lines mark the interior crossing when it exists.

Misleading teachers. We fix $N = 10$ and replace the uniform teacher by a teacher of weight $1 - \alpha$ that assigns probability ε to the correct response. Both teachers distribute their remaining probability uniformly over the incorrect responses. The targets are

$$F(\varepsilon) = \alpha q + (1 - \alpha)\varepsilon, \qquad R(\varepsilon) = \frac{q^\alpha \varepsilon^{1-\alpha}}{q^\alpha \varepsilon^{1-\alpha} + (1-q)^\alpha (1-\varepsilon)^{1-\alpha}}.$$

Figure 3 varies the expert weight at fixed $q = 0.99$, and its confidence at fixed $\alpha = 0.9$. The horizontal coordinate is $\log_{10} \varepsilon$; we evaluate the reverse target in the log domain to retain accuracy for extremely small probabilities.

Higher expert weight or confidence shifts the decline in reverse probability to smaller values of ε. This protection can be substantial: at $q = 0.99$, the reverse probability equals $1/2$ at approximately $\varepsilon = 10^{-18}$ for $\alpha = 0.9$, compared with 10^{-198} for $\alpha = 0.99$. Nevertheless, every fixed $\alpha < 1$ permits suppression as $\varepsilon \downarrow 0$, whereas forward aggregation remains above αq. The extreme ranges illustrate the absence of a uniform positive lower bound for reverse aggregation; they do not estimate how often such predictions arise in practice.

Long horizons. The expert selects the correct first token a with probability r. At every subsequent position, its probabilities for (a, b) are $(1 - \delta, \delta)$ after an initial a and $(1/2, 1/2)$ after an initial b. The other teacher is uniform at every prefix. With expert weight α, let $C_\alpha = (1-\delta)^\alpha + \delta^\alpha$ and $D_\alpha = 2^{1-\alpha}$. The first-token probabilities are

$$F_H = \alpha r + \frac{1-\alpha}{2}, \qquad R_H = \left[1 + \left(\frac{1-r}{r}\right)^\alpha \left(\frac{D_\alpha}{C_\alpha}\right)^{H-1}\right]^{-1}.$$

Figure 4 varies α, r, and δ separately around $(\alpha, r, \delta) = (0.9, 0.99, 0.01)$, evaluating every integer H from 1 to 1000.

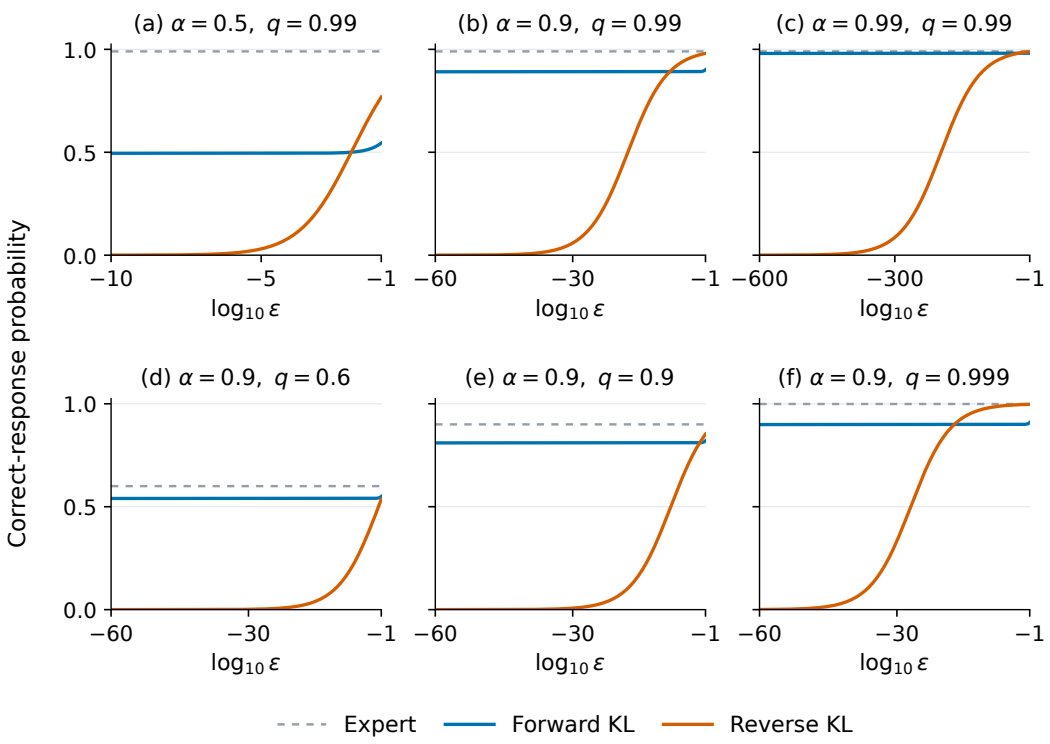

Figure 3: Sensitivity to expert weight and confidence with misleading teachers. Panels (a–c) fix $q = 0.99$ and vary α. Panels (d–f), together with (b), fix $\alpha = 0.9$ and compare $q \in \{0.6, 0.9, 0.99, 0.999\}$. Horizontal ranges differ: $\log_{10}\varepsilon \in [-10, -1]$ in (a), $[-600, -1]$ in (c), and $[-60, -1]$ otherwise. Moving left makes the other teacher more misleading.

Increasing the expert's weight or first-token confidence delays the reversal in the displayed settings. At $r = 0.99$ and $\delta = 0.01$, the first horizon with $R_H < 1/2$ is 10, 68, or 717 for $\alpha = 0.5$, 0.9, or 0.99, respectively. Making the continuations after a more diffuse also delays reversal. For $\delta = 0.3$, the reverse probability first falls below $1/2$ at $H = 555$, compared with $H = 68$ at $\delta = 0.01$. Forward aggregation is independent of H in every panel. Thus, in this construction, the deterioration arises from the difference between the continuation distributions, while expert confidence and weight determine how long the initial preference persists.

H Forward KL with function approximation

The tabular estimator learns a separate conditional distribution at each prefix. Function approximation allows these conditionals to share a common representation. Let $\Pi = \{\pi_\theta : \theta \in \Theta\}$ be a finite class of autoregressive policies, where

$$\pi_\theta(y|x) = \prod_{h=1}^{H} \pi_\theta(a_h|x, y_{<h}).$$

We make the standard realizablity assumption for the function class.

Assumption H.1. There exists $\theta^* \in \Theta$ such that, for every context x and every response $y \in \mathcal{Y}$,

$$\pi_{\theta^*}(y|x) = \pi^*_{\mathrm{F}}(y|x).$$

H.1 Algorithm Design

We aggregate the candidate policies at the token level, using sampled tokens when logits are unavailable and full teacher probabilities otherwise. Recall the token feedback $q_{t,j,h}(a)$ defined in Section 3:

$$q_{t,j,h}(a) := \begin{cases} \mathbb{1}(a_{t,j,h} = a) & \text{w/o logit,} \\ \frac{\exp(Z_{t,j,h}(a))}{\sum_{b \in \mathcal{A}} \exp(Z_{t,j,h}(b))}, & \text{w/ logit.} \end{cases}$$

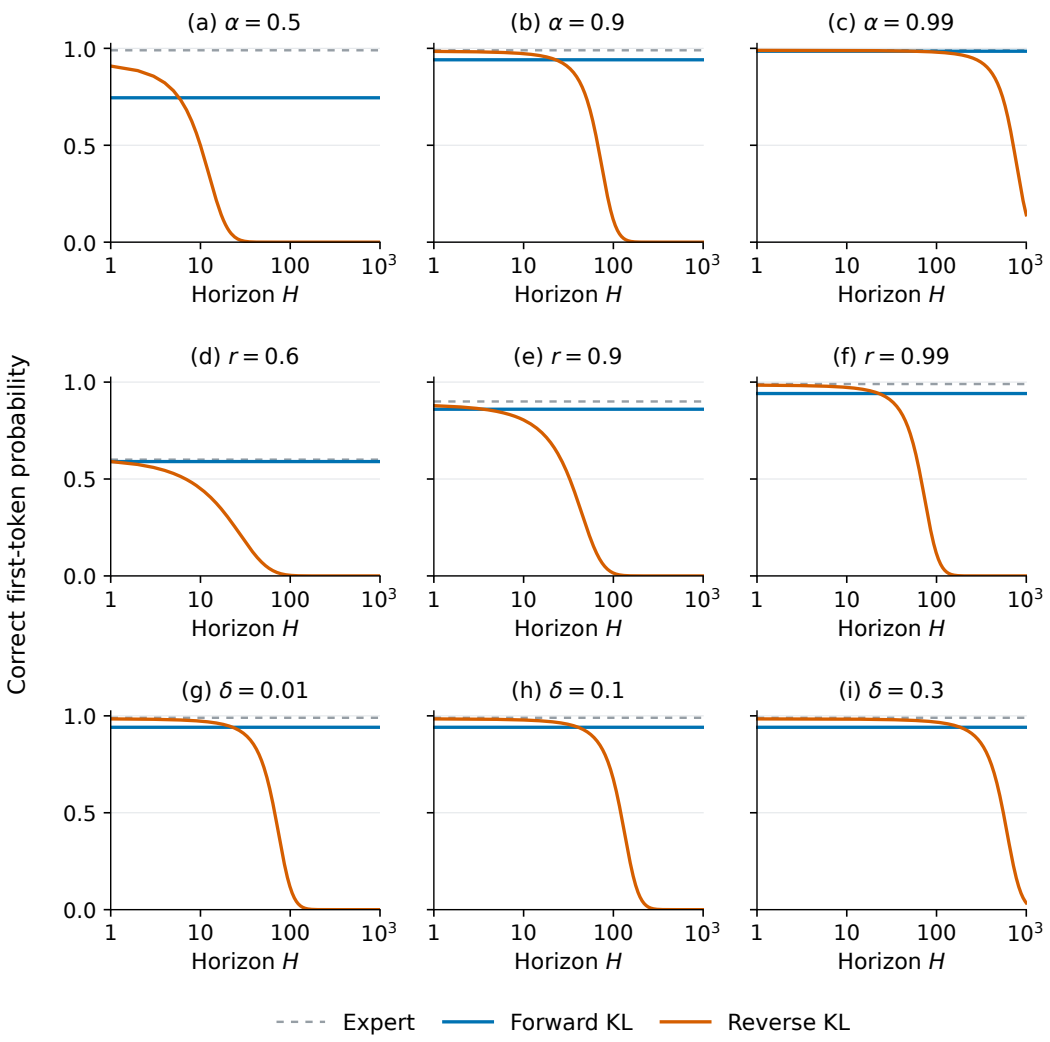

Figure 4: Sensitivity of first-token probabilities to the horizon. Top row: vary α with $r = 0.99$ and $\delta = 0.01$. Middle row: vary r with $\alpha = 0.9$ and $\delta = 0.01$. Bottom row: vary $\delta \in \{0.01, 0.1, 0.3\}$ with $\alpha = 0.9$ and $r = 0.99$. The baseline is repeated in (b), (f), and (g). The horizontal axis is logarithmic; curves connect values computed at integer horizons.

For each $\theta \in \Theta$, define its empirical loss in round t by

$$\ell_t(\pi_\theta) := -\frac{1}{m} \sum_{j=1}^{m} \sum_{h=1}^{H} \sum_{a \in \mathcal{A}} q_{t,j,h}(a) \log \pi_\theta(a | x_{t,j}, y_{t,j,<h}). \tag{H.1}$$

The main idea of the algorithm is to maintain a distribution $p_t(\theta)$ over the parameter space Θ. More specifically, choose a prior distribution p_0 on Θ with $p_0(\theta) > 0$ for every $\theta \in \Theta$, and initialize $L_0(\theta) = 0$. At the beginning of round t, for any prefix (x, u), the learner uses

$$\widehat{\pi}_t(a | x, u) := \sum_{\theta \in \Theta} p_{t-1}(\theta) \pi_\theta(a | x, u). \tag{H.2}$$

After observing the feedback, the learner updates

$$p_t(\theta) := \frac{p_{t-1}(\theta) \exp\big(-\ell_t(\pi_\theta)/H\big)}{\sum_{\theta' \in \Theta} p_{t-1}(\theta') \exp\big(-\ell_t(\pi_{\theta'})/H\big)}. \tag{H.3}$$

The factor $1/H$ normalizes the sum of the H token losses in each rollout. Intuitively, the update favors candidates with smaller cumulative empirical losses, relative to their prior weights.

Algorithm 3 Forward KL with Function Approximation

1: **Input:** Finite policy class $\Pi = \{\pi_\theta : \theta \in \Theta\}$, prior p_0 with $p_0(\theta) > 0$ for every θ, number of rounds T.
2: **for** $t = 1, \dots, T$ **do**
3: For any prefix (x, u), construct the policy $\widehat{\pi}_t(a|x, u)$ as in (H.2).
4: Observe the batch of teacher-generated rollouts $\{(x_{t,j}, y_{t,j})\}_{j=1}^m$ and, when available, the teacher logits $\{Z_{t,j,h}\}_{j,h}$.
5: For every $\theta \in \Theta$, compute $\ell_t(\pi_\theta)$ as in (H.1)
6: For every $\theta \in \Theta$, update the mixture weights $p_t(\theta)$ as in (H.3)
7: **end for**
8: **Output:** $\{\widehat{\pi}_t\}_{t=1}^T$.

This exponential reweighting resembles the posterior updates used in Thompson sampling. Unlike Thompson sampling, however, our algorithm predicts with a mixture of the candidate policies rather than a single sampled candidate. Evaluating this mixture at each decoding step introduces large computational cost, which is a key limitation of our forward-KL algorithm with function approximation.

H.2 Theoretical guarantee.

In this section, we first work on a finite function class Θ. We will apply the results to the setting with finite covering number later. The following theorem bounds the regret by the prior weight of a good candidate and its approximation error. Under realizability, the approximation term vanishes.

Theorem H.2. Let $T \geq 2$. Suppose the regret in (2.2) is defined with $D(p\|q) = \mathrm{KL}(q\|p)$. Under off-policy distillation protocol and the function approximation setting, both with and without access to teacher logits, Algorithm 3 satisfies

$$\mathbb{E}\big[\,\mathrm{Regret}(T)\big] \leq \inf_{\theta\in\Theta}\left\{\log\frac{1}{p_0(\theta)} + T\mathbb{E}_{x\sim\bar{\rho}}\mathrm{KL}\big(\pi^*_{\mathrm{F}}(\cdot|x)\,\big\|\,\pi_\theta(\cdot|x)\big)\right\},$$

provided the right-hand side is finite. In particular, under Assumption H.1, for the optimal parameter θ^*, Algorithm 3 satisfies

$$\mathbb{E}\big[\,\mathrm{Regret}(T)\big] \leq \log\frac{1}{p_0(\theta^*)}.$$

Moreover, for any $\delta > 0$, under Assumption H.1, the following inequality holds with probability at least $1 - \delta$:

$$\mathrm{Regret}(T) \leq 2H\log\frac{1}{p_0(\theta^*)} + 4H\left(1 + \log\frac{2}{\delta p_0(\theta^*)}\right)\log\frac{2}{\delta}.$$

Proof. Let $\mathcal{F}_t$ be the σ-algebra generated by all rollouts and feedback observed before round t:

$$\mathcal{F}_t := \sigma\Big((x_{s,j}, y_{s,j}), q_{s,j,h}(a) : s < t,\ j \in [m],\ h \in [H],\ a \in \mathcal{A}\Big).$$

Thus, both q_{t-1} and $\widehat{\pi}_t$ are $\mathcal{F}_t$-measurable. Recall the feedback counts $c_t(x, u, a)$ defined in (3.3). For any policy π, (H.1) can be rewritten as

$$\ell_t(\pi) = -\sum_{h=1}^{H}\sum_{(x,u,a)} c_t(x, u, a)\log\pi(a|x, u). \tag{H.4}$$

Easy to see, the following equation holds

$$\sum_{(x,u,a)} c_t(x, u, a) = 1$$

Therefore, we have

$$\sum_{h=1}^{H}\sum_{(x,u,a)}\frac{c_t(x, u, a)}{H} = 1. \tag{H.5}$$

For any $t \geq 1$, define the normalizing constant

$$C_t := \sum_{\theta \in \Theta} p_0(\theta) \exp\Big(-\frac{1}{H} \sum_{s=1}^{t} \ell_s(\pi_\theta) \Big),$$

and $C_0 = 1$. Direct iteration of (H.3) gives

$$p_t(\theta) = \frac{p_0(\theta) \exp\big(-H^{-1} \sum_{s=1}^{t} \ell_s(\pi_\theta) \big)}{C_t}.$$

Moreover, we have

$$\begin{aligned}
\frac{C_t}{C_{t-1}} &= \frac{1}{C_{t-1}} \sum_{\theta' \in \Theta} p_0(\theta') \exp\Big(-\frac{1}{H} \sum_{s=1}^{t-1} \ell_s(\pi_{\theta'}) - \frac{\ell_t(\pi_{\theta'})}{H} \Big) \\
&= \sum_{\theta' \in \Theta} \underbrace{\frac{p_0(\theta') \exp\big(-H^{-1} \sum_{s=1}^{t-1} \ell_s(\pi_{\theta'}) \big)}{C_{t-1}}}_{= p_{t-1}(\theta')} \exp\big(-\ell_t(\pi_{\theta'})/H\big) \\
&= \sum_{\theta' \in \Theta} p_{t-1}(\theta') \exp\big(-\ell_t(\pi_{\theta'})/H\big). \qquad \text{(H.6)}
\end{aligned}$$

The definition of ℓ_t in (H.4) gives

$$\begin{aligned}
\exp\big(-\ell_t(\pi_{\theta'})/H\big) &= \exp\Big(\sum_{h=1}^{H} \sum_{(x,u,a)} \frac{c_t(x,u,a)}{H} \log \pi_{\theta'}(a|x,u) \Big) \\
&= \prod_{h=1}^{H} \prod_{(x,u,a)} \pi_{\theta'}(a|x,u)^{c_t(x,u,a)/H}. \qquad \text{(H.7)}
\end{aligned}$$

Substituting (H.7) into (H.6), we have

$$\frac{C_t}{C_{t-1}} = \sum_{\theta' \in \Theta} p_{t-1}(\theta') \prod_{h=1}^{H} \prod_{(x,u,a)} \pi_{\theta'}(a|x,u)^{c_t(x,u,a)/H}. \qquad \text{(H.8)}$$

Using the generalized Hölder's inequality, we have

$$\begin{aligned}
\frac{C_t}{C_{t-1}} &\leq \prod_{h=1}^{H} \prod_{(x,u,a)} \Big(\sum_{\theta' \in \Theta} p_{t-1}(\theta') \pi_{\theta'}(a|x,u) \Big)^{c_t(x,u,a)/H} \\
&= \prod_{h=1}^{H} \prod_{(x,u,a)} \widehat{\pi}_t(a|x,u)^{c_t(x,u,a)/H} \\
&= \exp\big(-\ell_t(\widehat{\pi}_t)/H\big),
\end{aligned}$$

where the last equation holds using (H.4).
Taking the negative logarithm, and summing over $t \in [T]$ yields

$$\sum_{t=1}^{T} \ell_t(\widehat{\pi}_t) \leq -H \sum_{t=1}^{T} \log \frac{C_t}{C_{t-1}} = -H \log C_T, \qquad \text{(H.9)}$$

where the equality uses $C_0 = 1$.
Now fix any comparator policy π_θ such that $\theta \in \Theta$. Easy to see

$$\begin{aligned}
C_T &= \sum_{\theta'} p_0(\theta') \exp\Big(-\frac{1}{H} \sum_{t=1}^{T} \ell_t(\pi_{\theta'}) \Big) \\
&\geq p_0(\theta) \exp\Big(-\frac{1}{H} \sum_{t=1}^{T} \ell_t(\pi_\theta) \Big).
\end{aligned}$$

Therefore, we have

$$-\sum_{t=1}^{T}\ell_t(\pi_\theta) \le H\log C_T + H\log\frac{1}{p_0(\theta)}. \tag{H.10}$$

Summing up (H.9) and (H.10), we have

$$\sum_{t=1}^{T}\big[\ell_t(\widehat{\pi}_t) - \ell_t(\pi_\theta)\big] \le H\log\frac{1}{p_0(\theta)}. \tag{H.11}$$

As a result,

$$\begin{aligned}\sum_{t=1}^{T}\big[\ell_t(\widehat{\pi}_t) - \ell_t(\pi_{\mathrm{F}}^*)\big] &= \sum_{t=1}^{T}\big[\ell_t(\widehat{\pi}_t) - \ell_t(\pi_\theta)\big] + \sum_{t=1}^{T}\big[\ell_t(\pi_\theta) - \ell_t(\pi_{\mathrm{F}}^*)\big] \\ &\le H\log\frac{1}{p_0(\theta)} + \sum_{t=1}^{T}\big[\ell_t(\pi_\theta) - \ell_t(\pi_{\mathrm{F}}^*)\big].\end{aligned} \tag{H.12}$$

Using the same argument as in Appendix C, we have

$$\begin{aligned}\mathbb{E}[c_t(x,u,a)\mid\mathcal{F}_t] &= \frac{1}{I}\sum_{i\in\mathcal{I}}\rho_i(x)p_i(u|x)p_i(a|x,u) \\ &= \bar{\rho}(x)\pi_{\mathrm{F}}^*(u|x)\pi_{\mathrm{F}}^*(a|x,u).\end{aligned}$$

Consequently, using (H.4), for any $\mathcal{F}_t$-measurable policy π with finite loss,

$$\begin{aligned}&\mathbb{E}\big[\ell_t(\pi) - \ell_t(\pi_{\mathrm{F}}^*)\mid\mathcal{F}_t\big] \\ &= \sum_{h=1}^{H}\sum_{(x,u,a)}\mathbb{E}[c_t(x,u,a)\mid\mathcal{F}_t]\log\frac{\pi_{\mathrm{F}}^*(a|x,u)}{\pi(a|x,u)} \\ &= \sum_{h=1}^{H}\sum_{(x,u)}\bar{\rho}(x)\pi_{\mathrm{F}}^*(u|x)\mathrm{KL}\big(\pi_{\mathrm{F}}^*(\cdot|x,u)\,\big\|\,\pi(\cdot|x,u)\big) \\ &= \mathbb{E}_{x\sim\bar{\rho}}\mathrm{KL}\big(\pi_{\mathrm{F}}^*(\cdot|x)\,\big\|\,\pi(\cdot|x)\big).\end{aligned} \tag{H.13}$$

The last equality follows from the chain rule for KL divergence.
Taking expectations in (H.12) and applying (H.13) to both $\widehat{\pi}_t$ and π_θ, we conclude that

$$\begin{aligned}\mathbb{E}[\mathrm{Regret}(T)] &\le H\log\frac{1}{p_0(\theta)} + \sum_{t=1}^{T}\mathbb{E}_{x\sim\bar{\rho}}\mathrm{KL}\big(\pi_{\mathrm{F}}^*(\cdot|x)\,\big\|\,\pi_\theta(\cdot|x)\big) \\ &= H\log\frac{1}{p_0(\theta)} + T\mathbb{E}_{x\sim\bar{\rho}}\mathrm{KL}\big(\pi_{\mathrm{F}}^*(\cdot|x)\,\big\|\,\pi_\theta(\cdot|x)\big).\end{aligned}$$

Since the inequality holds for every fixed θ, taking the infimum gives

$$\mathbb{E}[\mathrm{Regret}(T)] \le \inf_{\theta\in\Theta}\left\{H\log\frac{1}{p_0(\theta)} + T\mathbb{E}_{x\sim\bar{\rho}}\mathrm{KL}\big(\pi_{\mathrm{F}}^*(\cdot|x)\,\big\|\,\pi_\theta(\cdot|x)\big)\right\}. \tag{H.14}$$

High-probability bound under realizability. Suppose $\pi_{\theta^*} = \pi_{\mathrm{F}}^*$ for some $\theta^*\in\Theta$. As in Appendix C, define

$$\begin{aligned}X_t &:= \ell_t(\pi_{\mathrm{F}}^*) - \ell_t(\widehat{\pi}_t), \\ r_t &:= -\mathbb{E}[X_t\mid\mathcal{F}_t].\end{aligned}$$

By (H.13), $\mathrm{Regret}(T) = \sum_{t=1}^{T} r_t$. The argument establishing (C.20) applies to any $\mathcal{F}_t$-measurable policy π, since it only uses policy normalization and the conditional mean of the feedback counts. Therefore,

$$\mathbb{E}\Big[\exp\Big(\frac{X_t}{H}\Big)\Big|\mathcal{F}_t\Big] = \mathbb{E}\left[\exp\left(\frac{\ell_t(\pi_{\mathrm{F}}^*) - \ell_t(\widehat{\pi}_t)}{H}\right)\,\middle|\,\mathcal{F}_t\right] \le 1. \tag{H.15}$$

To replace the lower bound provided by KT smoothing in the tabular proof, we control the mixture weight of θ^*. Define $W_0 = 1$ and

$$W_t := C_t \exp\left(\frac{1}{H}\sum_{s=1}^{t} \ell_s(\pi_{\mathrm{F}}^*)\right) = \frac{p_0(\theta^*)}{p_t(\theta^*)}.$$

The equality follows from the formula for $p_t(\theta^*)$ and realizability. By (H.6),

$$\begin{aligned}\frac{W_t}{W_{t-1}} &= \frac{C_t}{C_{t-1}} \exp\big(\ell_t(\pi_{\mathrm{F}}^*)/H\big) \\ &= \sum_{\theta\in\Theta} p_{t-1}(\theta) \exp\left(\frac{\ell_t(\pi_{\mathrm{F}}^*) - \ell_t(\pi_\theta)}{H}\right),\end{aligned}$$

where we apply (H.6). Since W_{t-1} and p_{t-1} are $\mathcal{F}_t$-measurable, (H.15) implies

$$\begin{aligned}\mathbb{E}[W_t \mid \mathcal{F}_t] &= W_{t-1} \sum_{\theta\in\Theta} p_{t-1}(\theta)\mathbb{E}\left[\exp\left(\frac{\ell_t(\pi_{\mathrm{F}}^*) - \ell_t(\pi_\theta)}{H}\right)\,\middle|\,\mathcal{F}_t\right] \\ &\le W_{t-1},\end{aligned}$$

where the last inequality holds due to (H.15). Thus, W_t is a nonnegative supermartingale with respect to $\{\mathcal{F}_{t+1}\}_{t=0}^{T}$. Using Ville's inequality (Lemma J.9), we have

$$\mathbb{P}\left(\max_{0\le t\le T} W_t \le \frac{2}{\delta}\right) \ge 1 - \frac{\delta}{2}. \tag{H.16}$$

Define

$$J_t := \mathbb{1}\left(\max_{0\le s<t} W_s \le \frac{2}{\delta}\right), \qquad L := \log\frac{2}{\delta p_0(\theta^*)}.$$

Whenever $J_t = 1$, the token-level mixture satisfies

$$\frac{\widehat{\pi}_t(a|x,u)}{\pi_{\mathrm{F}}^*(a|x,u)} \ge p_{t-1}(\theta^*) = \frac{p_0(\theta^*)}{W_{t-1}} \ge e^{-L}.$$

where the first inequality holds due to $\widehat{\pi}_t(a|x,u) := \sum_\theta p_{t-1}(\theta)\pi_\theta(a|x,u) \ge p_{t-1}(\theta^*)\pi_{\mathrm{F}}^*(a|x,u)$. The second inequality holds due to (H.16). Thus,

$$\log\frac{\widehat{\pi}_t(a|x,u)}{\pi_{\mathrm{F}}^*(a|x,u)} \ge -L$$

for every (x,u,a). Using the definition of X_t and the loss representation, we obtain

$$\begin{aligned}\frac{X_t}{H} &= \frac{\ell_t(\pi_{\mathrm{F}}^*) - \ell_t(\widehat{\pi}_t)}{H} \\ &= \sum_{h=1}^{H}\sum_{(x,u,a)} \frac{c_t(x,u,a)}{H}\log\frac{\widehat{\pi}_t(a|x,u)}{\pi_{\mathrm{F}}^*((a|x,u)} \\ &\ge -L\sum_{h=1}^{H}\sum_{(x,u,a)} \frac{c_t(x,u,a)}{H} \\ &= -L.\end{aligned}$$

The inequality uses $c_t(x,u,a) \ge 0$, and the last equality follows from (H.5).
When $J_t = 0$, we instead have $J_t X_t/H = 0 \ge -L$, since $L \ge 0$. Combining the two cases gives

$$\frac{J_t X_t}{H} \ge -L.$$

Moreover, J_t is $\mathcal{F}_t$-measurable,

$$\mathbb{E}[e^{J_t X_t/H} \mid \mathcal{F}_t] = 1 * \mathbb{1}(J_t = 0) + \mathbb{1}(J_t = 1)\mathbb{E}[e^{X_t/H} \mid \mathcal{F}_t]$$

$$= 1 - J_t + J_t \mathbb{E}[e^{X_t/H} \mid \mathcal{F}_t]$$
$$\leq 1,$$

where the last inequality holds due to (H.15).
Hence, Lemma C.2 and the argument leading to (C.22) apply to $J_t X_t$, whose conditional means are $-J_t r_t$. Replacing L_T by L and the failure probability by $\delta/2$, we obtain, for any fixed $\lambda \in (0, 1/H]$, with probability at least $1 - \delta/2$,

$$\sum_{t=1}^{T} J_t(X_t + r_t) \leq H(1+L)\lambda \sum_{t=1}^{T} J_t r_t + \frac{\log(2/\delta)}{\lambda}.$$

Choose $\lambda = [2H(1+L)]^{-1}$. Conditioned on the event in (H.16), and taking a union bound, we conclude that, with probability at least $1 - \delta$,

$$\mathrm{Regret}(T) \leq -\sum_{t=1}^{T} X_t + \frac{1}{2}\mathrm{Regret}(T) + 2H(1+L)\log\frac{2}{\delta}.$$

Finally, (H.11) with $\theta = \theta^*$ gives

$$-\sum_{t=1}^{T} X_t \leq H \log \frac{1}{p_0(\theta^*)}.$$

Substituting this bound and rearranging yields

$$\mathrm{Regret}(T) \leq 2H \log \frac{1}{p_0(\theta^*)} + 4H\left(1 + \log \frac{2}{\delta p_0(\theta^*)}\right)\log\frac{2}{\delta}.$$

□

H.3 Function class with Finite Covering Number

Let $\Pi = \{\pi_\theta : \theta \in \Theta\}$ be a class of autoregressive policies containing π^*_{F}. For $\varepsilon > 0$, we define the ε-covering number of the token log-probability class $\log \Pi$ as follows.

Definition H.3 (Covering number)**.** The covering number $\mathcal{N}(\varepsilon, \log\Pi, \|\cdot\|_\infty)$ is the smallest cardinality of a finite subset $\Pi_\varepsilon \subseteq \Pi$ such that, for every $\pi \in \Pi$, there exists $\widetilde{\pi} \in \Pi_\varepsilon$ satisfying

$$\|\log\pi - \log\widetilde{\pi}\|_\infty := \sup_{\substack{h\in[H],\, x\in\mathcal{X} \\ u\in\mathcal{A}^{h-1},\, a\in\mathcal{A}}} \left|\log\pi(a|x,u) - \log\widetilde{\pi}(a|x,u)\right| \leq \varepsilon.$$

When the function class Π has finite covering number $N_\varepsilon := \mathcal{N}(\varepsilon, \log\Pi, \|\cdot\|_\infty) < \infty$, we can choose an ε-cover $\Pi_\varepsilon \subseteq \Pi$ of cardinality N_ε, and run Algorithm 3 on Π_ε with a prior distribution p_0. Then, we have the following regret guarantee:

Theorem H.4. Let Π be a class of autoregressive policies containing π^*_{F}. Fix $\varepsilon > 0$ such that

$$N_\varepsilon := \mathcal{N}(\varepsilon, \log\Pi, \|\cdot\|_\infty) < \infty.$$

Let the algorithm be described as above and p_0 be the uniform prior. Under off-policy feedback, both with and without teacher logits, the expected regret can be bounded by

$$\mathbb{E}[\mathrm{Regret}(T)] \leq H \log N_\varepsilon + TH\varepsilon.$$

Moreover, for every $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\mathrm{Regret}(T) \leq 2H\big(\log N_\varepsilon + T\varepsilon\big) + 4H\Big(1 + \log N_\varepsilon + T\varepsilon + \log\frac{2}{\delta}\Big)\log\frac{2}{\delta}.$$

Proof. Since $\pi^*_{\mathrm{F}} \in \Pi$, the definition of the ε-cover ensures that there exists $\pi_{\bar{\theta}} \in \Pi_\varepsilon$ such that

$$\|\log\pi^*_{\mathrm{F}} - \log\pi_{\bar{\theta}}\|_\infty \leq \varepsilon. \tag{H.17}$$

In particular, for every prefix–token triple (x, u, a),

$$\log\frac{\pi^*_{\mathrm{F}}(a|x,u)}{\pi_{\bar{\theta}}(a|x,u)} \leq \varepsilon.$$

Using the autoregressive factorization, for every complete response $y = (a_1, \dots, a_H)$, we obtain

$$\begin{aligned}\log \frac{\pi_{\mathrm{F}}^*(y|x)}{\pi_{\bar{\theta}}(y|x)} &= \sum_{h=1}^{H} \log \frac{\pi_{\mathrm{F}}^*(a_h|x, y_{<h})}{\pi_{\bar{\theta}}(a_h|x, y_{<h})} \\ &\le H\varepsilon.\end{aligned}$$

Taking expectations over $x \sim \bar{\rho}$ and $y \sim \pi_{\mathrm{F}}^*(\cdot|x)$ therefore gives

$$\mathbb{E}_{x\sim\bar{\rho}}\mathrm{KL}\big(\pi_{\mathrm{F}}^*(\cdot|x) \,\big\|\, \pi_{\bar{\theta}}(\cdot|x)\big) \le H\varepsilon. \tag{H.18}$$

The uniform prior on the cover assigns $p_0(\bar{\theta}) = 1/N_\varepsilon$. Applying Theorem H.2, we have

$$\begin{aligned}\mathbb{E}[\mathrm{Regret}(T)] &\le H \log \frac{1}{p_0(\bar{\theta})} + T\mathbb{E}_{x\sim\bar{\rho}}\mathrm{KL}\big(\pi_{\mathrm{F}}^*(\cdot|x) \,\big\|\, \pi_{\bar{\theta}}(\cdot|x)\big) \\ &\le H \log N_\varepsilon + TH\varepsilon,\end{aligned}$$

where we use p_0 is uniform and (H.18). This has proved the expected regret bound.
For the high-probability bound, let Θ_ε index the N_ε policies in Π_ε, with $p_0(\theta) = 1/N_\varepsilon$ for every $\theta \in \Theta_\varepsilon$. We use the same filtration $\mathcal{F}_t$ as in the finite-class analysis and define

$$\begin{aligned}X_t &:= \ell_t(\pi_{\mathrm{F}}^*) - \ell_t(\widehat{\pi}_t), \\ r_t &:= -\mathbb{E}[X_t \mid \mathcal{F}_t].\end{aligned}$$

By (H.13), $\mathrm{Regret}(T) = \sum_{t=1}^{T} r_t$. For the policy $\pi_{\bar{\theta}}$ in (H.17),

$$\begin{aligned}\ell_t(\pi_{\bar{\theta}}) - \ell_t(\pi_{\mathrm{F}}^*) &= \sum_{h=1}^{H} \sum_{(x,u,a)} c_t(x,u,a) \log \frac{\pi_{\mathrm{F}}^*(a|x,u)}{\pi_{\bar{\theta}}(a|x,u)} \\ &\le \varepsilon \sum_{h=1}^{H} \sum_{(x,u,a)} c_t(x,u,a) = H\varepsilon.\end{aligned} \tag{H.19}$$

Combining (H.19) with (H.11) gives

$$\begin{aligned}-\sum_{t=1}^{T} X_t &= \sum_{t=1}^{T} [\ell_t(\widehat{\pi}_t) - \ell_t(\pi_{\mathrm{F}}^*)] \\ &= \sum_{t=1}^{T} [\ell_t(\widehat{\pi}_t) - \ell_t(\pi_{\bar{\theta}})] + \sum_{t=1}^{T} [\ell_t(\pi_{\bar{\theta}}) - \ell_t(\pi_{\mathrm{F}}^*)] \\ &\le H \log N_\varepsilon + \sum_{t=1}^{T} [\ell_t(\pi_{\bar{\theta}}) - \ell_t(\pi_{\mathrm{F}}^*)] \\ &\le H \log N_\varepsilon + TH\varepsilon.\end{aligned} \tag{H.20}$$

Define C_t and W_t the same as the last section, that is

$$\begin{aligned}C_t &:= \sum_{\theta\in\Theta_\varepsilon} p_0(\theta) \exp\bigg(-\frac{1}{H}\sum_{s=1}^{t} \ell_s(\pi_\theta)\bigg) = \frac{1}{N_\epsilon} \sum_{\theta\in\Theta_\varepsilon} \exp\bigg(-\frac{1}{H}\sum_{s=1}^{t} \ell_s(\pi_\theta)\bigg), \\ W_t &:= C_t \exp\bigg(\frac{1}{H}\sum_{s=1}^{t} \ell_s(\pi_{\mathrm{F}}^*)\bigg),\end{aligned}$$

with $C_0 = W_0 = 1$. With a similar argument to (C.20), we have

$$\mathbb{E}\bigg[\exp\bigg(\frac{\ell_t(\pi_{\mathrm{F}}^*) - \ell_t(\widehat{\pi}_t)}{H} \bigg) \,\bigg|\, \mathcal{F}_t \bigg] \le 1. \tag{H.21}$$

Using the normalizer-ratio identity (H.6), we have

$$\frac{W_t}{W_{t-1}} = \frac{C_t}{C_{t-1}} \exp\big(\ell_t(\pi_{\mathrm{F}}^*)/H\big)$$

$$= \sum_{\theta \in \Theta_\varepsilon} p_{t-1}(\theta) \exp\left(\frac{\ell_t(\pi_{\mathrm{F}}^*) - \ell_t(\pi_\theta)}{H}\right),$$

where the last equality holds due to the same argument as in (H.6). We also have

$$\begin{aligned} p_t(\bar{\theta}) &= \frac{N_\varepsilon^{-1} \exp\big(-H^{-1}\sum_{s=1}^t \ell_s(\pi_{\bar{\theta}})\big)}{C_t} \\ &= \frac{1}{N_\varepsilon W_t} \exp\left(-\frac{1}{H}\sum_{s=1}^t [\ell_s(\pi_{\bar{\theta}}) - \ell_s(\pi_{\mathrm{F}}^*)]\right) \\ &\geq \frac{e^{-t\varepsilon}}{N_\varepsilon W_t}, \end{aligned}$$

where the inequality uses (H.19). Furthermore, (H.17) implies $\pi_{\bar{\theta}}(a|x,u) \geq e^{-\varepsilon}\pi_{\mathrm{F}}^*(a|x,u)$. Therefore, the token-level mixture satisfies

$$\begin{aligned} \frac{\widehat{\pi}_t(a|x,u)}{\pi_{\mathrm{F}}^*(a|x,u)} &\geq p_{t-1}(\bar{\theta}) \frac{\pi_{\bar{\theta}}(a|x,u)}{\pi_{\mathrm{F}}^*(a|x,u)} \\ &\geq \frac{e^{-t\varepsilon}}{N_\varepsilon W_{t-1}}. \end{aligned}$$

Using the same concentration argument as the last section, we can consider the high-probability event

$$\left\{\max_{0 \leq s < t} W_s \leq \frac{2}{\delta}\right\}.$$

On this event, we have

$$\frac{\widehat{\pi}_t(a|x,u)}{\pi_{\mathrm{F}}^*(a|x,u)} \geq \frac{\delta e^{-T\varepsilon}}{2N_\varepsilon} = e^{-L},$$

where we define $L := \log(2N_\varepsilon/\delta) + T\varepsilon$. Repeating the stopped-process argument from the proof of Theorem H.2, we obtain, with probability at least $1-\delta$,

$$\mathrm{Regret}(T) \leq 2H\big(\log N_\varepsilon + T\varepsilon\big) + 4H\left(1 + \log N_\varepsilon + T\varepsilon + \log\frac{2}{\delta}\right)\log\frac{2}{\delta}.$$

This proves Theorem H.4. □

Example H.5. Suppose $\Theta \subseteq \mathbb{R}^d$ is a nonempty compact set contained in a Euclidean ball of radius R, where $d \geq 1$. Assume that $\pi_{\theta^*} = \pi_{\mathrm{F}}^*$ for some $\theta^* \in \Theta$. Moreover, suppose that, for some $L_0 > 0$,

$$\big|\log \pi_\theta(a|x,u) - \log \pi_{\theta'}(a|x,u)\big| \leq L_0\|\theta - \theta'\|_2$$

for every $\theta, \theta' \in \Theta$, $h \in [H]$, and $(x,u,a) \in \mathcal{X} \times \mathcal{A}^{h-1} \times \mathcal{A}$. Then, a standard Euclidean covering argument gives

$$\mathcal{N}(\varepsilon, \log \Pi, \|\cdot\|_\infty) \leq \left(1 + \frac{2RL_0}{\varepsilon}\right)^d.$$

For every $T \geq 2$, choose such a cover with $\varepsilon = 1/T$. By Theorem H.4, the expected regret satisfies

$$\mathbb{E}\big[\mathrm{Regret}(T)\big] \leq H\left[d\log\big(1 + 2RL_0T\big) + 1\right],$$

Moreover, with probability at least $1-\delta$, we have

$$\mathrm{Regret}(T) \leq \widetilde{O}(dH\log T).$$

I Reverse KL with Function Approximation

We now study on-policy distillation with function approximation. Fix any $h \in [H]$. For any prefix $x \in \mathcal{X}$, $u \in \mathcal{A}^{h-1}$, $a \in \mathcal{A}$, following the notation in Theorem 4.1, we define the target as

$$f^*(h, x, u, a) := \log g_h(a|x, u) = \sum_{i \in \mathcal{I}} w_i(x) \log p_i(a|x, u).$$

Specifically, we define

$$\mathcal{Z} := \bigcup_{h=1}^{H} \big(\{h\} \times \mathcal{X} \times \mathcal{A}^{h-1} \times \mathcal{A}\big),$$

and write $z_{t,j,h} := (h, x_{t,j}, y_{t,j,<h}, a_{t,j,h})$ for the query associated with the h-th token of rollout j in round t.

Throughout this section, we retain Assumption 4.3, with reference policy π_{ref} and bound $B > 0$. Moreover, let $\mathcal{F}$ be a finite class of functions $f : \mathcal{Z} \to \mathbb{R}$. We further impose the following realizability assumption.

Assumption I.1. $f^* \in \mathcal{F}$.

To measure how past observations constrain predictions at a new query, we introduce a new version of generalized Eluder dimension adapted to autoregressive trajectory batches. The underlying uncertainty compares the disagreement between two candidate score functions at the current query with their cumulative squared disagreement on previously observed data. Since the estimator is updated only after each round, this comparison uses observations from completed rounds, without incorporating feedback from the current batch.

Definition I.2. Fix $\lambda > 0$. For a sequence of trajectory batches, write

$$Z_t := (z_{t,j,h})_{j \in [m],\, h \in [H]}, \qquad Z_{<t} := (Z_1, \dots, Z_{t-1}).$$

For any query $z \in \mathcal{Z}$, define

$$D^2_{\mathcal{F}}(z; Z_{<t}) := \sup_{f_1, f_2 \in \mathcal{F}} \frac{\big(f_1(z) - f_2(z)\big)^2}{\lambda + \sum_{s=1}^{t-1} \frac{1}{mH} \sum_{j=1}^{m} \sum_{h=1}^{H} \big(f_1(z_{s,j,h}) - f_2(z_{s,j,h})\big)^2}. \tag{I.1}$$

The batched generalized Eluder dimension is

$$\dim_T(\mathcal{F}; \lambda) := \sup_{Z_{1:T}} \sum_{t=1}^{T} \frac{1}{mH} \sum_{j=1}^{m} \sum_{h=1}^{H} \min\left\{1,\, D^2_{\mathcal{F}}(z_{t,j,h}; Z_{<t})\right\}, \tag{I.2}$$

where the supremum is over all sequences of T trajectory batches, each containing m rollouts of horizon H.

I.1 Algorithm Design

Recall that, for every visited query $z_{t,j,h} = (h, x, u, a)$,

$$\begin{aligned} \mathbb{E}\big[l_{t,j,h} \,\big|\, \mathcal{F}_t,\, z_{t,j,h} = (h, x, u, a)\big] &= \sum_{i \in \mathcal{I}} w_i(x) \log p_i(a|x, u) \\ &= \log g_h(a|x, u) = f^*(h, x, u, a), \end{aligned} \tag{I.3}$$

where $\mathcal{F}_t$ denotes the history before round t. This identity motivates estimating f^* by least squares. For each $f \in \mathcal{F}$, define

$$\ell_t(f) := \frac{1}{mH} \sum_{j=1}^{m} \sum_{h=1}^{H} \big(f(z_{t,j,h}) - l_{t,j,h}\big)^2, \tag{I.4}$$

Moreover, we define

$$L_t(f) := \sum_{s=1}^{t} \ell_s(f), \qquad L_0(f) := 0.$$

At the beginning of round t, the learner computes

$$\bar{f}_{t-1} \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f). \tag{I.5}$$

Using the principle of optimism for online reinforcement learning, we therefore define a bonus using the batched generalized Eluder dimension, i.e., for any $z \in \mathcal{Z}$, the bonus function is defined as

$$b_{t-1}(z) := \sqrt{(4\beta + \lambda) D^2_{\mathcal{F}}(z; Z_{<t})}, \tag{I.6}$$

where $\lambda > 0$ and $\beta > 0$ are parameters to be specified later. Given the reference policy π_{ref} and the bound B from Assumption 4.3, as it implies

$$\big| f^*(h, x, u, a) - \log \pi_{\text{ref}}(a|x, u) \big| \le B,$$

we define the optimistic score by

$$\widehat{f}_{t-1}(h, x, u, a) := \min \Big\{ \bar{f}_{t-1}(h, x, u, a) + b_{t-1}(h, x, u, a),\, \log \pi_{\text{ref}}(a|x, u) + B \Big\}. \tag{I.7}$$

The following lemma establishes pointwise optimism and bounds the error of the optimistic score.

Lemma I.3. Under Assumptions 4.3 and I.1, fix $\delta \in (0, 1)$ and choose

$$\beta = 32 B^2 \log \frac{|\mathcal{F}|}{\delta}, \qquad \lambda > 0.$$

Then, with probability at least $1 - \delta$, simultaneously for every $t \in [T]$ and $z \in \mathcal{Z}$,

$$\big| \bar{f}_{t-1}(z) - f^*(z) \big| \le b_{t-1}(z),$$

Furthermore, the following inequality holds

$$0 \le \widehat{f}_{t-1}(z) - f^*(z) \le \min\big\{ 2 b_{t-1}(z), 2B \big\}.$$

To construct the output policy, we apply the backward propagation as in the tabular setting to the estimated scores $\widehat{f}_t$. More specifically, set $\widehat{V}_{t-1,H+1}(x, y) = 1$ and recursively define

$$\widehat{V}_{t-1,h}(x, u) := \sum_{a \in \mathcal{A}} \exp\big(\widehat{f}_{t-1}(h, x, u, a) \big) \widehat{V}_{t-1,h+1}(x, (u, a)), \tag{I.8}$$

for $h = H, \dots, 1$. We then output

$$\widehat{\pi}_t(a|x, u) := \frac{\exp\big(\widehat{f}_{t-1}(h, x, u, a) \big) \widehat{V}_{t-1,h+1}(x, (u, a))}{\widehat{V}_{t-1,h}(x, u)}. \tag{I.9}$$

This procedure is described in Algorithm 4.

Algorithm 4 Optimistic Reverse KL with Function Approximation

1: **Input:** Function class $\mathcal{F}$, reference policy π_{ref}, bound B, horizon H, batch size m, number of rounds T, and parameters $\lambda, \beta > 0$.
2: Initialize $L_0(f) = 0$ for every $f \in \mathcal{F}$.
3: **for** $t = 1, \dots, T$ **do**
4: Compute the regression estimate $\bar{f}_{t-1}$ using (I.5).
5: Compute b_{t-1} and $\widehat{f}_{t-1}$ using (I.6) and (I.7).
6: Set $\widehat{V}_{t-1,H+1}(x, y) = 1$ and compute the backward recursion (I.8).
7: Construct $\widehat{\pi}_t$ using (I.9).
8: Receive contexts $\{x_{t,j}\}_{j=1}^m$ according to the on-policy protocol and generate $y_{t,j} \sim \widehat{\pi}_t(\cdot|x_{t,j})$.
9: Observe teacher feedback $\{l_{t,j,h}\}_{j \in [m],\, h \in [H]}$.
10: Compute $\ell_t(f)$ using (I.4) and update $L_t(f) = L_{t-1}(f) + \ell_t(f)$ for every $f \in \mathcal{F}$.
11: **end for**
12: **Output:** $\{\widehat{\pi}_t\}_{t=1}^T$.

I.2 Theoretical Guarantee

Given Lemma I.3, we have the following guarantee on the regret of Algorithm 4.

Theorem I.4. Under Assumptions 4.3 and I.1, fix $\delta \in (0,1)$ and choose $\lambda > 0$ and

$$\beta \geq 32B^2 \log \frac{2|\mathcal{F}|}{\delta}$$

Then, with probability at least $1-\delta$, Algorithm 4 satisfies

$$\begin{aligned}\mathrm{Regret}(T) &:= \sum_{t=1}^{T} \mathbb{E}_{x\sim\bar{\rho}} \mathrm{KL}\big(\widehat{\pi}_t(\cdot|x) \,\big\|\, \pi^*_{\mathrm{R}}(\cdot|x)\big) \\ &\leq 4H^2(4\beta+\lambda)\dim_T(\mathcal{F};\lambda) + 16H^2B^2 \log\frac{4}{\delta}.\end{aligned}$$

In particular, if we choose $\lambda = B^2$ and $\beta = 32B^2\log(2|\mathcal{F}|/\delta)$, we have

$$\mathrm{Regret}(T) \leq O\bigg(B^2H^2 \dim_T(\mathcal{F};B^2)\log\frac{2|\mathcal{F}|}{\delta}\bigg).$$

Proof. Let $\mathcal{F}_t$ denote the history before round t. In particular, $\bar{f}_{t-1}$, b_{t-1}, and $\widehat{\pi}_t$ are $\mathcal{F}_t$-measurable. Applying Lemma I.3, we have with probability at least $1-\delta/2$, the following inequality holds simultaneously for every $t \in [T]$ and $z \in \mathcal{Z}$

$$0 \leq \widehat{f}_{t-1}(z) - f^*(z) \leq \min\{2b_{t-1}(z), 2B\}, \tag{I.10}$$

when $\beta \geq 4B^2\log(2|\mathcal{F}|/\delta)$. Let $\mathcal{E}$ be this high-probability event. For every t and z, define

$$\psi_t(z) := \min\{4b_{t-1}(z)^2, 4B^2\}. \tag{I.11}$$

Then we have $0 \leq \psi_t(z) \leq 4B^2$. On $\mathcal{E}$, (I.10) further gives

$$\big[\widehat{f}_{t-1}(z) - f^*(z)\big]^2 \leq \psi_t(z). \tag{I.12}$$

Fix a context x and a response y. Write

$$v_t(x,y) := \sum_{h=1}^{H} \Big[\widehat{f}_{t-1}(h,x,y_{<h},a_h) - f^*(h,x,y_{<h},a_h)\Big].$$

On $\mathcal{E}$, every summand is nonnegative, so $v_t(x,y) \geq 0$. Recall that for $y = (a_1,\ldots,a_H)$, the backward propagation gives

$$\begin{aligned}\widehat{\pi}_t(y|x) &= \prod_{h=1}^{H} \frac{e^{\widehat{f}_{t-1}(h,x,y_{<h},a_h)}\widehat{V}_{t-1,h+1}(x,y_{\leq h})}{\widehat{V}_{t-1,h}(x,y_{<h})} \\ &= \frac{\exp\big(\sum_{h=1}^{H}\widehat{f}_{t-1}(h,x,y_{<h},a_h)\big)}{\widehat{V}_{t-1,1}(x)},\end{aligned}$$

where we used $\widehat{V}_{t-1,H+1}(x,y) = 1$. Similarly, the reverse target satisfies

$$\pi^*_{\mathrm{R}}(y|x) = \frac{\exp\big(\sum_{h=1}^{H} f^*(h,x,y_{<h},a_h)\big)}{V_1(x)},$$

where $V_1(x) = \sum_y \exp\big(\sum_{h=1}^{H} f^*(h,x,y_{<h},a_h)\big)$. Taking the logarithm of their ratio therefore yields

$$\begin{aligned}\log\frac{\widehat{\pi}_t(y|x)}{\pi^*_{\mathrm{R}}(y|x)} &= \sum_{h=1}^{H}\big[\widehat{f}_{t-1}(h,x,y_{<h},a_h) - f^*(h,x,y_{<h},a_h)\big] \\ &\quad + \log V_1(x) - \log\widehat{V}_{t-1,1}(x) \\ &= v_t(x,y) + \log\frac{V_1(x)}{\widehat{V}_{t-1,1}(x)}.\end{aligned} \tag{I.13}$$

To express the ratio of normalizing constants as an expectation, expand the expectation over complete responses:

$$\begin{aligned}
&\mathbb{E}_{y\sim\widehat{\pi}_t(\cdot|x)}\left[e^{-v_t(x,y)}\right] \\
&= \sum_{y\in\mathcal{Y}} \widehat{\pi}_t(y|x) e^{-v_t(x,y)} \\
&= \sum_{y\in\mathcal{Y}} \frac{\exp\left(\sum_{h=1}^{H} \widehat{f}_{t-1}(h,x,y_{<h},a_h)\right)}{\widehat{V}_{t-1,1}(x)} \exp\left(\sum_{h=1}^{H}\left[f^*(h,x,y_{<h},a_h) - \widehat{f}_{t-1}(h,x,y_{<h},a_h)\right]\right) \\
&= \frac{1}{\widehat{V}_{t-1,1}(x)} \sum_{y\in\mathcal{Y}} \exp\left(\sum_{h=1}^{H} f^*(h,x,y_{<h},a_h)\right) \\
&= \frac{V_1(x)}{\widehat{V}_{t-1,1}(x)}.
\end{aligned} \tag{I.14}$$

The third equality follows by cancellation of the $\widehat{f}_{t-1}$ terms in the exponent. The last equality uses the definition of $V_1(x)$.
Taking expectation over $\widehat{\pi}_t(\cdot|x)$ in (I.13) and considering (I.14), we have

$$\begin{aligned}
\mathrm{KL}\big(\widehat{\pi}_t(\cdot|x) \,\big\|\, \pi^*_{\mathrm{R}}(\cdot|x)\big) &= \mathbb{E}_{\widehat{\pi}_t}[v_t(x,y)] + \log \mathbb{E}_{\widehat{\pi}_t}[e^{-v_t(x,y)}] \\
&\le \mathbb{E}_{\widehat{\pi}_t}[v_t(x,y)] + \mathbb{E}_{\widehat{\pi}_t}[e^{-v_t(x,y)}] - 1 \\
&\le \frac{1}{2}\mathbb{E}_{\widehat{\pi}_t}[v_t(x,y)^2].
\end{aligned} \tag{I.15}$$

Here the first inequality uses $\log u \le u - 1$ for $u > 0$, and the second uses $e^{-v} \le 1 - v + v^2/2$ for $v \ge 0$. All expectations in this display are conditional on x.
By the Cauchy–Schwarz inequality and (I.12),

$$\begin{aligned}
v_t(x,y)^2 &\le H \sum_{h=1}^{H} \left[\widehat{f}_{t-1}(h,x,y_{<h},a_h) - f^*(h,x,y_{<h},a_h)\right]^2 \\
&\le H \sum_{h=1}^{H} \psi_t(h,x,y_{<h},a_h).
\end{aligned}$$

Consequently, on $\mathcal{E}$,

$$\mathrm{Regret}(T) \le \frac{H}{2} \sum_{t=1}^{T} \mathbb{E}_{x\sim\bar{\rho},\, y\sim\widehat{\pi}_t(\cdot|x)}\left[\sum_{h=1}^{H} \psi_t(h,x,y_{<h},a_h)\right]. \tag{I.16}$$

Define

$$X_t := \frac{1}{mH} \sum_{j=1}^{m} \sum_{h=1}^{H} \psi_t(z_{t,j,h}).$$

Then X_t is $\mathcal{F}_{t+1}$-measurable and satisfies $0 \le X_t \le 4B^2$ almost surely. Under the on-policy protocol, each rollout has conditional marginal distribution

$$\begin{aligned}
\mathbb{P}\big(x_{t,j} = x,\, y_{t,j} = y \mid \mathcal{F}_t\big) &= \frac{1}{I} \sum_{i\in\mathcal{I}} \rho_i(x) \widehat{\pi}_t(y|x) \\
&= \bar{\rho}(x) \widehat{\pi}_t(y|x).
\end{aligned}$$

Since ψ_t is determined before round t, this implies

$$\mathbb{E}[X_t \mid \mathcal{F}_t] = \frac{1}{H} \mathbb{E}_{x\sim\bar{\rho},\, y\sim\widehat{\pi}_t(\cdot|x)}\left[\sum_{h=1}^{H} \psi_t(h,x,y_{<h},a_h)\right]. \tag{I.17}$$

This identity uses only the marginal law of each rollout, not independence within the batch. Combining (I.16) and (I.17), we obtain, on $\mathcal{E}$,

$$\mathrm{Regret}(T) \le \frac{H^2}{2} \sum_{t=1}^{T} \mathbb{E}[X_t \mid \mathcal{F}_t]. \tag{I.18}$$

Apply Lemma J.6 to $\{X_t\}_{t=1}^T$, with range bound $4B^2$ and failure probability $\delta/2$. With probability at least $1 - \delta/2$,

$$\sum_{t=1}^{T} \mathbb{E}[X_t \mid \mathcal{F}_t] \le 2 \sum_{t=1}^{T} X_t + 32B^2 \log \frac{4}{\delta}. \tag{I.19}$$

Taking a union bound, we conclude that, with probability at least $1 - \delta$,

$$\mathrm{Regret}(T) \le H^2 \sum_{t=1}^{T} X_t + 16H^2B^2 \log \frac{4}{\delta}. \tag{I.20}$$

Finally, by the definition of the bonus function, we have

$$\psi_t(z) = 4 \min\left\{(4\beta + \lambda) D^2_{\mathcal{F}}(z; Z_{<t}), B^2\right\}.$$

Our choice of β ensures $4\beta + \lambda \ge B^2$. Therefore,

$$\psi_t(z) \le 4(4\beta + \lambda) \min\left\{1, D^2_{\mathcal{F}}(z; Z_{<t})\right\}. \tag{I.21}$$

Indeed, when $D^2_{\mathcal{F}}(z; Z_{<t}) \le 1$, we use the first term in the minimum defining ψ_t. Otherwise, we use $\psi_t(z) \le 4B^2 \le 4(4\beta + \lambda)$.
Summing (I.21) over the observed queries gives

$$\begin{aligned}
\sum_{t=1}^{T} X_t &= \sum_{t=1}^{T} \frac{1}{mH} \sum_{j=1}^{m} \sum_{h=1}^{H} \psi_t(z_{t,j,h}) \\
&\le 4(4\beta + \lambda) \sum_{t=1}^{T} \frac{1}{mH} \sum_{j=1}^{m} \sum_{h=1}^{H} \min\left\{1, D^2_{\mathcal{F}}(z_{t,j,h}; Z_{<t})\right\} \\
&\le 4(4\beta + \lambda) \dim_T(\mathcal{F}; \lambda).
\end{aligned}$$

The last inequality follows from the definition of the generalized Eluder dimension, since the observed batches form a valid sequence of autoregressive trajectories. Substituting this inequality into (I.20) completes the proof of Theorem I.4. □

I.3 Proof of Lemma I.3

Proof. Let $\mathcal{F}_t$ denote the history before round t. Fix $f \in \mathcal{F}$, and define

$$\begin{aligned}
\Delta_{t,j,h}(f) &:= f(z_{t,j,h}) - f^*(z_{t,j,h}), \\
\xi_{t,j,h} &:= l_{t,j,h} - f^*(z_{t,j,h}).
\end{aligned}$$

The conditional-mean identity gives

$$\mathbb{E}[\xi_{t,j,h} \mid \mathcal{F}_t, z_{t,j,h}] = 0. \tag{I.22}$$

Moreover, conditioned on $z_{t,j,h} = z = (h, x, u, a)$, Assumption 4.3 implies

$$\xi_{t,j,h} \in \left[\log \pi_{\text{ref}}(a|x,u) - B - f^*(z),\ \log \pi_{\text{ref}}(a|x,u) + B - f^*(z)\right].$$

We use the abbreviation $\Delta := \Delta_{t,j,h}(f), \xi := \xi_{t,j,h}$. Then using Assumption 4.3 to f^* by observing it is an average over all the teachers, we have

$$\mathbb{E}[\xi \mid \mathcal{F}_t, z_{t,j,h}] = 0, \qquad |\xi| \le 2B.$$

Set $\eta = 1/(32B^2)$. We claim that

$$\mathbb{E}\left[e^{\eta(2\Delta\xi - \Delta^2/2)} \mid \mathcal{F}_t, z_{t,j,h}\right] \le 1. \tag{I.23}$$

In fact, if $|\Delta| \geq 8B$, then

$$2\Delta\xi - \frac{\Delta^2}{2} \leq 4B|\Delta| - \frac{\Delta^2}{2} \leq 0.$$

Thus, (I.23) holds in this case. Otherwise, if $|\Delta| < 8B$, then

$$|2\eta\Delta\xi| \leq 2\eta \cdot 8B \cdot 2B = 1.$$

Using $e^u \leq 1 + u + u^2$ for $|u| \leq 1$, we obtain

$$\begin{aligned}\mathbb{E}[e^{2\eta\Delta\xi} \mid \mathcal{F}_t, z_{t,j,h}] &\leq 1 + 2\eta\Delta\mathbb{E}[\xi \mid \mathcal{F}_t, z_{t,j,h}] + 4\eta^2\Delta^2\mathbb{E}[\xi^2 \mid \mathcal{F}_t, z_{t,j,h}] \\ &\leq 1 + 16\eta^2B^2\Delta^2 \\ &\leq \exp(16\eta^2B^2\Delta^2) \\ &= \exp(\eta\Delta^2/2).\end{aligned}$$

The second inequality uses the zero conditional mean and $\xi^2 \leq 4B^2$. The third uses $1 + v \leq e^v$, and the last equality follows from our choice of η. Multiplying both sides by $e^{-\eta\Delta^2/2}$ proves (I.23). To handle the dependence within a round, apply Jensen's inequality to the exponential function:

$$\begin{aligned}&\exp\left(\frac{\eta}{mH}\sum_{j=1}^{m}\sum_{h=1}^{H}\left[2\Delta_{t,j,h}(f)\xi_{t,j,h} - \frac{1}{2}\Delta_{t,j,h}(f)^2\right]\right) \\ &\leq \frac{1}{mH}\sum_{j=1}^{m}\sum_{h=1}^{H}\exp\left(\eta\left[2\Delta_{t,j,h}(f)\xi_{t,j,h} - \frac{1}{2}\Delta_{t,j,h}(f)^2\right]\right).\end{aligned}$$

Taking conditional expectations given $\mathcal{F}_t$ and applying (I.23) to each summand gives

$$\mathbb{E}\left[\exp\left(\frac{\eta}{mH}\sum_{j=1}^{m}\sum_{h=1}^{H}\left[2\Delta_{t,j,h}(f)\xi_{t,j,h} - \frac{1}{2}\Delta_{t,j,h}(f)^2\right]\right)\,\middle|\,\mathcal{F}_t\right] \leq 1. \tag{I.24}$$

We do not condition on the entire batch or assume independence among its observations. Define $M_0(f) = 1$ and

$$M_t(f) := \exp\left(\eta\sum_{s=1}^{t}\frac{1}{mH}\sum_{j=1}^{m}\sum_{h=1}^{H}\left[2\Delta_{s,j,h}(f)\xi_{s,j,h} - \frac{1}{2}\Delta_{s,j,h}(f)^2\right]\right).$$

Equation (I.24) implies

$$\mathbb{E}[M_t(f) \mid \mathcal{F}_t] \leq M_{t-1}(f).$$

Thus, $\{M_t(f)\}_{t=0}^T$ is a nonnegative supermartingale. By Ville's inequality (Lemma J.9),

$$\mathbb{P}\left(\max_{0\leq t\leq T} M_t(f) > \frac{|\mathcal{F}|}{\delta}\right) \leq \frac{\delta}{|\mathcal{F}|}.$$

Taking a union bound over $f \in \mathcal{F}$, we obtain an event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \geq 1-\delta$ on which, simultaneously for every $f \in \mathcal{F}$ and $t \in \{0, \ldots, T\}$,

$$\begin{aligned}\sum_{s=1}^{t}\frac{1}{mH}\sum_{j=1}^{m}\sum_{h=1}^{H}\left[2\Delta_{s,j,h}(f)\xi_{s,j,h} - \frac{1}{2}\Delta_{s,j,h}(f)^2\right] &\leq \frac{1}{\eta}\log\frac{|\mathcal{F}|}{\delta} \\ &= 32B^2\log\frac{|\mathcal{F}|}{\delta} \\ &\leq \beta.\end{aligned} \tag{I.25}$$

Expanding the difference of squared losses gives

$$(f(z_{s,j,h}) - l_{s,j,h})^2 - (f^*(z_{s,j,h}) - l_{s,j,h})^2 = \Delta_{s,j,h}(f)^2 - 2\Delta_{s,j,h}(f)\xi_{s,j,h}.$$

Therefore, on $\mathcal{E}$,

$$\begin{aligned} L_t(f) - L_t(f^*) &= \frac{1}{2}\sum_{s=1}^{t}\frac{1}{mH}\sum_{j=1}^{m}\sum_{h=1}^{H}\Delta_{s,j,h}(f)^2 \\ &\quad - \sum_{s=1}^{t}\frac{1}{mH}\sum_{j=1}^{m}\sum_{h=1}^{H}\left[2\Delta_{s,j,h}(f)\xi_{s,j,h} - \frac{1}{2}\Delta_{s,j,h}(f)^2\right] \\ &\geq \frac{1}{2}\sum_{s=1}^{t}\frac{1}{mH}\sum_{j=1}^{m}\sum_{h=1}^{H}\Delta_{s,j,h}(f)^2 - \beta, \end{aligned} \tag{I.26}$$

where the last inequality holds due to (I.25). It holds for any $f \in \mathcal{F}$, so it also applies to $\bar{f}_{t-1}$. Moreover, by (I.5) and the realizability assumption, we have

$$L_{t-1}(\bar{f}_{t-1}) \leq L_{t-1}(f^*).$$

Applying (I.26) at time $t-1$ consequently gives

$$\sum_{s<t}\frac{1}{mH}\sum_{j=1}^{m}\sum_{h=1}^{H}\left[\bar{f}_{t-1}(z_{s,j,h}) - f^*(z_{s,j,h})\right]^2 \leq 2\beta. \tag{I.27}$$

Since both $\bar{f}_{t-1}$ and f^* belong to $\mathcal{F}$, the definition of $D^2_{\mathcal{F}}$ implies, for every $z \in \mathcal{Z}$,

$$\begin{aligned} &\left|\bar{f}_{t-1}(z) - f^*(z)\right|^2 \\ &\leq D^2_{\mathcal{F}}(z; Z_{<t})\left(\lambda + \sum_{s<t}\frac{1}{mH}\sum_{j=1}^{m}\sum_{h=1}^{H}\left[\bar{f}_{t-1}(z_{s,j,h}) - f^*(z_{s,j,h})\right]^2\right) \\ &\leq (\lambda + 2\beta) D^2_{\mathcal{F}}(z; Z_{<t}) \\ &\leq b_{t-1}(z)^2. \end{aligned}$$

Taking square roots proves the pointwise confidence bound.
Finally, for $z = (h, x, u, a)$, Assumption 4.3 and the definition of f^* give

$$\log \pi_{\text{ref}}(a|x,u) - B \leq f^*(z) \leq \log \pi_{\text{ref}}(a|x,u) + B.$$

On $\mathcal{E}$, the confidence bound also gives

$$\bar{f}_{t-1}(z) + b_{t-1}(z) \geq f^*(z).$$

Therefore, $\widehat{f}_{t-1}(z) \geq f^*(z)$, which proves optimism. Moreover,

$$\begin{aligned} \widehat{f}_{t-1}(z) - f^*(z) &= \min\left\{\bar{f}_{t-1}(z) - f^*(z) + b_{t-1}(z),\, \log \pi_{\text{ref}}(a|x,u) + B - f^*(z)\right\} \\ &\leq \min\{2b_{t-1}(z), 2B\}. \end{aligned}$$

This completes the proof of Lemma I.3. □

J Auxilliary Lemmas

J.1 Uniform Bernstein Concentration in Bakhtiari et al. (2026)

First, we list several key results from Bakhtiari et al. (2026) that are used in our analysis. To start with, we present several definitions.

Definition J.1 (CGF-like, Definition 4 Bakhtiari et al. 2026)**.** A twice differentiable function $\psi : [0, \lambda_{\max}) \to \mathbb{R}_+$ is called CGF-like if it is convex, non-negative, and satisfies $\psi(0) = \psi'(0) = 0$.

Definition J.2 (sub-ψ process, Definition 5 Bakhtiari et al. 2026)**.** Let $\{\mathcal{F}_t\}_{t\in[T]}$ be a filtration, $\psi : [0, \lambda_{\max}) \to \mathbb{R}_+$ be a CGF-like function. Let $\{S_t\}_{t\in[T]}$ be a real-valued stochastic process adapted to $\{\mathcal{F}_t\}$, $\{V_t\}_{t\in[T]}$ be another process adapted to $\{\mathcal{F}_t\}$ taking values in $\mathbb{R}_+$. We say that $\{(S_t, V_t)\}_{t\in[T]}$ is a sub-ψ process if for all $t \in [T]$ and $\lambda \in [0, \lambda_{\max})$, there exists an $\mathcal{F}$-adapted supermartingale $\{L_t(\lambda)\}_{t\in[T]}$ such that

$$M_t(\lambda) := \exp\left(\lambda S_t - \sum_{s=1}^{t}\psi(\lambda)V_s\right) \leq L_t(\lambda), \quad \text{a.s.}$$

Definition J.3 (Definition 6 Bakhtiari et al. 2026)**.** A random process $\{(S_t, V_t)\}_{t\in[T]}$ is sub-gamma with parameter $v > 0$ if it is sub-ψ with the CGF-like function $\psi(\lambda) = \frac{\lambda^2}{2(1-v\lambda)}$ for some $v > 0$.

Lemma J.4 (Proposition 11 Bakhtiari et al. 2026)**.** Let $\{\mathcal{F}_t\}_{t\in[T]}$ be a filtration and let $(X_t)_{t\in[T]}$ be a real-valued stochastic process adapted to $\mathcal{F}$. Assume that there exists a constant $b > 0$ such that for all $t \in [T]$,

$$X_t \le \mathbb{E}[X_t|\mathcal{F}_{t-1}] + b \; a.s.$$

Then, for

$$S_t = \sum_{s=1}^{t}(X_s - \mathbb{E}[X_s|\mathcal{F}_{s-1}]), \quad V_t = \sum_{s=1}^{t}\mathrm{Var}(X_s|\mathcal{F}_{s-1}),$$

the process $\{(S_t, V_t)\}_{t\in[T]}$ is sub-gamma with parameter $b/3$.

Lemma J.5 (Theorem 10 Bakhtiari et al. 2026)**.** For any sub-gamma process $\{(S_t, V_t)\}_{t\in[T]}$ with parameter $v > 0$, and any $\rho > 0$, $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have for all $t \in [T]$,

$$S_t \le 4\sqrt{V_t \log(H_t/\delta)} + 11(v + \rho)\log(H_t/\delta),$$

where $H_t = \log(1 + V_t/\rho^2) + e$.

J.2 Other Auxilliary Lemmas

Lemma J.6 (Lemma B.3, Foster et al. 2024)**.** Let $\{\boldsymbol{X}_t\}_{t\in[T]}$ be a sequence of non-negative random variables adapted to a filtration $\{\mathcal{F}_t\}_{t\in[T]}$. Assume that $0 \le \boldsymbol{X}_t \le R$ almost surely for all $t \in [T]$ and some constant $R > 0$. Then, for any $\delta \in (0, 1)$ and all $T' \le T$, with probability at least $1 - \delta$, we have

$$\sum_{t=1}^{T'}\mathbb{E}[\boldsymbol{X}_t|\mathcal{F}_{t-1}] \le 2\sum_{t=1}^{T'}\boldsymbol{X}_t + 8R\log(2/\delta),$$

$$\sum_{t=1}^{T'}\boldsymbol{X}_t \le \frac{3}{2}\sum_{t=1}^{T'}\mathbb{E}[\boldsymbol{X}_t|\mathcal{F}_{t-1}] + 4R\log(2/\delta).$$

Lemma J.7. Let $\mu_f(\cdot) \propto \exp(f(\cdot))$, $\mu_g(\cdot) \propto \exp(g(\cdot))$ be two probability distributions. Then,

$$\mathrm{KL}(\mu_f\|\mu_g) = \mathbb{E}_{\mu_f}[f - g] + \log\left[\mathbb{E}_{\mu_f}[\exp(g - f)]\right].$$

Proof. Let $Z_f = \sum_x \exp(f(x))$, $Z_g = \sum_x \exp(g(x))$ be the normalization constants. Then,

$$\mu_f(x) = \frac{\exp(f(x))}{Z_f}, \; \mu_g(x) = \frac{\exp(g(x))}{Z_g}.$$

The KL-divergence is equal to

$$\begin{aligned}\mathrm{KL}(\mu_f\|\mu_g) &= \sum_x \mu_f(x)\log\frac{\mu_f(x)}{\mu_g(x)}\\ &= \sum_x \mu_f(x)\left[f(x) - g(x) + \log(Z_g/Z_f)\right]\\ &= \mathbb{E}_{\mu_f}[f - g] + \log(Z_g/Z_f).\end{aligned}$$

Moreover,

$$\begin{aligned}Z_g/Z_g &= \frac{\sum_x \exp(g(x))}{\sum_x \exp(f(x))}\\ &= \sum_x \frac{\exp(f(x))}{\sum_y \exp(f(y))}\cdot\exp(g(x) - f(x))\\ &= \mathbb{E}_{\mu_f}[\exp(g - f)].\end{aligned}$$

Thus, we complete the proof. □

Lemma J.8 (Stirling's inequality)**.** Let $\Gamma(\cdot)$ be the gamma function. For any $s > 0$, the following inequality holds

$$\left(s - \frac{1}{2}\right) \log s - s + \frac{1}{2} \log(2\pi) \leq \log \Gamma(s) \leq \left(s - \frac{1}{2}\right) \log s - s + \frac{1}{2} \log(2\pi) + \frac{1}{12s}.$$

Lemma J.9 (Ville's inequality)**.** Let $\{X_t\}_{t=0}^{T}$ be a non-negative supermartingale. Then, for any $v > 0$,

$$\mathbb{P}\Big[\max_{0 \leq t \leq T} X_t \geq v\Big] \leq \frac{\mathbb{E}[X_0]}{v}.$$

Proof. Please refer to Durrett (2019). □