LEARNING FROM FAILURES: WHEN CONTRASTIVE POST-TRAINING BEATS SUCCESS CONDITIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training a language model from rollouts with binary outcomes has two standard recipes: supervised fine-tuning on the successful rollouts (success conditioning), and pairwise contrastive learning on success–failure pairs. In practice the two are combined, and the combination is often reported to work best, but there is little theory of when and why. We analyze the combined family: a margin-based contrastive term plus a weighted success conditioning term on successful rollouts using a contextual bandit setup with binary rewards. We characterize when adding the contrastive term improves on success conditioning, both in value and other metrics such as those used in retrieval (e.g., precision, recall). We also propose another combined family that contains a DPO-style contrastive loss term that accounts for the base policy. We compare it with the combined loss that contains the more traditionally used triplet loss under different settings. We validate our theoretical results via synthetic experiments and also in the real-world context of generative retrieval.

1 INTRODUCTION

In reinforcement learning with binary outcome feedback, each trajectory receives a reward of 1 indicating success or 0 indicating failure. A policy can be trained by maximizing likelihood on successful trajectories alone or by contrasting successful and failed trajectories. These alternatives appear in applications such as success-only fine-tuning of language models e.g., (Zelikman et al., 2022; Singh et al., 2024; Russo, 2026) and preference optimization for language models e.g., (Rafailov et al., 2024). Combining the two objectives has also been of interest yielding gains in applications such as reasoning (Pang et al., 2024) and generative retrieval (Li et al., 2023a), motivating a general question: when does a combination of the two ideas i.e., maximizing likelihood of successful trajectories and contrasting successful and failed trajectories, improve a policy beyond training on successful trajectories alone?

Maximizing likelihood of successful trajectories (also known as Success conditioning, (Russo, 2026)) is already known to provably improve upon the base data collection policy (Corollary 4.5 in (Russo, 2026)). At the same time, prior work such as (Rafailov et al., 2024) demonstrate how for applications such as language model alignment, one could contrast a positive and negative answer via preference models (i.e., positive preferred over negative) and obtain strict improvements over the base policy. However, it is unclear if the techniques in the above literature can be generalized to characterize when the exact minimizer of a linear combination of success conditioning loss and contrastive loss can outperform success conditioning (or even the base policy) on the same rollout distribution. This distinction matters even without sampling or optimization error: a margin based contrastive loss can rank actions correctly yet assign less probability to better actions, and improvement over the base policy need not imply improvement over success conditioning. Motivated by the practical success of such combined losses, e.g, (Pang et al., 2024; Li et al., 2023a) we aim to theoretically understand and characterize their performance.

We study this comparison for binary rewards in contextual bandits and finite-horizon MDPs. A fixed base policy π_0 generates the data, and its success-conditioned policy π_0^+ is the population solution to maximum-likelihood training on successful trajectories. We analyze two different versions of the combined loss; (a) un-weighted, and (b) re-weighted. The un-weighted loss adds a pairwise hinge loss on independently sampled successful and failed trajectories (with a tunable margin of

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.

δ) to a maximum likelihood loss on successful trajectories only (with a tunable scalar multiplier λ). The re-weighted loss modifies the contrastive term of the un-weighted loss by dividing (i.e., re-weighting) the log-probability of the successful (failed respectively) trajectory under the policy π being optimized for by the log-probability of the successful (failed respectively) trajectory under the base policy π_0. This amounts to changing the margin δ in the usual contrastive loss by the gap between the log-probabilities of the successful and failed trajectories under π_0 making the margin more informed of the ranking induced by the base policy π_0. We characterize how these choices affect performance of the learned policies when measured under typical RL metrics such as "value" and other "ranking" based metrics. Our results identify conditions for guaranteed improvement over the base policy and the success conditioning policy and establish the limits of these guarantees, as summarized below.

1.1 Our Contributions

Let π_0 denote the base policy, π_0^+ its success-conditioned policy, and $\pi_{\mathrm{unw}}, \pi_{\mathrm{rw}}$ the population minimizers of the unweighted and reweighted objectives, respectively. Let $\delta \geq 0$ be the margin and $\lambda \geq$ be the scalar multiplier of the success conditioning loss. We refer the reader to Section 2 for some of the RL, Contextual bandit terminology and notations used below.

(1) Improvement over success conditioning - For contextual bandits, we establish thresholds $\delta^\star$, $\lambda^\star$ such that for $\delta \geq \delta^\star$ and $\lambda \geq \lambda^\star$ and for every state s, $V_{\pi_{\mathrm{unw}}}(s)$, $V_{\pi_{\mathrm{rw}}}(s) \geq V_{\pi_0^+}(s)$. The inequalities are strict when $\mathrm{Var}_{a\sim\pi_0^+(\cdot|s)}[Q(s,a)] > 0$. We also construct counterexamples with small values of the margin δ to show that neither comparison holds unconditionally.

(2) The role of reference-policy reweighting - For finite-horizon MDPs, for every $\lambda, \delta \geq 0$, the re-weighted policy π_{rw} satisfies $V_{\pi_{\mathrm{rw}}}(s) \geq V_{\pi_0}(s)$. In contrast, for the un-weighted policy π_{unw} we construct counterexamples where such a guarantee does not hold. In contextual bandits, we further show $V_{\pi_{\mathrm{rw}}}(s) \geq V_{\pi_{\mathrm{unw}}}(s)$ for every state s, when $\pi_0(\cdot \mid s)$ and $Q(s,\cdot)$ have the same ordering; the inequality reverses when their orderings are opposite.

(3) Limits of generalization to finite-horizon MDPs - We construct a finite-horizon MDP for which $V_{\pi_0^+}(\mu) > \max\{V_{\pi_{\mathrm{unw}}}(\mu), V_{\pi_{\mathrm{rw}}}(\mu)\}$, for every finite $\lambda \geq 0$ and every $\delta \geq 0$. A separate construction shows that the even when the orderings of $\pi_0(\cdot \mid s)$ and $Q(s,\cdot)$ are same, comparison between π_{rw} and π_{unw} (mentioned for contextual bandits above) need not extend to MDPs.

(4) Ranking guarantees distinct from value guarantees - We show that for sufficiently large λ, δ, π_{rw} also improves over π_0^+ in other metrics such as precision and recall that are relevant for applications such as ranking and retrieval. This aligns with our experiments in Section 4 and the performance gains that have been observed for such combines losses in the retrieval literature (Li et al., 2023a).

(5) Empirical evidence - We perform semi-synthetic bandit and MS MARCO generative-retrieval experiments in Section 4 and examine the effects of λ, δ, and the ordering of base probabilities relative to rewards. We demonstrate gains from combined objectives and distinguish the objective preferred for expected reward from that preferred for retrieval ranking.

1.2 Related Work

Learning from successful and failed outcomes. STaR and ReST$^{\mathrm{EM}}$ train on successful model-generated solutions (Zelikman et al., 2022; Singh et al., 2024), while Russo (2026) characterizes success conditioning as a conservative policy-improvement update. Negative examples can improve probability redistribution (Tajwar et al., 2024) and, when constructed at the step level, credit assignment (Setlur et al., 2024). In robotics, failed demonstrations constrain inverse reinforcement learning (Shiarlis et al., 2016), while GCR and TW-CRL use success–failure contrasts to learn rewards (Biza et al., 2025; Li et al., 2025). These works establish the utility of learning from both outcomes, but do not characterize the population minimizers of our combined policy objective relative to success conditioning. We provide this comparison directly, without an intermediate learned reward model: we identify conditions for strict improvement, establish the role of the base policy, and construct counterexamples showing why bandit guarantees need not extend to sequential decisions.

Contrastive objectives and ranking. DPO uses reference-policy likelihood ratios (Rafailov et al., 2024), SLiC-HF combines a likelihood-margin loss with cross-entropy regularization (Zhao et al., 2023), and Iterative Reasoning Preference Optimization adds likelihood training on correct responses to DPO (Pang et al., 2024). GPO unifies several preference losses and analyzes their regularization (Tang et al., 2024). Our contribution is not the combination of these ingredients, but explicit conditions under which reference reweighting and the contrastive margin yield higher value than success-only training. In generative retrieval, MINDER learns multiple identifier views (Li et al., 2023b), and LTRGR supplements generation with ranking supervision (Li et al., 2023a). Beyond these application-specific methods, we establish ranking-inversion guarantees and separate them from expected-reward guarantees, explaining why the loss preferred for policy value need not be preferred for retrieval.

2 PRELIMINARIES

In this section, we introduce key definitions and results used throughout rest of the paper. The set $\{1, \ldots, n\}$ is denoted by $[n]$.

Finite-horizon MDPs. A finite-horizon MDP is a tuple $(\mathcal{S}, \mathcal{A}, P, S_0, \mu, \mathcal{T})$, where $\mathcal{S}, \mathcal{A}$ are finite state and action sets, $S_0 \sim \mu$, $P(s' \mid s, a)$ is the transition probability, and $\mathcal{T} \subseteq \mathcal{S}$ is the set of terminal states. Each episode terminates within H decisions, producing a trajectory $\tau = (S_0, A_0, \ldots, S_{h-1}, A_{h-1}, S_h)$ with $h \leq H$ and $S_h \in \mathcal{T}$. We write $S_i(\tau), A_i(\tau)$ for its states and actions. The episode receives a binary outcome $\mathcal{R}(\tau) \in \{0, 1\}$ whose conditional distribution depends only on the terminal state; temporal history can be included in the state when necessary.

A policy $\pi(\cdot \mid s)$ is a probability distribution over $\mathcal{A}$, and $\mathbb{P}_\pi$ denotes the trajectory law induced by μ, π, P. Its value and action-value functions are the success probabilities when starting from s, and when additionally taking a as the first action before following π, respectively:

$$V_\pi(s) = \mathbb{P}_\pi(\mathcal{R}(\tau) = 1 \mid S_0 = s)\,, \qquad Q_\pi(s, a) = \mathbb{P}_\pi(\mathcal{R}(\tau) = 1 \mid S_0 = s, A_0 = a)\,.$$

Both take values in $[0, 1]$, and $V_\pi(\mu) = \mathbb{E}_{s\sim\mu} V_\pi(s)$. For notational convenience, terminated trajectories may be padded with dummy steps (with same outcome) giving every trajectory length H.

Contextual bandits - A contextual bandit is a special case of finite-horizon MDPs defined above with $H = 1$, i.e., the MDP has a single state S_0. It is often used to model sequential decision making problems where a learner interacts with an environment over $T \in \mathbb{N}$ rounds. At each round $t \in [T]$, the environment presents the learner a context (state) sampled (independently over rounds) from μ. In response, the learner plays an action and receives a binary reward. The learner's goal is to interact with the environment over T rounds and use the collected trajectories, rewards to estimate a policy π^T with high value. Alternatively, the learner focuses on minimizing regret defined as $Reg(T) = \max_\pi V_\pi - V_{\pi^T}$.

Offline contextual bandits - In this setting we assume that the learner only gets to see the trajectories $\{(s_t, a_t, r_t) : t \in [T]\}$ generated using a prior data collection policy π_0, i.e., each $a_t \sim \pi_0(. \mid S_t = s_t)$. The learner can use these trajectories to estimate π^T with a goal of minimizing the regret defined above.

Language models as contextual bandits - Let $\mathcal{V}$ be a vocabulary of tokens. A Language Model (LM) is an auto-regressive policy $\pi(Y \mid X)$ which when given an input text context s sampled from some distribution μ, sequentially generates tokens $y_1, \ldots, y_H \in \mathcal{V}$ such that $y_h \sim \pi(Y = y_h \mid X = (s, y_{<h}))$. Given an input text s and the generated response $\bar{y} = (y_1, \ldots, y_H)$, one can use a suitable reward model $\mathcal{R}(s, y) \sim \text{Bernoulli}(Q(s, y))$ to measure the quality of the response as specified by some downstream application Bai et al. (2022). There is a natural contextual bandit problem associated with an LM where the action set $\mathcal{A}$ comprises all possible responses.

We call a state $s \in \mathcal{S}$ to be *non-degenerate* under a policy π, if $V_\pi(s) \in (0, 1)$. Unless otherwise specified, we assume all states to be non-degenerate in the rest of the paper. The following definitions and observations apply both in the cases of MDPs and contextual bandits.

Definition 2.1 (Reward conditioned policies)**.** For a policy π, we define the success conditioned policy π^+ following (Russo, 2026) as $\pi^+(a \mid s) = \mathbb{P}_\pi(A_i(\tau) = a \mid S_i(\tau) = s, \mathcal{R}(\tau) = 1)$. In other words, it is the probability of selecting a from the state s under a given policy π with the condition

that (s, a) belong to a successful trajectory. Similarly, we can define the failure conditioned policy as $\pi^-(a \mid s) = \mathbb{P}_\pi(A_t(\tau) = a \mid S_t(\tau) = s, \mathcal{R}(\tau) = 0)$.

We now make some observations regarding the success and failure conditioned trajectories, similar to (Russo, 2026), stated in the following lemma.

Lemma 2.2 (Closed form expressions for reward conditioned policies)**.** *For an MDP policy π and a state s non degenerate under π we have:*

$$\pi^+(a \mid s) = \frac{\pi(a \mid s)\, Q_\pi(s,a)}{V_\pi(s)},\ \pi^-(a \mid s) = \frac{\pi(a \mid s)\,\big(1 - Q_\pi(s,a)\big)}{1 - V_\pi(s)}, \tag{1}$$

The above closed form expressions are applications of Bayes rule, proofs of which are provided in Appendix B. Moreover, we can also write π^+ and π^- as minimizers of next-action prediction loss on successful and failed trajectories.

Lemma 2.3 (Reward conditioning as minimizer of next-action prediction loss)**.** *Let π be a policy of an finite-horizon MDP such that $0 < V_\pi < 1$. We show that,*

$$\begin{aligned}
\pi^+ &\in \arg\min_{\hat{\pi}} \Big\{ -\mathbb{E}_{\tau \sim \mathbb{P}_\pi} \Big[\sum_{h \in [H]} \ln \hat{\pi}(A_h(\tau) \mid S_h(\tau)) \;\Big|\; \mathcal{R}(\tau) = 1 \Big] \Big\}, \\
\pi^- &\in \arg\min_{\hat{\pi}} \Big\{ -\mathbb{E}_{\tau \sim \mathbb{P}_\pi} \Big[\sum_{h \in [H]} \ln \hat{\pi}(A_h(\tau) \mid S_h(\tau)) \;\Big|\; \mathcal{R}(\tau) = 0 \Big] \Big\}.
\end{aligned} \tag{2}$$

where the minimization is over the space of all possible policies. Note that, here the expectation is with respect to the distribution $\mathbb{P}_\pi$ on trajectories that arise as a result of sampling actions as per policy π and transitioning states as per the state transition distribution P.

Policy optimization via contrastive loss - We consider MDPs with a base policy π_0, from which trajectories are rolled out (with state transitions as per P) and data is collected. These observed rollouts give us state, action pairs from both successful and unsuccessful trajectories. Recent work in the Reinforcement Learning (RL) and Supervised Fine-Tuning (SFT) like Pang et al. (2024) utilizes the collection of observed rollouts to further fine-tune the base policy π to a new policy $\hat{\pi}$ via contrastive learning based approaches. Among such formulations, the Direct Preference Optimization (DPO) loss from (Rafailov et al., 2024) and the triplet hinge loss from (Li et al., 2023a) have received considerable attention lately due to the performance improvements observed in practical settings. The DPO loss re-weighs the log-likelihoods of the optimizing policy, i.e., $\ln \hat{\pi}(\cdot \mid s)$ with that of the base policy (i.e., $\ln \pi(\cdot \mid s)$) in the contrastive term while the triplet loss directly uses the un-weighted log-likelihoods of the optimizer. Following their approach, we now formally define the un-weighted and re-weighted losses as follows:

$$\mathcal{L}_{\mathrm{unw}}(\hat{\pi}, \pi) \coloneqq \mathop{\mathbb{E}}_{\substack{(s,a_+)\sim\tau_+ \\ (s,a_-)\sim\tau_-}} [\delta + \ln \hat{\pi}(a_- \mid s) - \ln \hat{\pi}(a_+ \mid s)]_+ - \lambda \mathop{\mathbb{E}}_{(s,a_+)\sim\tau_+} \ln \hat{\pi}(a_+ \mid s) \tag{3}$$

$$\mathcal{L}_{\mathrm{rw}}(\hat{\pi}, \pi) \coloneqq \mathop{\mathbb{E}}_{\substack{(s,a_+)\sim\tau_+ \\ (s,a_-)\sim\tau_-}} \left[\delta + \ln \frac{\hat{\pi}(a_- \mid s)}{\pi(a_- \mid s)} - \ln \frac{\hat{\pi}(a_+ \mid s)}{\pi(a_+ \mid s)}\right]_+ - \lambda \mathop{\mathbb{E}}_{(s,a_+)\sim\tau_+} \ln \hat{\pi}(a_+ \mid s) \tag{4}$$

where τ_+, τ_- are successful and failed trajectories both sampled using the base policy π (and state transitions as per P) and $\delta \geq 0$ is a margin hyperparameter forcing further separation between the actions a^+ and a^- belonging to τ^+ and τ^- respectively. Note that, in the contrastive terms (i.e., the first expectation) of 3 and 4, a^+ and a^- are sampled independently of each other for the common state s. We define $\pi_{\mathrm{unw}}(\pi)$ and $\pi_{\mathrm{rw}}(\pi)$ to be the minimizers (with respect to $\hat{\pi}$) of $\mathcal{L}_{\mathrm{unw}}(\hat{\pi}, \pi)$ (from eq. (3)) and $\mathcal{L}_{\mathrm{rw}}(\hat{\pi}, \pi)$ (from eq. (4)) respectively.

3 Main results

In this section we present our main technical results. Unless otherwise specified, we will assume a base policy π_0. For this base policy, we can define the success and failure conditioning policies π_0^+ and π_0^- (as per Definition 2.1) and the unweighted and re-weighted contrastive policies $\pi_{\mathrm{unw}}(\pi_0)$

and $\pi_{\mathrm{rw}}(\pi_0)$ as minimizers of losses $\mathcal{L}_{\mathrm{unw}}$ (3) and $\mathcal{L}_{\mathrm{rw}}$ (4) respectively. Since the base policy π_0 is fixed for this section, for brevity, we will denote $\pi_{\mathrm{unw}}(\pi_0)$ and $\pi_{\mathrm{rw}}(\pi_0)$ by π_{unw} and π_{rw} respectively. The goal of this section is to compare the performance of π_{unw} and π_{rw} with that of π_0^+ and π_0 under different metrics. Note that, from Corollary 4.5 in Russo (2026), we already know that, in any given finite-horizon MDP, $V_{\pi_0^+}(s) \geq V_{\pi_0}(s)$, for every state s. Here $V_\pi(s)$ is the value of a policy π at state s (defined in Section 2).

We now give a brief overview of our results in this section.

(1) $\pi_{\mathrm{unw}}, \pi_{\mathrm{rw}}$ vs. π_0 - In Theorem 3.1, we obtain a result similar to Corollary 4.5 of Russo (2026) for π_{rw}, i.e., we show that, for any given epsiodic MDP and any fixed $\lambda, \delta \geq 0$, $V_{\pi_{\mathrm{rw}}}(s) \geq V_{\pi_0}(s)$ for every state s. Unfortunately, this result does not hold in general for the unweighted policy π_{unw}. In Example 3.3, we construct a contextual bandit instance where this inequality fails. We strengthen this example in Lemma B.4, by showing that under simple assumptions on the base policy π_0, for all λ, δ sufficiently small, the value $V_{\pi_{\mathrm{unw}}}$ of the corresponding unweighted policy is strictly smaller than V_{π_0}.

(2) $\pi_{\mathrm{unw}}, \pi_{\mathrm{rw}}$ vs. π_0^+ - Interestingly for λ, δ large enough, in Theorem 3.2, we show that for contextual , both $V_{\pi_{\mathrm{unw}}}(s) > V_{\pi_0^+}(s)$ and $V_{\pi_{\mathrm{rw}}}(s) > V_{\pi_0^+}(s)$ for every s. Unfortunately, the situation for smaller values of λ, δ is not as clear. In fact, since Example 3.3 exhibits $V_{\pi_{\mathrm{unw}}} < V_{\pi_0}$, for this same example we obtain that $V_{\pi_{\mathrm{unw}}} < V_{\pi_0^+}$ since $V_{\pi_0}(s) \leq V_{\pi_0^+}(s)$ for all states s (by Corollary 4.5, Russo (2026)). In Lemma 3.4, we construct a contextual bandit instance where for any given $\lambda \geq 0$ and δ sufficiently small, we are able to show that $V_{\pi_{\mathrm{rw}}} < V_{\pi_0^+}$. In summary, if one can appropriately tune λ, δ, the contrastive loss based policies $\pi_{\mathrm{unw}}, \pi_{\mathrm{rw}}$ are better than π_0^+ in terms of value.

(3) π_{unw} vs. π_{rw} - In Theorem 3.6, for contextual bandit instances, we are able to identify situations where one of these policies always dominates the other. More precisely, when π_0 and Q_{π_0} are aligned (anti-aligned respectively), then $V_{\pi_{\mathrm{rw}}} \geq V_{\pi_{\mathrm{unw}}}$ ($V_{\pi_{\mathrm{unw}}} \geq V_{\pi_{\mathrm{rw}}}$ respectively).

3.1 π_{rw} DOMINATES π_0 IN MDPS

In case of learning with the re-weighted loss $\mathcal{L}_{\mathrm{rw}}$, the "re-weighting" by $\ln \pi_0$ introduces a form of an adaptive margin adjustment for the different (s, a_+, a_-) tuples. The contrastive term (first expectation in 4) increases the margin of separation more in tuples where the base policy π_0 already puts a higher weight-age on the successful action. In such cases, the higher margin induced by $\ln \pi_0(a_+ \mid s) - \ln \pi_0(a_- \mid s)$ encourages a higher separation and makes the learned policy improve further over the base. This phenomenon is observed for any finite-horizon MDPs. We state this formally in the following theorem. Proof can be found in Appendix B.

Theorem 3.1. *Let λ, δ be any scalars ≥ 0 and $\mathcal{L}_{\mathrm{rw}}$ be the corresponding re-weighted loss (eq. (4)). For any finite-horizon MDP with non-degenerate states, let π_{rw} be the state-wise minimizer of $\mathcal{L}_{\mathrm{rw}}$. Then, we have $V_{\pi_{\mathrm{rw}}}(s) \geq V_{\pi_0}(s)$ for every state s.*

3.2 LARGE λ, δ IMPLIES $\pi_{\mathrm{unw}}, \pi_{\mathrm{rw}}$ DOMINATE π_0^+ IN CONTEXTUAL BANDITS

Note that, as $\lambda \to \infty$, both the losses $\mathcal{L}_{\mathrm{unw}}, \mathcal{L}_{\mathrm{rw}}$ converge to the success conditioned loss (top row in 2) whose optimizer is the success conditioned policy π_0^+. The question however arises whether the contrastive terms in $\mathcal{L}_{\mathrm{unw}}, \mathcal{L}_{\mathrm{rw}}$ help in improving $\pi_{\mathrm{unw}}, \pi_{\mathrm{rw}}$ strictly beyond π_0^+. We show that, for contextual bandits, for appropriately large λ and δ, the value function corresponding to both $\pi_{\mathrm{unw}}, \pi_{\mathrm{rw}}$ can improve strictly beyond the value function of π_0^+ in the case of contextual bandits. The result is formally stated below. Proof can be found in Appendix B.

Theorem 3.2. *Consider a contextual bandit problem and any context s with non-constant $Q(s, \cdot)$. Then there exist scalars $\lambda^*_{\mathrm{rw}}, \lambda^*_{\mathrm{unw}} > 0$, $\delta^*_{\mathrm{rw}}, \delta^*_{\mathrm{unw}} \geq 0$, such that for all $\lambda \geq \lambda^*_{\mathrm{rw}}$, $\delta \geq \delta^*_{\mathrm{rw}}$, the corresponding policies π_{unw} and π_{rw} satisfy the following:*

1. $V_{\pi_{\mathrm{rw}}}(s) \geq V_{\pi_0^+}(s)$
2. $V_{\pi_{\mathrm{unw}}}(s) \geq V_{\pi_0^+}(s)$

In particular, the difference between both the value function differences $V_{\pi_{\mathrm{rw}}} - V_{\pi_0^+}$ and $V_{\pi_{\mathrm{unw}}} - V_{\pi_0^+}$ can be lower bounded by a constant proportional to $\mathrm{Var}_{a\sim\pi_0^+(\cdot,s)}(Q(s,a))$.

3.3 π_{unw} CAN BE WORSE THAN π_0, π_0^+

Lemma 3.3. *Consider a contextual bandit instance with two actions a_1, a_2 and a single state s. Using definitions in Section 2, we know that $Q(s,a) = \mathbb{P}(\mathcal{R}(s,a) = 1 | S_0 = s, A = a)$. We assume that $Q(s,a_1) > Q(s,a_2)$. For a base policy π, if $\delta < \ln\left(\frac{\pi(a_1|s)}{\pi(a_2|s)}\right)$, then $V_{\pi_{\mathrm{unw}}} < V_\pi$*

The training data might contain "misdirectional labels" due to stochasticity of reward. Comparisons favoring a_1 increase the log-probability gap until it reaches δ and saturates, while comparisons favoring a_2 continue to push the gap down. The pure contrastive objective is thus minimized when $\ln \pi_{\mathrm{unw}}(a_1 \mid s) - \ln \pi_{\mathrm{unw}}(a_2 \mid s) = \delta$. If the base policy's log-probability gap already exceeds δ, this update actually reduces the probability of choosing the better action. A stronger version with positive λ is stated and proved in lemma B.4 of Appendix B.3.

In these examples, the base policy π_0 is already assigns higher weight-age to the higher value actions, more than what level of separation the margin δ demands in the contrastive loss component and as a result π_{unw} fails to improve on π_0. This motivates the need to define the re-weighted loss $\mathcal{L}_{\mathrm{rw}}$ (eq. (4)) which adapts the margin according to π_0 in its contrastive term.

3.4 π_{rw} CAN BE WORSE THAN π_0^+

Similar to the previous contextual bandit instances, in Lemma 3.4, we construct an instance where π_{rw} is worse than the success conditioning policy π_0^+. Proof can be found in Appendix B

Lemma 3.4. *Consider a two-action contextual bandit with a single state s and base policy π_0 with full support. Suppose we have two actions a_1, a_2, with $1 > Q(s,a_1) > Q(s,a_2) > 0$,. Then, for any $\lambda \geq 0$ and any margin $0 \leq \delta < \ln\big(Q(s,a_1)/Q(s,a_2)\big)$, the optimizer π_{rw} of the corresponding re-weighted hinge loss satisfies $V_{\pi_{\mathrm{rw}}} < V_{\pi_0^+}$.*

The above suggests that, if the margin δ in the contrastive losses $\mathcal{L}_{\mathrm{rw}}, \mathcal{L}_{\mathrm{unw}}$ is not high enough, the respective optimizers cannot beat π_0^+ in terms of expected reward. In contrast, recall from Theorem 3.2 that for high enough margin and appropriate λ, both π_{rw} and π_{unw} can improve over π_0^+.

3.5 COMPARISON BETWEEN π_{unw} AND π_{rw}

Theorems 3.1, 3.2 and Lemmas B.4, 3.4 give rise to the natural question: which of the two losses $(\mathcal{L}_{\mathrm{unw}}, \mathcal{L}_{\mathrm{rw}})$ should one use? On the one hand, we have seen in Theorem 3.1 that π_{rw} is at least always guaranteed to improve over the base policy π_0 regardless of the hyperparameters λ, δ involved in its definition. On the other hand, we have seen in Theorem 3.2 that both the losses $\mathcal{L}_{\mathrm{unw}}, \mathcal{L}_{\mathrm{rw}}$ (for appropriately chosen λ, δ) can beat π_0^+.

In the following result, we identify a situation, which can help in clearly determining between the two choices.

Definition 3.5 (Aligned/ Anti-aligned policies)**.** We say that a policy π is *aligned* if for every state s and actions a, a',

$$(Q(s,a) - Q(s,a'))(\pi_0(a \mid s) - \pi_0(a' \mid s)) \geq 0.$$

If for every state s and all pairs of actions a, a', the reverse inequality holds, then we call the policy π *anti-aligned.*

Theorem 3.6, uses Definition 3.5 to select between $\pi_{\mathrm{unw}}, \pi_{\mathrm{rw}}$. Proof can be found in Appendix B.

Theorem 3.6. *Consider a contextual bandit problem and a base policy π_0. For any fixed scalars $\lambda > 0\, \delta \geq 0$, let π_{unw} and π_{rw} be the corresponding un-weighted and re-weighted policies as defined in Section 2 If π_0 is aligned then $V_{\pi_{\mathrm{rw}}}(s) \geq V_{\pi_{\mathrm{unw}}}(s)$ for every state s. Conversely, if π_0 and Q are anti-aligned then $V_{\pi_{\mathrm{rw}}}(s) \leq V_{\pi_{\mathrm{unw}}}(s)$ for every state s.*

Therefore, if π_0 already puts higher weight-age on higher reward actions, then π_{rw} with appropriately chosen λ, δ is guaranteed to be better than both π_0^+ and π_{unw}.

3.6 Generalizing our results from Contextual bandits to MDPs

While we were able to show Theorem 3.1 for general finite-horizon MDPs, Theorems 3.2 and 3.6 were proved only in the case of contextual bandits. While this already helps us in important practical settings such as fine tuning language models to maximize rewards, its not clear to what extent these results might hold for general RL problems involving finite-horizon MDPs. Below we demonstrate constructions of MDPs where these theorems fail. In Lemma 3.7, we construct an MDP such that for any values of the hyperparameters $\lambda, \delta \geq 0$, the corresponding contrastive policies π_{unw} and π_{rw} are inferior to π_0^+. Details of the construction and complete proof is presented in Appendix B.

Lemma 3.7. *There exists a finite finite-horizon tabular MDP (with a DAG structure) with binary terminal reward, a full-support base policy π_0, and an initial distribution μ, such that the following holds simultaneously for π_{rw} and π_{unw}. For every finite $\lambda \geq 0$, every $\delta \geq 0$, the statewise optimizer π_λ of the contrastive-plus-SFT objective (from eqs.* (3) *and* (4)*) satisfies $V_{\pi_0^+}(\mu) - V_{\pi_\lambda}(\mu) > 0$*

Remark 3.8. While the one-step reward $Q(s, a)$ is independent of the deployed policy for contextual bandits, for trajectory rollout in MDPs it is dependent on the policy since $Q_\pi(s, a) = \sum_{s' \in \mathcal{S}} \pi(a \mid s)\mathbb{P}(s' \mid s, a)V_\pi(s')$. Thus, one-step guarantees obtained by minimizing state-wise action-wise losses do not directly give us trajectory-level guarantees in the general case. It should be noted that, as $\lambda \to \infty$, both the optimizers π_{rw} and π_{unw} tend to π_0^+ anyways, but they neeed not always approach from "the above" in case of MDPs, and the entire trajectory of V_{π_λ}, $\lambda : 0 \to 1$ can stay below $V_{\pi_0^+}$.

In Lemma 3.9 below we show that even Theorem 3.6 does not hold for general epsiodic MDPs. In particular, we show that π_{unw} can beat π_{rw} in expected trajectory level value despite the base policy π_0 being aligned with rewards. Proof is provided in Appendix B. This also critically uses the dependence between the action value function Q_π and the deployed policy π that was highlighted in the remark above.

Lemma 3.9. *There exists a finite finite-horizon MDP with two decision states, binary terminal rewards, and a full-support base policy π_0 (that is strictly aligned with Q_0 at both state) and some hyperparameter $\lambda, \delta > 0$ such that the state-wise hinge optimizers satisfy $V_{\pi_{\mathrm{unw}}}(s_0) > V_{\pi_{\mathrm{rw}}}(s_0)$.*

3.7 Optimization of metrics other than value

In practical applications, value is not the only objective a policy needs to maximize. For e.g.,in generative retrieval Li et al. (2023a); Tay et al. (2022) the policy $\pi(\cdot \mid q)$ assigns probabilities to document identifiers given a text query q, and these scores are used to construct a ranked list of retrieved documents., the policy π takes as input a text query q and and outputs a list $\widehat{\mathcal{D}}_K$ of K documents from a corpus $\mathcal{D}$ using their likelihood given by $\pi(\cdot \mid q)$ (distribution over the document corpus). Given a binary reward $\mathcal{R}(q, d)$ (for e.g., relevance) distributed as a bernoulli variable with mean $Q(q, d)$, the value of π is $V_\pi = \mathbb{E}_{q\sim\mu} \sum_{d\in\mathcal{D}} Q(q, d)\pi(d \mid q)$. This does not capture the key retrieval metrics that one cares about. For e.g., one might want to maximize one of the following metrics. Let $\mathcal{D}(q)$ be the set of all documents relevant to query q (say all $d \in \mathcal{D}$ such that $Q(q, d) > \eta$ for some fixed threshold η). For each $k \in [K]$, define Recall@$k = \mathbb{E}_{q\sim\mu}\left(|\mathcal{D}(q) \cap \widehat{\mathcal{D}}_k|/|\mathcal{D}(q)|\right)$, and Precision@$k = \mathbb{E}_{q\sim\mu}\left(|\mathcal{D}(q) \cap \widehat{\mathcal{D}}_k|/k\right)$. Here $\widehat{\mathcal{D}}_k \subset \widehat{\mathcal{D}}_K$ are the top k documents in $\widehat{\mathcal{D}}_K$ as per the likelihood $\pi(\cdot \mid q)$. We observe that these metrics exhibit two properties; (a) they depend on a document only through their position in the sorted list, and (b) they do not decrease when a relevant item is moved above an irrelevant one. We call a metric with these two properties *swap-monotone*. Next, we prove that π_{unw} and π_{rw} improve such swap monotone metrics in contextual bandit settings (such as our retrieval problem).

Inversions and ranking using a policy - For any contextual bandit problem, fix a state s and assume the action set $\mathcal{A} = \{a_1, \ldots, a_N\}$ is ordered such that $Q(s, a_1) \geq Q(s, a_2) \geq \cdots Q(s, a_N)$. Considering the score vector given by likelihood $\pi(\cdot \mid s)$, the *inversions* are the pairs it gets strictly wrong, namely $\mathrm{Inv}(\pi) = \big\{(i, j) :\ Q(s, a_i) > Q(s, a_j),\ \pi(a_i \mid s) < \pi(a_j \mid s)\big\}$. Inversions order rankings without reference to a metric: It is easy to see that if $\mathrm{Inv}(\pi_1) \subseteq \mathrm{Inv}(\pi_2)$, then the list of actions sorted by $\pi_1(\cdot \mid s)$ is at least as good as the list sorted by $\pi_2(\cdot \mid s)$ for *every* swap-monotone metric at once (we prove this formally in the proof of Theorem 3.10). This is because, list with more

inversions can be turned into one with fewer by adjacent swaps, each moving a more relevant item above a less relevant one; by swap monotonicity, no such swap lowers the expected metric. And thus at the state s, π_1 has a higher swap monotone score than π_2. The list sorted by relevance has no inversions and is therefore optimal for all these metrics simultaneously, which is the probability ranking principle of Robertson (1997). Next, we show that, π_{unw} is always guaranteed to improve over π_0^+ and π_{rw} is always guaranteed to improve over π_0 in terms of minimizing inversions.

Theorem 3.10. *Consider a contextual bandit with state s and actions $a_1, \cdots, a_n$ such that $Q(s, a_1) \geq \cdots \geq Q(s, a_n)$. Then the following holds:*

1. *Fix any $\lambda > 0$ and let π_{unw} and π_{rw} be the un-weighted and re-weighted contrastive policies respectively. Then, $\mathrm{Inv}(\pi_{\mathrm{unw}}) \subseteq \mathrm{Inv}(\pi_0^+) \subseteq \mathrm{Inv}(\pi_0)$ and $\mathrm{Inv}(\pi_{\mathrm{rw}}) \subseteq \mathrm{Inv}(\pi_0)$.*
2. *$\exists \lambda^* > 0, \delta^* > 0$ such that $\forall \lambda \geq \lambda^*$ and $\delta \geq \delta^*$, we have $\mathrm{Inv}(\pi_{\mathrm{rw}}) \subseteq \mathrm{Inv}(\pi_0^+)$*

4 Experiments: applications in Generative Retrieval

In this section, we empirically compare the policies $\pi_{\mathrm{unw}}, \pi_{\mathrm{rw}}, \pi_0^+, \pi_0$ in the framework of generative retrieval, where training must improve the passages returned by the model and their ranking. We organize the evaluation around three questions: *(1)* Do $\pi_{\mathrm{unw}}, \pi_{\mathrm{rw}}$ improve over either of π_0 and π_0^+? *(2)*How does the comparison between $\pi_{\mathrm{unw}}, \pi_{\mathrm{rw}}$ change as the base policy π_0 changes? *(3)* Do the policies $\pi_{\mathrm{unw}}, \pi_{\mathrm{rw}}$ lead to better retrieval performance? These are asking the question of how well our theoretical guarantees from theorems 3.2, 3.6 and 3.10 hold in a real world setting, like generative retrieval. We also validate our theory in a synthetic setup, where we can directly control the *alignment* between π_0 and Q, which is discussed in Appendix D.1.

Retrieval setup. We use a compact-trie implementation of MINDER (Li et al., 2023b), with a BART-large generator (Lewis et al., 2020) that produces pseudo-query identifiers for the 8.84 million passages in MS MARCO (Nguyen et al., 2016). We start from three checkpoints of the same supervised training trajectory: EARLY at step 1,000, MIDPOINT at step 3,000, and FINAL. We train on the trained split and evaluate on the 6,980 MS MARCO dev-small queries. Appendix D gives the complete details. For each query, we construct a fixed identifier pool via beam search. The frozen Qwen3-Reranker-0.6B (Zhang et al., 2025) assigns a soft reward to passages, the identifier reward $\overline{Q}(q, z)$is the mean reward of its associated passages. Let $\pi_0(z \mid q)$ denote the base model's full-sequence probability normalized over the selected pool, and let We draw positive and negative identifiers independently from $\pi_0^+(\cdot \mid q)$ and $\pi_0^-(\cdot \mid q)$ respectively.

Objectives and methodology. We finetune the BART model using the contrastive losses $\mathcal{L}_{\mathrm{rw}}$ eq. (4) and $\mathcal{L}_{\mathrm{unw}}$ eq. (3) and base policy's weightage on picking the identifier z for the action q. We use the convex combination loss, with weight $\omega \in (0, 1)$ ($\lambda = \omega/(1-\omega)$ for the corresponding SFT weight λ from eqs. (3) and (4)) to avoid numerical issues.we see that*early*, *midpoint* and correspond *final* have their corresponding expected rewards $0.3940, 0.4376, 0.4743$ respectively, which is a real-world analogue of π_0 getting more "aligned" with the rewards and we ask the question whether the guarantees of Theorem 3.6 would hold under a more realistic setting like this. Using each one of the 3 checkpoints as π_0, we finetune the BART model using both the contrastive losses $\mathcal{L}_{\mathrm{unw}}, \mathcal{L}_{\mathrm{rw}}$ at $\omega \in \{0, 1/9, \ldots, 8/9\}$ and margins $\delta \in \{0.1, 2.575, 5.05, 7.525, 10\}$, and train one shared SFT endpoint ($\omega = 1$).

Expected reward of contrastive and SFT losses We first show the comparisons of how the average Qwen reward $V_{\pi_{\mathrm{rw}}}, V_{\pi_{\mathrm{unw}}}$ varies across this grid for the early, and final checkpoints and that compares to π_0^+ in fig. 1. We make the following observations: **(1)** We largely observe a "bell" shaped curve, which mean that at optimum combination, the mixed losses $\mathcal{L}_{\mathrm{unw}}, \mathcal{L}_{\mathrm{rw}}$ are better than their pure SFT or pure contrastive endpoints **(2)** High margin curves have higher reward than low margin ones. These two observations validate theorem 3.2 in a real-world empirical setup where we observe that high margin contrastive losses show significant gains in reward over SFT.

Retrieval comparison of contrastive and SFT losses - Here we wish to see if the gains of contrastive losses over SFT hold for retrieval metrics as well. We give the Recall@100 vs ω plot in fig. 5. We observe that: **(1)** The best performance across $3/4$ losses, checkpoints, the best performance was found for $\omega \in (0, 1)$ indicating the usefulness in combining the two losses. **(2)** Low

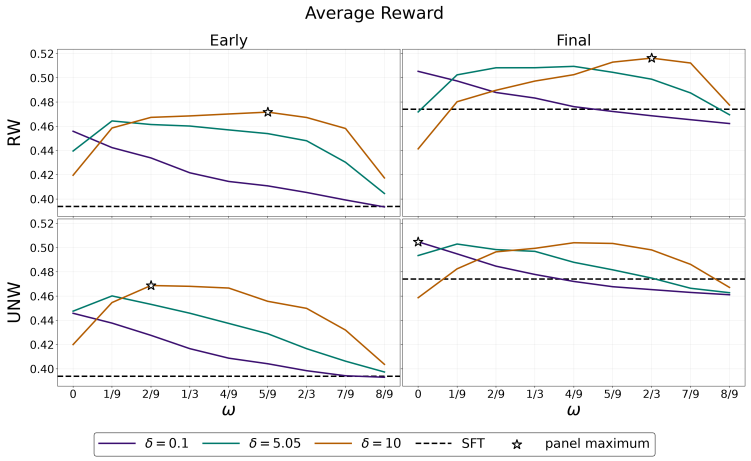


Figure 1: Variation of $V_{\pi_{\mathrm{rw}}}$ and $V_{\pi_{\mathrm{unw}}}$ with ω across different intial π_0. Each curve is one margin δ.

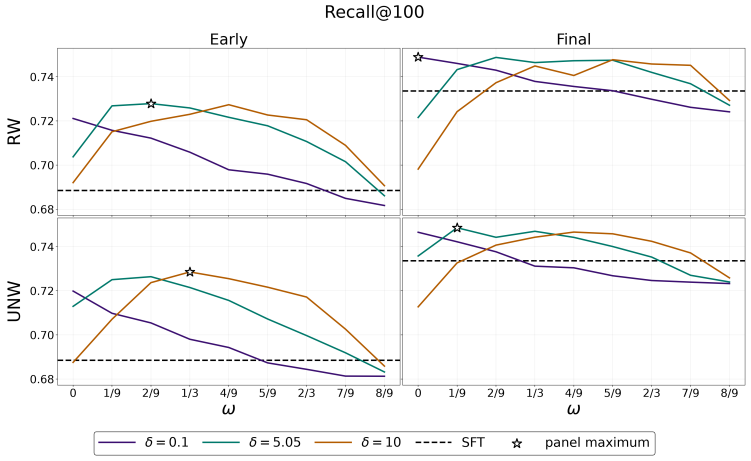


Figure 2: Variation of Recall with ω across different intial π_0. Each curve is one margin δ.

margin curves start high and degrade fast. This is significant, because this implies that to achieve an optimal trade-off between reward and ranking, a higher margin is preferable, since the low margin, low ω point are higher in Recall but had low reward, whereas high margin curves with appropriate λ achieve both high reward and ranking accuracy—best performance overall

Comparison between π_{rw} and π_{unw} as π_0 gets better - We report the average reward for baselines and the grid-winner (best across all λ, δ in validation) of the two hinge-based contrastive losses across checkpoints in table 1. We observe that as the baseline gets more reward-aligned, the performance of π_{rw} gets better, and it outperforms π_{unw} in Final, while lagging in early. SFT also improves with π_0. Both $\pi_{\mathrm{rw}}, \pi_{\mathrm{unw}}$ outperfirm std. baselines like RankNet Köppel et al. (2019) and DPO.

Table 1: Qwen rewards.

	Average reward		
Method	Early	Midpoint	Final
SFT	0.3939	0.4370	0.4741
DPO	0.1101	0.2380	0.1679
RankNet	0.1451	0.2426	0.2159
π_{rw}	0.4686	**0.4847**	**0.5163**
π_{unw}	**0.4688**	0.4748	0.5072

5 Conclusion

We establish losses combining contrastive training with success conditioning improves on success conditioning and how this depends on the hyperparameters and base policy. Our results distinguish reward improvement from ranking improvement and identify which guarantees extend from bandits to MDPs. Experiments in generative retrieval illustrate these distinctions, showing that learning from failures must be matched to the evaluation objective.

AI USE STATEMENT

In this work, we used generative AI tools for: (1) Polishing the paper, texts, fixing typos (2) Aid with strengthening and verifying mathematical arguments and proofs (3) For running experiments and AI-assisted coding. But generative AI don't form a core componet of our methodology.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments.

REFERENCES

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset, 2018. URL https://arxiv.org/abs/1611.09268.

Ondrej Biza, Thomas Weng, Lingfeng Sun, Karl Schmeckpeper, Tarik Kelestemur, Yecheng Jason Ma, Robert Platt, Jan-Willem van de Meent, and Lawson L. S. Wong. On-robot reinforcement learning with goal-contrastive rewards, 2025. URL https://arxiv.org/abs/2410.19989.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2005/08/icml_ranking.pdf.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pp. 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

Marius Köppel, Alexander Segner, Martin Wagener, Lukas Pensel, Andreas Karwath, and Stefan Kramer. Pairwise learning to rank by neural networks revisited: Reconstruction, theoretical analysis and practical performance, 2019. URL https://arxiv.org/abs/1909.02768.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703/.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. Learning to rank in generative retrieval, 2023a. URL https://arxiv.org/abs/2306.15222.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. Multiview identifiers enhanced generative retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6636–6648. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.acl-long.366. URL https://aclanthology.org/2023.acl-long.366/.

Yuxuan Li, Yijing Gao, Ning Yang, and Shanshan Xia. Tw-crl: Time-weighted contrastive reward learning for efficient inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2025. URL https://api.semanticscholar.org/CorpusID:277628097.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.neurips.cc/paper_files/paper/2024/hash/e099c1c9699814af0be873a175361713-Abstract-Conference.html.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated MAchine Reading COmprehension dataset. *arXiv preprint arXiv:1611.09268v1*, 2016. URL https://arxiv.org/abs/1611.09268v1.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization, 2024. URL https://arxiv.org/abs/2404.19733.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

S. E. Robertson. *The probability ranking principle in IR*, pp. 281–286. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. ISBN 1558604545.

Daniel Russo. Success conditioning as policy improvement: The optimization problem solved by imitating success, 2026. URL https://arxiv.org/abs/2601.18175.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization, 2017. URL https://arxiv.org/abs/1502.05477.

Amrith Setlur, Saurabh Garg, Xinyang Geng, Naman Garg, Virginia Smith, and Aviral Kumar. Learning to reason by failing: Offline RL on sub-optimal rollouts scales synthetic data by 8x. In *AI for Math Workshop @ ICML 2024*, 2024. URL https://openreview.net/forum?id=v2PV1yCFJk.

Kyriacos Shiarlis, Joao Messias, and Shimon Whiteson. Inverse reinforcement learning from failure. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, pp. 1060–1068, Richland, SC, 2016. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450342391.

Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron T Parisi, Abhishek Kumar, Alexander A Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Fathy Elsayed,

Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura A Culp, Lechao Xiao, Maxwell Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=lNAyUngGFK. Expert Certification.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of llms should leverage suboptimal, on-policy data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Remi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Avila Pires, and Bilal Piot. Generalized preference optimization: A unified approach to offline alignment. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47725–47742. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/tang24b.html.

Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. Transformer memory as a differentiable search index, 2022. URL https://arxiv.org/abs/2202.06991.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488. Curran Associates, Inc., 2022. doi: 10.52202/068431-1126. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/639a9a172c044fbb64175b5fad42e9a5-Paper-Conference.pdf.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. Slichf: Sequence likelihood calibration with human feedback. *ArXiv*, abs/2305.10425, 2023. URL https://api.semanticscholar.org/CorpusID:258741082.

A Appendix

B Proofs of technical results

Throughout the section, we have used x and s interchangeably to denote a state/context. Generally, a or a_i denotes actions.

B.1 Proofs of Preliminary Lemmas

Lemma 2.2 (Closed form expressions for reward conditioned policies)**.** *For an MDP policy π and a state s non degenerate under π we have:*

$$\pi^+(a \mid s) = \frac{\pi(a \mid s)\, Q_\pi(s,a)}{V_\pi(s)},\ \pi^-(a \mid s) = \frac{\pi(a \mid s)\,\big(1 - Q_\pi(s,a)\big)}{1 - V_\pi(s)}, \tag{1}$$

Proof. Consider a trajectory started from $S_0 = x$ and subsequently generated by π. By the definition of the policy and the action-value function, $\mathbb{P}_\pi\big(A_0 = a, \mathcal{R}(\tau) = 1 \mid S_0 = x\big) = \pi(a \mid x) Q_\pi(x,a)$. Similarly, since the terminal reward is binary, $\mathbb{P}_\pi\big(A_0 = a, \mathcal{R}(\tau) = 0 \mid S_0 = x\big) = \pi(a \mid x)\big(1 - Q_\pi(x,a)\big)$. The corresponding probabilities of the conditioning events are

$$\mathbb{P}_\pi\big(\mathcal{R}(\tau) = 1 \mid S_0 = x\big) = V_\pi(x),\ \mathbb{P}_\pi\big(\mathcal{R}(\tau) = 0 \mid S_0 = x\big) = 1 - V_\pi(x).$$

Both denominators are positive by the assumed non-degeneracy of x. Bayes' rule therefore gives

$$\pi^+(a \mid x) = \mathbb{P}_\pi\big(A_0 = a \mid S_0 = x, \mathcal{R}(\tau) = 1\big) = \frac{\pi(a \mid x) Q_\pi(x,a)}{V_\pi(x)},$$
$$\pi^-(a \mid x) = \mathbb{P}_\pi\big(A_0 = a \mid S_0 = x, \mathcal{R}(\tau) = 0\big) = \frac{\pi(a \mid x)\big(1 - Q_\pi(x,a)\big)}{1 - V_\pi(x)}.$$

By the Markov property, the same conditional action laws apply at every visit to x. □

Lemma 2.3 (Reward conditioning as minimizer of next-action prediction loss)**.** *Let π be a policy of an finite-horizon MDP such that $0 < V_\pi < 1$. We show that,*

$$\begin{aligned}
\pi^+ &\in \arg\min_{\hat\pi} \Big\{ -\mathbb{E}_{\tau \sim \mathbb{P}_\pi}\Big[\sum_{h \in [H]} \ln \hat\pi(A_h(\tau) \mid S_h(\tau)) \;\Big|\; \mathcal{R}(\tau) = 1 \Big] \Big\}, \\
\pi^- &\in \arg\min_{\hat\pi} \Big\{ -\mathbb{E}_{\tau \sim \mathbb{P}_\pi}\Big[\sum_{h \in [H]} \ln \hat\pi(A_h(\tau) \mid S_h(\tau)) \;\Big|\; \mathcal{R}(\tau) = 0 \Big] \Big\}.
\end{aligned} \tag{2}$$

where the minimization is over the space of all possible policies. Note that, here the expectation is with respect to the distribution $\mathbb{P}_\pi$ on trajectories that arise as a result of sampling actions as per policy π and transitioning states as per the state transition distribution P.

Proof. We treat the first problem; the second is identical with $\mathcal{R}(\tau) = 0$, $1 - Q_\pi$ and $1 - V_\pi$ in place of $\mathcal{R}(\tau) = 1$, Q_π and V_π. The conditioning event has probability $\mathbb{P}_\pi(\mathcal{R}(\tau) = 1) = V_\pi(\mu) > 0$, and every summand $-\ln \hat\pi(A_t \mid S_t)$ is nonnegative, with the conventions $-\ln 0 = +\infty$ and $0 \cdot (+\infty) = 0$; hence the objective is well defined in $[0, +\infty]$, and expectations of sums of nonnegative terms may be taken term by term.

For a decision state x and $t < H$, let $E_x^t = \{t < P,\ S_t = x\}$ be the event that x is visited at time t, and suppose that $\mathbb{P}_\pi(E_x^t) > 0$. By the Markov property, conditionally on E_x^t the continuation $(A_t, S_{t+1}, \ldots, S_P)$ together with the outcome $\mathcal{R}(\tau)$ is distributed as a trajectory started from $S_0 = x$ and generated by π, together with its outcome. Hence, by the definition of the policy and of the action-value function, exactly as in the proof of Lemma 2.2,

$$\mathbb{P}_\pi\big(A_t = a,\ \mathcal{R}(\tau) = 1 \;\big|\; E_x^t\big) = \pi(a \mid x)\, Q_\pi(x,a),\ \mathbb{P}_\pi\big(\mathcal{R}(\tau) = 1 \;\big|\; E_x^t\big) = V_\pi(x), \tag{5}$$

the second identity by summing the first over a.

Since a trajectory occupies exactly one state at each decision step,

$$-\sum_{t<P} \ln \hat{\pi}(A_t \mid S_t) = \sum_x \sum_{t<H} \mathbf{1}\{E_x^t\}\big(-\ln \hat{\pi}(A_t \mid x)\big),$$

where the outer sum runs over the decision states and the (x,t) summand is 0 off E_x^t. Fix x and t. If $\mathbb{P}_\pi(E_x^t) = 0$, the (x,t) summand vanishes almost surely. Otherwise, splitting E_x^t according to the value of A_t and using (5),

$$\begin{aligned}\mathbb{E}_\pi\big[\mathbf{1}\{E_x^t\}\big(-\ln \hat{\pi}(A_t \mid x)\big) \;\big|\; \mathcal{R}(\tau)=1\big] &= \sum_a \frac{\mathbb{P}_\pi\big(E_x^t,\ A_t=a,\ \mathcal{R}(\tau)=1\big)}{V_\pi(\mu)}\big(-\ln \hat{\pi}(a \mid x)\big) \\ &= \frac{\mathbb{P}_\pi(E_x^t)}{V_\pi(\mu)} \sum_a \pi(a \mid x)\, Q_\pi(x,a)\big(-\ln \hat{\pi}(a \mid x)\big).\end{aligned}$$

Summing over t and x, and writing $d_\pi(x) = \mathbb{E}_\pi \sum_{t<P} \mathbf{1}\{S_t = x\} = \sum_{t<H} \mathbb{P}_\pi(E_x^t)$ for the expected number of visits to x, we obtain

$$\begin{aligned}-\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t<P} \ln \hat{\pi}(A_t \mid S_t) \;\Big|\; \mathcal{R}(\tau)=1\Big] &= \sum_x \frac{d_\pi(x)}{V_\pi(\mu)} \sum_a \pi(a \mid x)\, Q_\pi(x,a)\big(-\ln \hat{\pi}(a \mid x)\big) \\ &= \sum_{x:\, d^+(x)>0} d^+(x) \sum_a \pi^+(a \mid x)\big(-\ln \hat{\pi}(a \mid x)\big),\end{aligned} \tag{6}$$

where $d^+(x) = d_\pi(x)\, V_\pi(x)/V_\pi(\mu)$. Indeed, if $V_\pi(x) = 0$ then $\pi(a \mid x) Q_\pi(x,a) = 0$ for every a, so the x term vanishes; otherwise $\pi(a \mid x) Q_\pi(x,a) = V_\pi(x)\, \pi^+(a \mid x)$, where $\pi^+(\cdot \mid x)$ is given by the first formula in (1), which only requires $V_\pi(x) > 0$ (by (5) and Bayes' rule, it is the conditional law of A_t given E_x^t and $\mathcal{R}(\tau) = 1$ for every t with $\mathbb{P}_\pi(E_x^t) > 0$). By (5), $d^+(x) = \mathbb{E}_\pi[\sum_{t<P} \mathbf{1}\{S_t = x\} \mid \mathcal{R}(\tau) = 1]$ is the expected number of visits to x on successful trajectories; at the states with $d^+(x) = 0$ the objective does not depend on $\hat{\pi}(\cdot \mid x)$.

For each x with $d^+(x) > 0$, the inner sum in (6) is the cross-entropy

$$\sum_a \pi^+(a \mid x)\big(-\ln \hat{\pi}(a \mid x)\big) = \mathrm{H}\big(\pi^+(\cdot \mid x)\big) + \mathrm{KL}\big(\pi^+(\cdot \mid x) \,\big\|\, \hat{\pi}(\cdot \mid x)\big),$$

where H is the Shannon entropy and the divergence equals $+\infty$ when $\hat{\pi}(\cdot \mid x)$ vanishes at an action charged by $\pi^+(\cdot \mid x)$. The objective is therefore

$$\sum_x d^+(x)\, \mathrm{H}\big(\pi^+(\cdot \mid x)\big) + \sum_x d^+(x)\ \mathrm{KL}\big(\pi^+(\cdot \mid x) \,\big\|\, \hat{\pi}(\cdot \mid x)\big).$$

The first sum does not depend on $\hat{\pi}$ and is finite, since each entropy is at most $\ln|\mathcal{A}|$ and $\sum_x d^+(x) = \mathbb{E}_\pi[P \mid \mathcal{R}(\tau) = 1] \le H$; by Gibbs' inequality the second is nonnegative and vanishes at $\hat{\pi} = \pi^+$. Hence π^+ is a minimizer, which is (2). (Gibbs' inequality also shows that every minimizer agrees with π^+ at every state with $d^+(x) > 0$.) □

B.2 Auxiliary lemmas

Lemma B.1 (covariance identity)**.** *For a probability vector p and vectors f, g, $\mathrm{Cov}_p(f,g) = \frac{1}{2}\sum_{a,a'} p_a p_{a'}(f_a - f_{a'})(g_a - g_{a'})$. In particular $\mathrm{Cov}_p(f,g) \ge 0$ whenever f is a nondecreasing function of g, with strict inequality if f is strictly increasing in g and g is nonconstant on the support of p.*

Proof. Expand the right-hand side and use $\sum_a p_a = 1$. □

Lemma B.2 (performance-difference identity)**.** *In an episodic MDP, for any two policies π, π' and any initial law, $V_\pi(\mu) - V_{\pi'}(\mu) = \sum(\cdot \mid s) d_{\pi,\mu}(s)\, \langle \pi(\cdot \mid s) - \pi'(\cdot \mid s), Q_{\pi'}(s,\cdot)\rangle$. In particular $|V_\pi(\mu) - V_{\pi'}(\mu)| \le \frac{1}{2}\sum(\cdot \mid s) d_{\pi,\mu}(s) \|\pi(\cdot \mid s) - \pi'(\cdot \mid s)\|_1$, and $\sum(\cdot \mid s) d_{\pi,\mu}(s) = \mathbb{E}_{\pi,\mu} T \le H$.*

Proof. Let $(S_0, A_0, \dots, S_T)$ be a trajectory of π from μ, S_T terminal. Extend $V_{\pi'}$ to terminal states by $V_{\pi'}(s) = \mathbb{P}(\mathcal{R} = 1 \mid s)$, so that $\mathbb{E}V_{\pi'}(S_T) = V_\pi(\mu)$ and $\mathbb{E}V_{\pi'}(S_0) = V_{\pi'}(\mu)$. Telescoping, $V_{\pi'}(S_T) - V_{\pi'}(S_0) = \sum_{t<T}\big(V_{\pi'}(S_{t+1}) - V_{\pi'}(S_t)\big)$, a finite sum; conditioning each increment on (S_t, A_t) and using $\mathbb{E}[V_{\pi'}(S_{t+1}) \mid S_t, A_t] = Q_{\pi'}(S_t, A_t)$ and $\sum_a \pi'(a \mid s)Q_{\pi'}(s,a) = V_{\pi'}(s)$,

$$\begin{aligned} V_\pi(\mu) - V_{\pi'}(\mu) &= \sum_{t\geq 0} \mathbb{E}\big[\mathbf{1}\{t < T\}\big(Q_{\pi'}(S_t, A_t) - V_{\pi'}(S_t)\big)\big] \\ &= \sum (\cdot \mid s) d_{\pi,\mu}(s)\Big(\sum_a \pi(a \mid s) Q_{\pi'}(s,a) - V_{\pi'}(s)\Big), \end{aligned}$$

which is the claim (Kakade & Langford, 2002). The bound follows because each inner product is at most $\frac{1}{2}\|\pi(\cdot \mid s) - \pi'(\cdot \mid s)\|_1 \operatorname{osc}(Q_{\pi'}(s,\cdot))$ and $\operatorname{osc} \leq 1$. □

Lemma B.3 (Hinge optimizer at large margin)**.** *Consider a contexual bandit and fix a* non-degenerate *context x and abase policy π_0. Let $P \coloneqq \{a : Q(x,a) > 0\}$, define $q_{\min} = \min_{a\in P} Q(x,a)$ and define*

$\lambda^* = \max_{a\in P}\left(\frac{\pi_0^-(a_i)}{\pi_0^+(a|x)} - \pi_0^-(P \mid x)\right) = \frac{V_{\pi_0}/q_{\min} - \pi(P_x|x)}{1 - V_{\pi_0}}$. *For $\lambda > \lambda^*$, define*

$$\pi^\lambda(a \mid x) = \begin{cases} \pi_0^+(a \mid x) + \dfrac{\pi_0^-(P \mid x)\pi_0^+(a \mid x) - \pi_0^-(a_i)}{\lambda}, & a \in P, \\ 0, & a \notin P. \end{cases} \tag{7}$$

This is a probability distribution, strictly positive on P. For either readout, let

$$D_{\mathrm{rw}}(\lambda) \coloneqq \operatorname{osc}_{a\in P}\left\{\log \pi^\lambda(a \mid x) - \log \pi_0(a \mid x)\right\},\ D_{\mathrm{unw}}(\lambda) \coloneqq \operatorname{osc}_{a\in P} \log \pi^\lambda(a \mid x).$$

If $\delta \geq \max\{D_{\mathrm{rw}}(\lambda), D_{\mathrm{unw}}(\lambda)\}$, then every policy π satisfies

$$\mathcal{L}_{\mathrm{rw}}(\pi) - \inf_{\hat\pi} \mathcal{L}_{\mathrm{rw}}(\hat\pi) \geq \lambda \operatorname{KL}\big(\pi^\lambda \| \pi\big),\ \mathcal{L}_{\mathrm{unw}}(\pi) - \inf_{\hat\pi} \mathcal{L}_{\mathrm{unw}}(\hat\pi) \geq \lambda \operatorname{KL}\big(\pi^\lambda \| \pi\big) \tag{8}$$

If P is the entire action set, the unique optimal policy is π^λ. If P is proper, no finite score vector minimizes the risk, but every minimizing sequence induces policies converging to π^λ.

Proof. The definitions of the success- and failure-conditioned policies give, for every $a \in P$,

$$\frac{\pi_0^-(a \mid x)}{\pi_0^+(a \mid x)} - \pi_0^-(P \mid x) = \frac{V_{\pi_0}/Q(x,a) - \pi_0(P \mid x)}{1 - V_{\pi_0}}.$$

Taking the maximum proves the stated expression for λ^*. Moreover, $V_{\pi_0} \geq q_{\min}\pi_0(P \mid x)$, so $\lambda^* \geq 0$. For every $a \in P$,

$$\frac{\pi^\lambda(a \mid x)}{\pi_0^+(a \mid x)} = 1 - \frac{1}{\lambda}\left(\frac{\pi_0^-(a \mid x)}{\pi_0^+(a \mid x)} - \pi_0^-(P \mid x)\right) \geq 1 - \frac{\lambda^*}{\lambda} > 0.$$

Since $\pi_0^+(P \mid x) = 1$, summing (7) over P gives

$$\sum_{a\in P} \pi^\lambda(a \mid x) = 1 + \frac{\pi_0^-(P \mid x) - \pi_0^-(P \mid x)}{\lambda} = 1.$$

Thus π^λ is a probability distribution with support P.

We next derive a lower bound for each loss. Every hinge is at least its affine argument. Apply this inequality to the pairs whose two actions belong to P, and discard the nonnegative hinge terms whose failure action lies outside P. There are no terms with a success action outside P, because π_0^+ vanishes there.

For the unweighted loss, collecting the coefficients of $\log \pi(a \mid x)$ gives

$$\pi_0^-(a \mid x) - \big(\lambda + \pi_0^-(P \mid x)\big)\pi_0^+(a \mid x) = -\lambda\pi^\lambda(a \mid x).$$

Consequently,

$$\mathcal{L}_{\mathrm{unw}}(\pi) \geq \delta\pi_0^-(P \mid x) - \lambda \sum_{a \in P} \pi^\lambda(a \mid x) \log \pi(a \mid x).$$

For the reweighted loss, the same calculation leaves the additional base-policy term:

$$\begin{aligned}\mathcal{L}_{\mathrm{rw}}(\pi) \geq \delta\pi_0^-(P \mid x) &- \lambda \sum_{a \in P} \pi^\lambda(a \mid x) \log \pi(a \mid x) \\ &+ \lambda \sum_{a \in P} \big(\pi^\lambda(a \mid x) - \pi_0^+(a \mid x)\big) \log \pi_0(a \mid x).\end{aligned}$$

In both inequalities, use

$$-\lambda \sum_{a \in P} \pi^\lambda(a \mid x) \log \pi(a \mid x) = -\lambda \sum_{a \in P} \pi^\lambda(a \mid x) \log \pi^\lambda(a \mid x) + \lambda\, \mathrm{KL}\big(\pi^\lambda \| \pi\big).$$

It remains to show that the resulting policy-independent lower bounds are the respective infima.

First suppose that P is the entire action set. At $\pi = \pi^\lambda$, the margin condition ensures that every unweighted score difference and every reweighted score difference has absolute value at most δ. Hence every hinge equals its affine argument, and both lower bounds are attained. Subtracting these attained bounds proves (8). Since $\lambda > 0$, equality at the minimum forces $\mathrm{KL}(\pi^\lambda \| \pi) = 0$, hence $\pi = \pi^\lambda$. This proves uniqueness.

Now suppose that P is proper. Consider strictly positive policies $\hat{\pi}$ of the form

$$\hat{\pi}(a \mid x) = \begin{cases} \hat{\pi}(P \mid x)\pi^\lambda(a \mid x), & a \in P, \\ \dfrac{1 - \hat{\pi}(P \mid x)}{1 - \pi_0(P \mid x)} \pi_0(a \mid x), & a \notin P, \end{cases}$$

and let $\hat{\pi}(P \mid x)$ increase to one.

Within P, their score differences are exactly those of π^λ, for both readouts. The margin condition therefore makes every internal hinge equal its affine argument. Meanwhile, the score of every action outside P tends to $-\infty$, whereas the scores on P remain finite. Thus every hinge pairing a success action in P with a failure action outside P is eventually zero.

It follows that both affine lower bounds are equalities along these policies once $\hat{\pi}(P \mid x)$ is sufficiently close to one. Furthermore,

$$\mathrm{KL}\big(\pi^\lambda \| \hat{\pi}\big) = -\log \hat{\pi}(P \mid x) \longrightarrow 0.$$

Therefore the policy-independent lower bounds are again the respective infima. Subtracting them proves (8).

Every finite-score policy is strictly positive and hence assigns positive mass outside P. It therefore differs from π^λ, so the KL divergence in (8) is strictly positive. Neither loss can attain its infimum at finite scores.

Finally, along any minimizing sequence for either loss, (8) forces the KL divergence from π^λ to tend to zero. Pinsker's inequality then gives convergence of the induced policies to π^λ in total variation. □

B.3 Proofs concerning comparisons of $\pi_0^+, \pi_0, \pi_{\mathrm{rw}}, \pi_{\mathrm{unw}}$

Theorem 3.1. *Let λ, δ be any scalars ≥ 0 and $\mathcal{L}_{\mathrm{rw}}$ be the corresponding re-weighted loss (eq. (4)). For any finite-horizon MDP with non-degenerate states, let π_{rw} be the state-wise minimizer of $\mathcal{L}_{\mathrm{rw}}$. Then, we have $V_{\pi_{\mathrm{rw}}}(s) \geq V_{\pi_0}(s)$ for every state s.*

Proof. Fix a decision state s. First suppose that the state-wise minimum is attained at finite scores. We claim that

$$Q_0(s,a) > Q_0(s,a') \implies \frac{\pi_{\mathrm{rw}}(a \mid s)}{\pi_0(a \mid s)} \geq \frac{\pi_{\mathrm{rw}}(a' \mid s)}{\pi_0(a' \mid s)}.$$

Suppose otherwise. The reweighted score of a is then strictly smaller than that of a'. Increase the unnormalized reweighted score of a at rate $1/\pi_0(a \mid s)$ and decrease that of a' at rate $1/\pi_0(a' \mid s)$, keeping the other unnormalized scores fixed and renormalizing the resulting policy. We show that the right directional derivative of $\mathcal{L}_{\mathrm{rw}}$ along this perturbation is strictly negative.

Consider first the contrastive terms involving a third action. When the third action is the success action, the hinge argument involving a as the failure action is strictly smaller than the corresponding argument involving a'. Moreover,

$$\frac{\pi_0^-(a \mid s)}{\pi_0(a \mid s)} = \frac{1 - Q_0(s,a)}{1 - V_{\pi_0}(s)} < \frac{1 - Q_0(s,a')}{1 - V_{\pi_0}(s)} = \frac{\pi_0^-(a' \mid s)}{\pi_0(a' \mid s)}.$$

The perturbation increases the smaller hinge argument and decreases the larger one. Since hinge slopes are nonnegative and nondecreasing, the sum of these two directional derivatives is nonpositive.

When the third action is the failure action, the hinge argument involving a as the success action is strictly larger than the corresponding argument involving a'. Here,

$$\frac{\pi_0^+(a \mid s)}{\pi_0(a \mid s)} = \frac{Q_0(s,a)}{V_{\pi_0}(s)} > \frac{Q_0(s,a')}{V_{\pi_0}(s)} = \frac{\pi_0^+(a' \mid s)}{\pi_0(a' \mid s)}.$$

The perturbation decreases the larger hinge argument and increases the smaller one. Their combined directional derivative is again nonpositive. These comparisons remain valid at hinge kinks: the right slope at a strictly smaller argument is at most the left slope at a strictly larger argument.

The self-pair terms are constant. For the two terms pairing a and a' with each other, the hinge with success action a and failure action a' is strictly active, because $\delta \geq 0$ and the score of a is smaller. Its slope is therefore one, while the slope of the reverse hinge is at most one. Their combined directional derivative is at most

$$\begin{aligned}
&\left(\frac{1}{\pi_0(a \mid s)} + \frac{1}{\pi_0(a' \mid s)}\right)\left(\pi_0^+(a' \mid s)\pi_0^-(a \mid s) - \pi_0^+(a \mid s)\pi_0^-(a' \mid s)\right) \\
&\quad = -\frac{\left(\pi_0(a \mid s) + \pi_0(a' \mid s)\right)\left(Q_0(s,a) - Q_0(s,a')\right)}{V_{\pi_0}(s)\left(1 - V_{\pi_0}(s)\right)} < 0.
\end{aligned}$$

Thus the contrastive term has strictly negative directional derivative.

The SFT term, including its coefficient λ, has directional derivative

$$\lambda\left(\frac{\pi_{\mathrm{rw}}(a \mid s)}{\pi_0(a \mid s)} - \frac{\pi_{\mathrm{rw}}(a' \mid s)}{\pi_0(a' \mid s)} - \frac{Q_0(s,a) - Q_0(s,a')}{V_{\pi_0}(s)}\right) \leq 0.$$

Consequently, the right directional derivative of the entire loss is strictly negative, contradicting optimality. This proves the claimed ordering for every $\lambda, \delta \geq 0$.

We next address limiting optimal policies. If π_{rw} assigns positive probability only to actions with $Q_0(s,a) = 1$, then

$$\sum_a \pi_{\mathrm{rw}}(a \mid s) Q_0(s,a) = 1 \geq V_{\pi_0}(s).$$

Otherwise, π_{rw} retains an action whose base quality is strictly below one. Every action with positive base quality must then also have positive probability under π_{rw}: if such an action were omitted, its reweighted score relative to the retained action would tend to $-\infty$ along the minimizing sequence. The hinge pairing the omitted success action with the retained failure action has a strictly positive weight and would diverge, contradicting minimization.

Hence, in this remaining case, every omitted action has base quality zero. The limiting policy minimizes the loss restricted to its positive-probability actions. Indeed, conditioning any strictly positive policy on those actions leaves their internal hinges unchanged, removes nonnegative external hinge terms, and cannot increase the SFT risk. Conversely, any strictly positive policy on the retained actions can be extended by vanishing mass on the omitted actions; their failure scores then tend to $-\infty$, so the external hinges vanish. The full and restricted infima therefore coincide, and continuity on the retained support proves restricted optimality.

The preceding directional argument applies on that support. Every omitted action has both zero base quality and zero likelihood ratio, so the same ordering extends to all actions.

Whenever this ordering applies, normalization gives

$$\sum_a \big(\pi_{\mathrm{rw}}(a \mid s) - \pi_0(a \mid s)\big) Q_0(s,a)$$
$$= \frac{1}{2} \sum_{a,a'} \pi_0(a \mid s)\pi_0(a' \mid s) \left(\frac{\pi_{\mathrm{rw}}(a \mid s)}{\pi_0(a \mid s)} - \frac{\pi_{\mathrm{rw}}(a' \mid s)}{\pi_0(a' \mid s)} \right) (Q_0(s,a) - Q_0(s,a')) \geq 0.$$

Together with the case where all retained actions have quality one, this proves, at every decision state,

$$\sum_a \pi_{\mathrm{rw}}(a \mid s) Q_0(s,a) \geq V_{\pi_0}(s).$$

Finally, apply backward induction over the finite horizon. The two policies have identical values at terminal states. If the value under π_{rw} is at least the base value at every successor state, the Bellman equation and the preceding state-wise inequality give

$$V_{\pi_{\mathrm{rw}}}(s) \geq \sum_a \pi_{\mathrm{rw}}(a \mid s) Q_0(s,a) \geq V_{\pi_0}(s).$$

Backward induction establishes the claim at every state. □

Theorem 3.2. *Consider a contextual bandit problem and any context s with non-constant $Q(s,\cdot)$. Then there exist scalars $\lambda^*_{\mathrm{rw}}, \lambda^*_{\mathrm{unw}} > 0$, $\delta^*_{\mathrm{rw}}, \delta^*_{\mathrm{unw}} \geq 0$, such that for all $\lambda \geq \lambda^*_{\mathrm{rw}}$, $\delta \geq \delta^*_{\mathrm{rw}}$, the corresponding policies π_{unw} and π_{rw} satisfy the following:*

1. $V_{\pi_{\mathrm{rw}}}(s) \geq V_{\pi_0^+}(s)$
2. $V_{\pi_{\mathrm{unw}}}(s) \geq V_{\pi_0^+}(s)$

In particular, the difference between both the value function differences $V_{\pi_{\mathrm{rw}}} - V_{\pi_0^+}$ and $V_{\pi_{\mathrm{unw}}} - V_{\pi_0^+}$ can be lower bounded by a constant proportional to $\mathrm{Var}_{a \sim \pi_0^+(\cdot,s)}(Q(s,a))$.

Proof. Fix the context s and call an action a *positive-quality* if $Q(s,a) > 0$. By definition, $\pi_0^+(a \mid s) = \pi_0(a \mid s) Q(s,a) / V_{\pi_0}(s)$ and $\pi_0^-(a \mid s) = \pi_0(a \mid s)\big(1 - Q(s,a)\big) / \big(1 - V_{\pi_0}(s)\big)$, so $\pi_0^+(\cdot \mid s)$ is supported exactly on the positive-quality actions, and

$$\sum_{a:\, Q(s,a)>0} \pi_0(a \mid s)\, Q(s,a) = V_{\pi_0}(s), \qquad \sum_{a:\, Q(s,a)>0} \pi_0^-(a \mid s) = \frac{\sum_{a:\, Q(s,a)>0} \pi_0(a \mid s) - V_{\pi_0}(s)}{1 - V_{\pi_0}(s)}. \tag{i}$$

Let $\lambda^*_{\mathrm{rw}}, \delta^*_{\mathrm{rw}}, \lambda^*_{\mathrm{unw}}, \delta^*_{\mathrm{unw}}$ be the thresholds of Lemma B.3: for $\star \in \{\mathrm{rw}, \mathrm{unw}\}$, $\lambda \geq \lambda^*_\star$ and $\delta \geq \delta^*_\star$, the optimizer $\pi_\star$ of the corresponding hinge loss puts zero mass on every action with $Q(s,a) = 0$ and, on every positive-quality action, mass

$$\pi_0^+(a \mid s) + \frac{1}{\lambda}\Big(\pi_0^+(a \mid s) \sum_{a':\, Q(s,a')>0} \pi_0^-(a' \mid s) - \pi_0^-(a \mid s)\Big). \tag{ii}$$

The same expression serves both readouts, so it suffices to prove the claims for the policy (ii). Substituting the definitions and the second identity in (i), the bracket in (ii) equals

$$\frac{\pi_0(a \mid s)}{V_{\pi_0}(s)\big(1 - V_{\pi_0}(s)\big)} \left(Q(s,a)\Big(\sum_{a':\, Q(s,a')>0} \pi_0(a' \mid s) - V_{\pi_0}(s) \Big) - V_{\pi_0}(s)\big(1 - Q(s,a)\big) \right)$$
$$= \frac{\pi_0(a \mid s)\Big(Q(s,a) \sum_{a':\, Q(s,a')>0} \pi_0(a' \mid s) - V_{\pi_0}(s) \Big)}{V_{\pi_0}(s)\big(1 - V_{\pi_0}(s)\big)}.$$

Multiplying by $Q(s,a)$, summing over the positive-quality actions, and using the first identity in (i), the value of the policy (ii) exceeds $V_{\pi_0^+}(s)$ by exactly

$$\frac{1}{\lambda\, V_{\pi_0}(s)\big(1-V_{\pi_0}(s)\big)}\Bigg(\sum_{a'\colon Q(s,a')>0}\pi_0(a'\mid s)\sum_{a\colon Q(s,a)>0}\pi_0(a\mid s)\,Q(s,a)^2-V_{\pi_0}(s)^2\Bigg). \tag{iii}$$

Items 1 and 2. By the Cauchy–Schwarz inequality with the positive weights $\pi_0(a \mid s)$ on the positive-quality actions,

$$V_{\pi_0}(s)^2=\Big(\sum_{a\colon Q(s,a)>0}\pi_0(a\mid s)\,Q(s,a)\Big)^2\le\sum_{a'\colon Q(s,a')>0}\pi_0(a'\mid s)\sum_{a\colon Q(s,a)>0}\pi_0(a\mid s)\,Q(s,a)^2,$$

with equality if and only if $Q(s,\cdot)$ is constant on the positive-quality actions. Hence (iii) is nonnegative, which is $V_{\pi_{\mathrm{rw}}}(s)\ge V_{\pi_0^+}(s)$ and $V_{\pi_{\mathrm{unw}}}(s)\ge V_{\pi_0^+}(s)$ on the respective ranges of (λ,δ), with strict inequality whenever $Q(s,\cdot)$ is non-constant on the positive-quality actions.

The variance bound. The variance of a random variable is at most its second moment about any fixed number. Taking the number $V_{\pi_0}(s)\big/\sum_{a'\colon Q(s,a')>0}\pi_0(a'\mid s)$, and using $\pi_0^+(a\mid s)=\pi_0(a\mid s)Q(s,a)/V_{\pi_0}(s)$ together with $Q(s,a)\le 1$,

$$\begin{aligned}\mathrm{Var}_{a\sim\pi_0^+(\cdot|s)}\big(Q(s,a)\big)&\le\frac{1}{V_{\pi_0}(s)}\sum_{a\colon Q(s,a)>0}\pi_0(a\mid s)\,Q(s,a)\Big(Q(s,a)-\frac{V_{\pi_0}(s)}{\sum_{a'\colon Q(s,a')>0}\pi_0(a'\mid s)}\Big)^2\\&\le\frac{1}{V_{\pi_0}(s)}\sum_{a\colon Q(s,a)>0}\pi_0(a\mid s)\Big(Q(s,a)-\frac{V_{\pi_0}(s)}{\sum_{a'\colon Q(s,a')>0}\pi_0(a'\mid s)}\Big)^2.\end{aligned}$$

Expanding the square and using (i), the last sum equals

$$\sum_{a\colon Q(s,a)>0}\pi_0(a\mid s)\,Q(s,a)^2-\frac{V_{\pi_0}(s)^2}{\sum_{a'\colon Q(s,a')>0}\pi_0(a'\mid s)},$$

and therefore

$$\begin{aligned}&\sum_{a'\colon Q(s,a')>0}\pi_0(a'\mid s)\sum_{a\colon Q(s,a)>0}\pi_0(a\mid s)\,Q(s,a)^2-V_{\pi_0}(s)^2\\&\ge V_{\pi_0}(s)\sum_{a'\colon Q(s,a')>0}\pi_0(a'\mid s)\ \mathrm{Var}_{a\sim\pi_0^+(\cdot|s)}\big(Q(s,a)\big).\end{aligned}$$

Inserting this into (iii), and then using $V_{\pi_0}(s)\le\sum_{a'\colon Q(s,a')>0}\pi_0(a'\mid s)$ (by (i) and $Q(s,a)\le 1$), we obtain, on the respective ranges of (λ,δ):

$$\begin{aligned}&\min\Big\{V_{\pi_{\mathrm{rw}}}(s)-V_{\pi_0^+}(s),V_{\pi_{\mathrm{unw}}}(s)-V_{\pi_0^+}(s)\Big\}\\&\quad\ge\frac{\sum_{a'\colon Q(s,a')>0}\pi_0(a'\mid s)}{\lambda\big(1-V_{\pi_0}(s)\big)}\ \mathrm{Var}_{a\sim\pi_0^+(\cdot|s)}\big(Q(s,a)\big)\\&\quad\ge\frac{V_{\pi_0}(s)}{\lambda\big(1-V_{\pi_0}(s)\big)}\ \mathrm{Var}_{a\sim\pi_0^+(\cdot|s)}\big(Q(s,a)\big),\end{aligned}$$

which is the claimed lower bound; it is strictly positive exactly when $Q(s,\cdot)$ is non-constant on the positive-quality actions. □

Lemma B.4. *Consider a contextual bandit instance with a single state s and two actions a_1,a_2 with $1>Q(s,a_1)>Q(s,a_2)>0$. Let the base policy π_0 be such that $\pi_0(a_1\mid x)>\pi_0(a_2\mid x)>0$. Then there exists $\lambda^*>0$, $\delta^*>0$ such that $V_{\pi_{\mathrm{unw}}}<V_{\pi_0}$ for every $\delta<\delta^*$ and $\lambda<\lambda^*$.*

Proof. We write Q_1 and Q_2 in short to denote $Q(x,a_1)$ and $Q(x,a_2)$ respectively. Write $A=\pi_0^+(1)\pi_0^-(2)$ and $B=\pi_0^+(2)\pi_0^-(1)$ for the probabilities of a correctly and of a wrongly ordered

pair, and $d_0 = \log(\pi_0(1)/\pi_0(2))$ which means $\pi_0(1) = \sigma(d_0)$. Note that we have, $A/B = Q_1(1-Q_2)/(Q_2(1-Q_1)) > 1$.

Let u_{unw} is the optimizer for $\ln \hat{\pi}$ in $\mathcal{L}_{\text{unw}}$ in eq. (3), giving $\hat{\pi}(a \mid x) = \frac{\exp(u_{\text{unw}}(x,a))}{\sum_{a' \in \mathcal{A}} \exp(u_{\text{unw}}(x,a'))}$. We write u_1, u_2 to denote $u_{\text{unw}}(x, a_1), u_{\text{unw}}(x, a_2)$ respectively. Writing score gap $g \coloneqq u_1 - u_2$ the un-weighted policy has mass $\sigma(g)$ on the better action, and its combined risk is, up to a constant, can be written as

$$\mathcal{L}_{\text{unw}}(g) = A[\delta - g]_+ + B[\delta + g]_+ + \lambda\big(\log(1+e^g) - \pi_0^+(1)g\big) :$$

The self-pairs in the contrastive term add constant factors ,and $\mathrm{KL}(\pi_0^+ \| \sigma(g)) = \log(1+e^g) - \pi_0^+(1)g + \text{const}$.

$\mathcal{L}_{\text{unw}}$ is convex and $\mathcal{L}_{\text{unw}}(g) \to \infty$ if $|g| \to \infty$ which means it has interioir optimizers. For $\lambda > 0$ the objective is strictly convex and the optimizer $\hat{g}$ is unique. By convexity all of them lie strictly to the left of d_0 if and only if the left derivative of $\mathcal{L}_{\text{unw}}$ at d_0 is positive. Note that, for the two action bandit, for any given policy π we have: $V_\pi - V_0 = (\pi(1) - \pi_0(1))(Q_1 - Q_2)$ Now, by our "score-gap" notation, we have: $\pi_0(1) = \sigma(d_0)$ and $\hat{\pi}(1) = \sigma(\hat{g})$ for the learned un-weighted policy $\hat{\pi}$ and its corresponding "score-gap" $\hat{g}$, This gives us the equivalence:

$$V_{\pi_{\text{unw}}} < V_0 \iff \sigma(\hat{g}) < \sigma(d_0) \iff \hat{g} < d_0.$$

Now, suppose $d_0 > \delta$, i.e. $\pi_0(1) > \sigma(\delta)$. Near d_0 the hinge $[\delta - g]_+$ vanishes and $[\delta + g]_+$ has slope 1, so the contrastive part contributes B to the left derivative of $\mathcal{L}_{\text{unw}}$ at d_0, while the KL anchor contributes $\lambda(\sigma(d_0) - \pi_0^+(1)) = -\lambda(\pi_0^+(1) - \pi_0(1))$ with $\pi_0^+(1) - \pi_0(1) = \pi_0(1)(Q_1 - V_0)/V_0 = \pi_0(1)\pi_0(2)(Q_1 - Q_2)/V_0 > 0$. Hence we can write the positivity condition on left-derivative of $\mathcal{L}_{\text{unw}}$ as follows:

$$\mathcal{L}'^{,-}_{\text{unw}}(d_0) = B - \lambda(\pi_0^+(1) - \pi_0(1)) > 0 \iff \lambda < \lambda^* = \frac{B}{\pi_0^+(1) - \pi_0(1)} = \frac{Q_2(1-Q_1)}{(1-V_0)(Q_1 - Q_2)},$$

where the last expression uses $B = \pi_0(1)\pi_0(2)\, Q_2(1-Q_1)/(V_0(1-V_0))$. Thus, we see that, defining $\delta^* \coloneqq \ln(\pi_0(1)/\pi_0(2))$ and $\lambda^* \coloneqq \frac{Q_2(1-Q_1)}{(1-V_0)(Q_1-Q_2)}$ gives us that $V_{\pi_{\text{unw}}} < V_{\pi_0}, \forall\, (\lambda, \delta) \in (0, \lambda^*) \times (0, \delta^*)$.

□

Lemma 3.4. *Consider a two-action contextual bandit with a single state s and base policy π_0 with full support. Suppose we have two actions a_1, a_2, with $1 > Q(s, a_1) > Q(s, a_2) > 0$,. Then, for any $\lambda \geq 0$ and any margin $0 \leq \delta < \ln\big(Q(s,a_1)/Q(s,a_2)\big)$, the optimizer π_{rw} of the corresponding re-weighted hinge loss satisfies $V_{\pi_{\text{rw}}} < V_{\pi_0^+}$.*

Proof. Every policy on the two actions has $\pi(a_2 \mid s) = 1 - \pi(a_1 \mid s)$, hence $V_\pi = Q(s, a_2) + \pi(a_1 \mid s)\big(Q(s,a_1) - Q(s,a_2)\big)$. Since $Q(s,a_1) > Q(s,a_2)$, the claim $V_{\pi_{\text{rw}}} < V_{\pi_0^+}$ is equivalent to

$$\pi_{\text{rw}}(a_1 \mid s) \;<\; \pi_0^+(a_1 \mid s) = \frac{\pi_0(a_1 \mid s)\, Q(s, a_1)}{V_{\pi_0}}. \tag{$*$}$$

By definition of the re-weighted hinge loss, the positive action is drawn from π_0^+ and the negative action from the failure-conditioned policy $a_i \mapsto \pi_0(a_i \mid s)\big(1 - Q(s,a_i)\big)/\big(1 - V_{\pi_0}\big)$, independently. Expanding the expectation over the four pairs, the loss of a policy π equals

$$\begin{aligned}
&\pi_0^+(a_1 \mid s)\, \frac{\pi_0(a_2 \mid s)\big(1 - Q(s,a_2)\big)}{1 - V_{\pi_0}} \Big[\delta - \ln \frac{\pi(a_1 \mid s)}{\pi_0(a_1 \mid s)} + \ln \frac{\pi(a_2 \mid s)}{\pi_0(a_2 \mid s)}\Big]_+ \\
&+ \pi_0^+(a_2 \mid s)\, \frac{\pi_0(a_1 \mid s)\big(1 - Q(s,a_1)\big)}{1 - V_{\pi_0}} \Big[\delta + \ln \frac{\pi(a_1 \mid s)}{\pi_0(a_1 \mid s)} - \ln \frac{\pi(a_2 \mid s)}{\pi_0(a_2 \mid s)}\Big]_+ \\
&- \lambda\Big(\pi_0^+(a_1 \mid s) \ln \pi(a_1 \mid s) + \pi_0^+(a_2 \mid s) \ln \pi(a_2 \mid s)\Big) \;+\; \text{a constant},
\end{aligned}$$

where the constant is the contribution of the two pairs with equal positive and negative action, whose hinge argument is $\delta \geq 0$.

The two hinge arguments are δ minus and δ plus the same quantity

$$\ln \frac{\pi(a_1 \mid s)}{\pi(a_2 \mid s)} - \ln \frac{\pi_0(a_1 \mid s)}{\pi_0(a_2 \mid s)},$$

which is a continuous, strictly increasing function of $\pi(a_1 \mid s)$ and, by $(*)$ together with $\pi_0^+(a_2 \mid s) = \pi_0(a_2 \mid s)Q(s,a_2)/V_{\pi_0}$, equals $\ln\big(Q(s,a_1)/Q(s,a_2)\big)$ at $\pi(a_1 \mid s) = \pi_0^+(a_1 \mid s)$. Because $\ln\big(Q(s,a_1)/Q(s,a_2)\big) > \delta \geq 0$, every policy with $\pi(a_1 \mid s) \geq \pi_0^+(a_1 \mid s)$ — and, by continuity, every policy with $\pi(a_1 \mid s)$ slightly below $\pi_0^+(a_1 \mid s)$ — has a negative first hinge argument and a positive second one. On this range the first hinge vanishes, the second equals its argument, and the loss is a differentiable function of $\pi(a_1 \mid s)$ whose derivative is

$$\frac{\pi_0^+(a_2 \mid s)\,\pi_0(a_1 \mid s)\big(1 - Q(s,a_1)\big)}{\big(1 - V_{\pi_0}\big)\,\pi(a_1 \mid s)\big(1 - \pi(a_1 \mid s)\big)} \;+\; \lambda\,\frac{\pi(a_1 \mid s) - \pi_0^+(a_1 \mid s)}{\pi(a_1 \mid s)\big(1 - \pi(a_1 \mid s)\big)},$$

the second term following from $\pi_0^+(a_1 \mid s) + \pi_0^+(a_2 \mid s) = 1$. The first term is strictly positive, because π_0 has full support and $Q(s,a_1), Q(s,a_2) \in (0,1)$; the second is nonnegative whenever $\pi(a_1 \mid s) \geq \pi_0^+(a_1 \mid s)$, because $\lambda \geq 0$. Hence the loss is strictly increasing in $\pi(a_1 \mid s)$ on $[\pi_0^+(a_1 \mid s), 1)$, and its derivative is continuous and strictly positive at $\pi(a_1 \mid s) = \pi_0^+(a_1 \mid s)$, so a policy with $\pi(a_1 \mid s)$ slightly below $\pi_0^+(a_1 \mid s)$ has strictly smaller loss than every policy with $\pi(a_1 \mid s) \geq \pi_0^+(a_1 \mid s)$. Therefore no minimizer of the re-weighted hinge loss satisfies $\pi(a_1 \mid s) \geq \pi_0^+(a_1 \mid s)$; in particular $(*)$ holds for π_{rw}, which proves the lemma. □

Theorem 3.6. *Consider a contextual bandit problem and a base policy π_0. For any fixed scalars $\lambda > 0\,\delta \geq 0$, let π_{unw} and π_{rw} be the corresponding un-weighted and re-weighted policies as defined in Section 2 If π_0 is aligned then $V_{\pi_{\mathrm{rw}}}(s) \geq V_{\pi_{\mathrm{unw}}}(s)$ for every state s. Conversely, if π_0 and Q are anti-aligned then $V_{\pi_{\mathrm{rw}}}(s) \leq V_{\pi_{\mathrm{unw}}}(s)$ for every state s.*

Proof. Both objectives decouple across states, so we fix a state s and prove the inequality there. By the standing assumption of Section 2, $\pi_0(a \mid s) > 0$ and $0 < Q(s,a) < 1$ for every $a \in \mathcal{A}$, so $\pi_0^+(\cdot \mid s)$ and $\pi_0^-(\cdot \mid s)$ have full support. Comonotonicity of π_0 and Q at s says that

$$\big(Q(s,a) - Q(s,a')\big)\big(\pi_0(a \mid s) - \pi_0(a' \mid s)\big) \;\geq\; 0 \qquad \text{for all } a, a' \in \mathcal{A}, \tag{i}$$

and the second hypothesis of the theorem reverses this inequality. We treat the comonotone case; reversing every sign below gives the other one.

A policy that gives zero mass to some action has infinite objective at s: for $\lambda > 0$ its supervised term is infinite, and for $\lambda = 0$ the pair with that action as the success and any action of positive mass as the failure has an infinite hinge. Both minimizations therefore run over policies with full support, and every such policy can be written, for any fixed $t \in [0,1]$, as

$$\pi_t^u(a \mid s) \coloneqq \frac{\pi_0(a \mid s)^t\, e^{u(a)}}{\sum_{a' \in \mathcal{A}} \pi_0(a' \mid s)^t\, e^{u(a')}} \tag{ii}$$

for a score $u : \mathcal{A} \to \mathbb{R}$, unique up to an additive constant. Inside the hinge of (3) the policy enters through $\log \pi(a^- \mid s) - \log \pi(a^+ \mid s)$, and inside the hinge of (4) through $\log \frac{\pi(a^- \mid s)}{\pi_0(a^- \mid s)} - \log \frac{\pi(a^+ \mid s)}{\pi_0(a^+ \mid s)}$; at $\pi = \pi_0^u$ the first difference equals $u(a^-) - u(a^+)$, and at $\pi = \pi_1^u$ so does the second, the normalizer of (ii) cancelling. Hence both hinge terms equal

$$C(u) \coloneqq \sum_{a, a' \in \mathcal{A}} \pi_0^+(a \mid s)\,\pi_0^-(a' \mid s)\,\big[\delta + u(a') - u(a)\big]_+, \tag{iii}$$

while the supervised term at π_t^u is

$$-\lambda \sum_a \pi_0^+(a \mid s) \log \pi_t^u(a \mid s) = \lambda\Big[\log \sum_a \pi_0(a \mid s)^t e^{u(a)} - \sum_a \pi_0^+(a \mid s)\,u(a)\Big] - \lambda t \sum_a \pi_0^+(a \mid s) \log \pi_0(a \mid s),$$

whose last term does not depend on u. Up to additive constants, the objectives (3) and (4) at s are therefore the members $t = 0$ and $t = 1$ of the family

$$F_t(u) \coloneqq C(u) + \lambda\Big[\log \sum_{a \in \mathcal{A}} \pi_0(a \mid s)^t e^{u(a)} - \sum_{a \in \mathcal{A}} \pi_0^+(a \mid s)\,u(a)\Big], \qquad t \in [0,1], \tag{iv}$$

with $\pi_{\mathrm{unw}}(\cdot \mid s) = \pi_0^u$ for a minimizer u of F_0 and $\pi_{\mathrm{rw}}(\cdot \mid s) = \pi_1^u$ for a minimizer u of F_1. Every F_t is unchanged when a constant is added to u, so we minimize over the scores with $\sum_a u(a) = 0$. On this set F_t is coercive: with $\bar{u} = \max_a u(a)$ and $\underline{u} = \min_a u(a)$, the pair formed by an action attaining $\underline{u}$ as the success and one attaining $\bar{u}$ as the failure gives $C(u) \geq \min_a \pi_0^+(a \mid s)\, \min_a \pi_0^-(a \mid s)\, (\delta + \bar{u} - \underline{u})$, and the bracket in (iv) is at least $t \min_a \log \pi_0(a \mid s) + \min_a \pi_0^+(a \mid s)\, (\bar{u} - \underline{u})$. It is convex, and for $\lambda > 0$ strictly convex, because the Hessian of the bracket at u is the covariance matrix of π_t^u, positive definite on nonconstant vectors as π_t^u has full support. So for $\lambda > 0$ each F_t has a unique minimizer, π_{rw} and π_{unw} are well defined, and both have full support. For $\lambda = 0$ the objective is C itself, the same for both readouts, and, as in Section 2, both policies are read from one common minimizer u of C: $\pi_{\mathrm{unw}}(\cdot \mid s) = \pi_0^u$ and $\pi_{\mathrm{rw}}(\cdot \mid s) = \pi_1^u$.

The case $\lambda = 0$ is now a covariance computation. By (ii), $\pi_{\mathrm{rw}}(a \mid s) = \pi_{\mathrm{unw}}(a \mid s)\, \pi_0(a \mid s) \big/ \sum_{a'} \pi_{\mathrm{unw}}(a' \mid s)\, \pi_0(a' \mid s)$, so

$$
\begin{aligned}
&V_{\pi_{\mathrm{rw}}}(s) - V_{\pi_{\mathrm{unw}}}(s) \\
=&\frac{\sum_a \pi_{\mathrm{unw}}(a \mid s)\pi_0(a \mid s)Q(s,a) - \Big(\sum_a \pi_{\mathrm{unw}}(a \mid s)Q(s,a)\Big)\Big(\sum_a \pi_{\mathrm{unw}}(a \mid s)\pi_0(a \mid s)\Big)}{\sum_a \pi_{\mathrm{unw}}(a \mid s)\pi_0(a \mid s)} \\
=&\frac{\frac{1}{2}\sum_{a,a'} \pi_{\mathrm{unw}}(a \mid s)\, \pi_{\mathrm{unw}}(a' \mid s)\big(Q(s,a) - Q(s,a')\big)\big(\pi_0(a \mid s) - \pi_0(a' \mid s)\big)}{\sum_a \pi_{\mathrm{unw}}(a \mid s)\pi_0(a \mid s)} \;\geq\; 0
\end{aligned}
$$

by (i).

For $\lambda > 0$ we move t from 0 to 1 in (iv) and follow the value of the minimizer. At the hinge the minimizer is not a differentiable function of t, so we smooth inside the proof and pass to the limit at the end. For $\varepsilon > 0$ replace $[z]_+$ in (iii) by $\varepsilon \log(1 + e^{z/\varepsilon})$, which is convex, has derivative $\sigma(z/\varepsilon)$ with $\sigma(z) = (1 + e^{-z})^{-1}$, and satisfies

$$
0 < \varepsilon \log(1 + e^{z/\varepsilon}) - [z]_+ = \varepsilon \log\big(1 + e^{-|z|/\varepsilon}\big) \leq \varepsilon \log 2. \tag{v}
$$

Write C_ε and $F_{t,\varepsilon}$ for C and F_t after this replacement. Since $C_\varepsilon \geq C$, the argument above applies verbatim: $F_{t,\varepsilon}$ is smooth, strictly convex and coercive on $\{\sum_a u(a) = 0\}$, with a unique minimizer $u_{t,\varepsilon}$ there; let $\pi_{t,\varepsilon} \coloneqq \pi_t^{u_{t,\varepsilon}}$, a policy with full support. Two facts finish the proof. First, for $t \in \{0, 1\}$, with $\pi_0 \coloneqq \pi_{\mathrm{unw}}(\cdot \mid s)$ and $\pi_1 \coloneqq \pi_{\mathrm{rw}}(\cdot \mid s)$,

$$
\sum_{a \in \mathcal{A}} \pi_t(a) \log \frac{\pi_t(a)}{\pi_{t,\varepsilon}(a \mid s)} \;\leq\; \frac{\varepsilon \log 2}{\lambda}. \tag{vi}
$$

Second, for every fixed $\varepsilon > 0$ the map $t \mapsto \pi_{t,\varepsilon}$ is continuously differentiable and

$$
\frac{d}{dt} V_{\pi_{t,\varepsilon}}(s) = \sum_{\{a,a'\}} \omega_{aa'}(t)\, \big(Q(s,a) - Q(s,a')\big)\big(\log \pi_0(a \mid s) - \log \pi_0(a' \mid s)\big), \qquad \omega_{aa'}(t) \geq 0, \tag{vii}
$$

the sum running over unordered pairs of distinct actions. Since $\log$ is increasing, (i) makes every term of (vii) nonnegative, and integrating over $t \in [0, 1]$ gives $V_{\pi_{1,\varepsilon}}(s) \geq V_{\pi_{0,\varepsilon}}(s)$ for every $\varepsilon > 0$. By (vi) and Pinsker's inequality, $\pi_{t,\varepsilon} \to \pi_t$ as $\varepsilon \downarrow 0$ for $t \in \{0, 1\}$, so

$$
V_{\pi_{\mathrm{rw}}}(s) - V_{\pi_{\mathrm{unw}}}(s) = \lim_{\varepsilon \downarrow 0} \big(V_{\pi_{1,\varepsilon}}(s) - V_{\pi_{0,\varepsilon}}(s)\big) \;\geq\; 0.
$$

For (vi), fix $t \in \{0, 1\}$ and let u_t be the minimizer of F_t on $\{\sum_a u(a) = 0\}$, so that $\pi_t = \pi_t^{u_t}$. Since (v) gives $0 \leq F_{t,\varepsilon} - F_t \leq \varepsilon \log 2$ pointwise (the pair weights $\pi_0^+(a \mid s)\pi_0^-(a' \mid s)$ sum to one),

$$
F_t(u_{t,\varepsilon}) - F_t(u_t) \;\leq\; F_{t,\varepsilon}(u_{t,\varepsilon}) - F_t(u_t) \;\leq\; F_{t,\varepsilon}(u_t) - F_t(u_t) \;\leq\; \varepsilon \log 2.
$$

In the other direction, C is convex and finite and the bracket in (iv) is differentiable with gradient $\pi_t^u - \pi_0^+(\cdot \mid s)$, so minimality of u_t provides a subgradient g of C at u_t with $g + \lambda\big(\pi_t - \pi_0^+(\cdot \mid s)\big) = 0$. For every score u, convexity gives $C(u) \geq C(u_t) + \sum_a g(a)\big(u(a) - u_t(a)\big)$, and the log-partition function satisfies the exact identity

$$
\log \sum_a \pi_0(a \mid s)^t e^{u(a)} - \log \sum_a \pi_0(a \mid s)^t e^{u_t(a)} - \sum_a \pi_t(a)\big(u(a) - u_t(a)\big) = \sum_a \pi_t(a) \log \frac{\pi_t(a)}{\pi_t^u(a \mid s)},
$$

as one sees by writing $\log \pi_t(a) - \log \pi_t^u(a \mid s) = u_t(a) - u(a) + \log \sum_{a'} \pi_0(a' \mid s)^t e^{u(a')} - \log \sum_{a'} \pi_0(a' \mid s)^t e^{u_t(a')}$ and averaging under π_t. Adding the two,

$$F_t(u) - F_t(u_t) \;\geq\; \sum_a \big(g(a) + \lambda \pi_t(a) - \lambda \pi_0^+(a \mid s)\big)\big(u(a) - u_t(a)\big) + \lambda \sum_a \pi_t(a) \log \frac{\pi_t(a)}{\pi_t^u(a \mid s)} = \lambda \sum_a \pi_t(a) \log \frac{\pi_t(a)}{\pi_t^u(a \mid s)},$$

and $u = u_{t,\varepsilon}$ gives (vi).

For (vii), fix $\varepsilon > 0$ and let t range over an open interval containing $[0, 1]$. The minimizer is characterized by the stationarity condition

$$\nabla C_\varepsilon(u_{t,\varepsilon}) + \lambda\big(\pi_{t,\varepsilon}(\cdot \mid s) - \pi_0^+(\cdot \mid s)\big) = 0, \tag{viii}$$

a smooth equation in (u, t) between zero-sum vectors whose derivative in u is $K + \lambda\big(\operatorname{diag}(\pi_{t,\varepsilon}(\cdot \mid s)) - \pi_{t,\varepsilon}(\cdot \mid s)\, \pi_{t,\varepsilon}(\cdot \mid s)^\top\big)$, with $K \coloneqq \nabla^2 C_\varepsilon(u_{t,\varepsilon})$. Differentiating (iii) twice and grouping the two orientations of each pair shows that K is the graph Laplacian

$$(Kv)(a) = \sum_{a' \neq a} k_{aa'}\big(v(a) - v(a')\big), \qquad k_{aa'} = \frac{1}{\varepsilon}\Big[\pi_0^+(a \mid s)\pi_0^-(a' \mid s)\, \sigma'\Big(\tfrac{\delta + u_{t,\varepsilon}(a') - u_{t,\varepsilon}(a)}{\varepsilon}\Big) + \pi_0^+(a' \mid s)\pi_0^-(a \mid s)\, \sigma'\Big(\tfrac{\delta + u_{t,\varepsilon}(a) - u_{t,\varepsilon}(a')}{\varepsilon}\Big)\Big]$$

so K is symmetric, $K\mathbf{1} = 0$ and $v^\top K v = \sum_{\{a,a'\}} k_{aa'}\big(v(a) - v(a')\big)^2 \geq 0$; the second summand of the derivative is the covariance matrix of $\pi_{t,\varepsilon}(\cdot \mid s)$, positive definite on nonzero zero-sum vectors. By the implicit function theorem, $t \mapsto u_{t,\varepsilon}$ and hence $t \mapsto \pi_{t,\varepsilon}$ are continuously differentiable; write $\dot u$ and $\dot\pi$ for their derivatives at a fixed t. Since $\pi_{t,\varepsilon}(\cdot \mid s)$ is the softmax of $u_{t,\varepsilon} + t \log \pi_0(\cdot \mid s)$, the chain rule gives $\dot\pi = \big(\operatorname{diag}(\pi_{t,\varepsilon}) - \pi_{t,\varepsilon}\pi_{t,\varepsilon}^\top\big)\big(\dot u + \log \pi_0(\cdot \mid s)\big)$, that is, $\dot\pi = \operatorname{diag}(\pi_{t,\varepsilon}(\cdot \mid s))\, y$ for the centred vector

$$y \coloneqq \dot u + \log \pi_0(\cdot \mid s) - \Big(\sum_a \pi_{t,\varepsilon}(a \mid s)\big(\dot u(a) + \log \pi_0(a \mid s)\big)\Big)\mathbf{1}.$$

Differentiating (viii) in t gives $K\dot u + \lambda \dot\pi = 0$, and $K\dot u = K\big(y - \log \pi_0(\cdot \mid s)\big)$ because $K\mathbf{1} = 0$; so

$$My = K \log \pi_0(\cdot \mid s), \qquad M \coloneqq K + \lambda \operatorname{diag}\big(\pi_{t,\varepsilon}(\cdot \mid s)\big).$$

M is symmetric positive definite, as $v^\top M v = v^\top K v + \lambda \sum_a \pi_{t,\varepsilon}(a \mid s)\, v(a)^2 > 0$ for $v \neq 0$; let $G \coloneqq M^{-1}$. Then $y = GK \log \pi_0(\cdot \mid s)$ and

$$\dot\pi = \operatorname{diag}(\pi_{t,\varepsilon})\, G\, \big(M - \lambda \operatorname{diag}(\pi_{t,\varepsilon})\big) \log \pi_0(\cdot \mid s) = \Omega\, \log \pi_0(\cdot \mid s), \qquad \Omega \coloneqq \operatorname{diag}(\pi_{t,\varepsilon}) - \lambda \operatorname{diag}(\pi_{t,\varepsilon})\, G \operatorname{diag}(\pi_{t,\varepsilon}).$$

Three properties of Ω give (vii). It is symmetric. Its rows sum to zero: $M\mathbf{1} = \lambda\, \pi_{t,\varepsilon}(\cdot \mid s)$ because $K\mathbf{1} = 0$, hence $G\, \pi_{t,\varepsilon}(\cdot \mid s) = \mathbf{1}/\lambda$ and $\Omega\mathbf{1} = \pi_{t,\varepsilon}(\cdot \mid s) - \operatorname{diag}(\pi_{t,\varepsilon})\mathbf{1} = 0$. And its off-diagonal entries $\Omega_{aa'} = -\lambda\, \pi_{t,\varepsilon}(a \mid s)\, \pi_{t,\varepsilon}(a' \mid s)\, G_{aa'}$ are nonpositive, because G is entrywise nonnegative: if x is a column of G, then Mx is a unit vector, and if x had a negative minimum at a coordinate a, then $(Mx)(a) = \lambda \pi_{t,\varepsilon}(a \mid s)\, x(a) + \sum_{a' \neq a} k_{aa'}\big(x(a) - x(a')\big) < 0$, which is impossible. For a symmetric matrix with zero row sums, $f^\top \Omega g = -\sum_{\{a,a'\}} \Omega_{aa'}\big(f(a) - f(a')\big)\big(g(a) - g(a')\big)$ for all vectors f, g, so

$$\frac{d}{dt} V_{\pi_{t,\varepsilon}}(s) = \sum_a Q(s,a)\, \dot\pi(a) = Q(s,\cdot)^\top\, \Omega\, \log \pi_0(\cdot \mid s)$$

is (vii) with $\omega_{aa'}(t) = \lambda\, \pi_{t,\varepsilon}(a \mid s)\, \pi_{t,\varepsilon}(a' \mid s)\, G_{aa'} \geq 0$. □

C MDP COUNTEREXAMPLES

Lemma 3.7. *There exists a finite finite-horizon tabular MDP (with a DAG structure) with binary terminal reward, a full-support base policy π_0, and an initial distribution μ, such that the following holds simultaneously for π_{rw} and π_{unw}. For every finite $\lambda \geq 0$, every $\delta \geq 0$, the statewise optimizer π_λ of the contrastive-plus-SFT objective (from eqs.* (3) *and* (4)*) satisfies $V_{\pi_0^+}(\mu) - V_{\pi_\lambda}(\mu) > 0$*

Proof. Fix $\varepsilon \in (0, 1/5)$. The MDP has decision states u, v, y and two absorbing states, success and failure; an entry $\mathrm{Bernoulli}(q)$ means immediate absorption in success with probability q and in failure otherwise.

State	Action 1	Action 2	$\pi_0(1 \mid s)$
u	$\mathrm{Bernoulli}(1-\varepsilon)$	move to y	$1/2$
v	$\mathrm{Bernoulli}(1-3\varepsilon)$	$\mathrm{Bernoulli}(\varepsilon/2)$	$1/2$
y	success	failure	ε

Every episode contains at most two decisions. The action values of π_0 and its state values are

$$Q(y,\cdot) = (1,0),\ V_{\pi_0}(y) = \varepsilon; \quad Q(u,\cdot) = (1-\varepsilon, \varepsilon),\ V_{\pi_0}(u) = \tfrac{1}{2};$$
$$Q(v,\cdot) = (1-3\varepsilon, \varepsilon/2),\ V_{\pi_0}(v) = \tfrac{2-5\varepsilon}{4},$$

where $Q(u,2) = V_{\pi_0}(y)$ because action 2 moves to y and π_0 is followed from there. Every decision state has base value in $(0,1)$ and π_0 has full support. The initial distribution is

$$\mu(u) = \frac{2-7\varepsilon}{2-5\varepsilon}, \quad \mu(v) = \frac{2\varepsilon}{2-5\varepsilon}, \quad \mu(y) = 0, \quad \text{so that} \quad \mu(u)\,\varepsilon = \mu(v)\Big(1 - \frac{7\varepsilon}{2}\Big).$$

Step 1: the optimizers at u and v. For $s \in \{u, v\}$ write $\alpha_s = \pi_0^+(1 \mid s)$, $\beta_s = \pi_0^-(1 \mid s)$, $p_s = \pi_\lambda(1 \mid s)$. From the values above,

$$\alpha_u = 1-\varepsilon, \quad \beta_u = \varepsilon, \quad \alpha_v = 1 - \frac{\varepsilon}{2-5\varepsilon}, \quad \beta_v = \frac{6\varepsilon}{2+5\varepsilon},$$

and, since $1 < 2-5\varepsilon < 2$ and $2+5\varepsilon < 3$, we have $\varepsilon/2 < 1-\alpha_v < \varepsilon$ and $\beta_v > 2\varepsilon$, hence

$$\alpha_v > \alpha_u, \quad (1-\alpha_v)\beta_v > \varepsilon^2 = (1-\alpha_u)\beta_u, \quad \alpha_v - \beta_v < 1 - \tfrac{5\varepsilon}{2} < 1 - 2\varepsilon = \alpha_u - \beta_u. \tag{9}$$

The base is uniform at u and v, so the reweighted and unweighted scores there differ by the constant $\ln 2$ and both losses have the same optimizer. As a function of the log-odds of action 1, the state loss is convex, and its derivative is $\lambda\big(\pi_\lambda(1 \mid s) - \alpha_s\big)$ plus the hinge contribution

$$-\alpha_s(1-\beta_s) \text{ on } (-\infty, -\delta), \quad \beta_s - \alpha_s \text{ on } (-\delta, \delta), \quad (1-\alpha_s)\beta_s \text{ on } (\delta, \infty).$$

For nonpositive log-odds, $\pi_\lambda(1 \mid s) \leq \frac{1}{2} < \alpha_s$ and $\beta_s < \alpha_s$ make this derivative negative, so the minimizer has positive log-odds. On $(0, \delta)$, at the kink δ, and on (δ, ∞) the stationarity conditions are $p_s - \alpha_s = (\alpha_s - \beta_s)/\lambda$; $p_s = \sigma(\delta)$ with $-(1-\alpha_s)\beta_s/\lambda \leq \sigma(\delta) - \alpha_s \leq (\alpha_s - \beta_s)/\lambda$; and $p_s - \alpha_s = -(1-\alpha_s)\beta_s/\lambda$. Exactly one of them holds, by convexity, and therefore

$$p_s - \alpha_s = \mathrm{median}\Big\{-\frac{(1-\alpha_s)\beta_s}{\lambda},\ \sigma(\delta) - \alpha_s,\ \frac{\alpha_s - \beta_s}{\lambda}\Big\}, \tag{10}$$

where for $\lambda = 0$ the outer entries are read as $\mp\infty$: the pure hinge is minimized exactly at log-odds δ, because $\alpha_s(1-\beta_s) > (1-\alpha_s)\beta_s$. By (9), each entry of (10) at u strictly exceeds the corresponding entry at v, and the median strictly increases under a strict componentwise increase; hence

$$(p_u - \alpha_u) - (p_v - \alpha_v) > 0 \quad \text{for every } \lambda, \delta \geq 0. \tag{11}$$

Step 2: values. At y, $\pi_0^+(1 \mid y) = \varepsilon \cdot 1/\varepsilon = 1$, so $V_{\pi_0^+}(y) = 1 \geq V_{\pi_\lambda}(y)$ for whatever policy π_λ uses at y. Some choice is needed there: for $\lambda > 0$ the loss at y has infimum 0, approached only as $\pi_\lambda(1 \mid y) \to 1$, and for $\lambda = 0$ its minimizers are not unique; the argument uses only $V_{\pi_\lambda}(y) \leq 1$. The one-step Bellman equations give

$$\begin{aligned} V_{\pi_0^+}(u) - V_{\pi_\lambda}(u) &= (\alpha_u - p_u)(1-\varepsilon) + (1-\alpha_u) - (1-p_u)V_{\pi_\lambda}(y) \\ &= \varepsilon\,(p_u - \alpha_u) + (1-p_u)\big(1 - V_{\pi_\lambda}(y)\big), \\ \implies V_{\pi_0^+}(v) - V_{\pi_\lambda}(v) &= (\alpha_v - p_v)\Big(1 - \frac{7\varepsilon}{2}\Big). \end{aligned}$$

Weighting by μ and using $\mu(u)\varepsilon = \mu(v)(1-7\varepsilon/2)$,

$$V_{\pi_0^+}(\mu) - V_{\pi_\lambda}(\mu) = \mu(u)\Big[\varepsilon\big((p_u-\alpha_u)-(p_v-\alpha_v)\big) + (1-p_u)\big(1-V_{\pi_\lambda}(y)\big)\Big] > 0$$

by (11) and $V_{\pi_\lambda}(y) \le 1$. This holds for both readouts, every finite $\lambda, \delta \ge 0$, and every admissible choice of the policy at y. □

Lemma 3.9. *There exists a finite finite-horizon MDP with two decision states, binary terminal rewards, and a full-support base policy π_0 (that is strictly aligned with Q_0 at both state) and some hyperparameter $\lambda, \delta > 0$ such that the state-wise hinge optimizers satisfy $V_{\pi_{\mathrm{unw}}}(s_0) > V_{\pi_{\mathrm{rw}}}(s_0)$.*

Proof. Fix scalars $0 < q < 1$, $\frac{1}{2} < c < 1$, $\frac{1}{2} < t < 1$, and r with $c + (1-c)q < r < \frac{c+(1-c)q^2}{c+(1-c)q}$. the interval is nonempty, the difference of its endpoints being $c(1-c)(1-q)^2/(c+(1-c)q) > 0$, and $r < 1$ because $q^2 < q$. The MDP has two decision states and two absorbing states, success and failure;

State	Action 1	Action 2	$\pi_0(1 \mid s)$
s_0	Bernoulli(r)	move to s_1	t
s_1	success	Bernoulli(q)	c

Thus, we see: the action values of the base policy π_0 and its state values are

$$Q(s_1,\cdot) = (1,\, q), \quad Q(s_0,\cdot) = \big(r,\; c+(1-c)q\big),$$
$$V_{\pi_0}(s_1) = c + (1-c)q, \quad V_{\pi_0}(s_0) = t\,r + (1-t)\big(c+(1-c)q\big),$$

where $Q(s_0, 2) = V_{\pi_0}(s_1)$ because action 2 moves to s_1 and π_0 is followed from there. At both states action 1 has the strictly larger action value, $Q(s_0,1) = r > c + (1-c)q = Q(s_0,2)$ and $Q(s_1,1) = 1 > q = Q(s_1,2)$, and, since $\pi_0(1 \mid s_0) = t > \frac{1}{2}$ and $\pi_0(1 \mid s_1) = c > \frac{1}{2}$, also the strictly larger base probability: π_0 is strictly aligned with $Q(s,\cdot)$ at both states.

The construction works as follows. Under the base critic, action 1 is the better action at s_0; but the SFT value of s_1, is given by:

$$V_{\pi_0^+}(s_1) = \frac{c+(1-c)q^2}{c+(1-c)q} > r,$$

This exceeds r, so once the policy at s_1 is SFT, action 2 at s_0 is the better action. We show that both π_{rw} and π_{unw} learn the same score gap at s_0, and that the reweighted readout, tilted by the aligned base, converts that gap into more mass on action 1.

We choose

$$0 < \delta < \log\frac{1}{q}, \quad 0 < \lambda < \min\Big\{\pi_0^+(1 \mid s_0) - \pi_0^-(1 \mid s_0),\; \pi_0^+(2 \mid s_0)\,\pi_0^-(1 \mid s_0)\Big\}.$$

The first entry is positive because $r > V_{\pi_0}(s_0)$ (action 1 beats the other action's quality), which gives $\pi_0^+(1 \mid s_0) = tr/V_{\pi_0}(s_0) > t > t(1-r)/(1-V_{\pi_0}(s_0)) = \pi_0^-(1 \mid s_0)$ (the second is positive because r and $c+(1-c)q$ lie in $(0,1)$.)

Since $Q_0(s_1,1) = 1$, the failure-conditioned policy is $\pi_0^-(\cdot \mid s_1) = (0,1)$, and the contrastive term reduces to

$$\pi_0^+(1 \mid s_1)\,[\delta + u_2 - u_1]_+ \;+\; \pi_0^+(2 \mid s_1)\,\delta,$$

which is minimal exactly when $u_1 - u_2 \ge \delta$. The SFT policy meets this for both $\mathcal{L}_{\mathrm{rw}}$ and $\mathcal{L}_{\mathrm{unw}}$: its un-weighted gap is $\log\frac{\pi_0^+(1|s_1)}{\pi_0^+(2|s_1)} = \log\frac{c}{(1-c)q} > \log\frac{1}{q} > \delta$ (as $c > 1-c$), and its re-weighted gap is $\log\frac{\pi_0^+(1|s_1)/c}{\pi_0^+(2|s_1)/(1-c)} = \log\frac{1}{q} > \delta$. Since the SFT term is uniquely minimized at π_0^+, both combined optimizers at s_1 equal $\pi_0^+(\cdot \mid s_1)$; they are unique, have full support, and

$$V_{\pi_{\mathrm{rw}}}(s_1) = V_{\pi_{\mathrm{unw}}}(s_1) = V_{\pi_0^+}(s_1).$$

The score gap is given by

$$g = \begin{cases} \ln\hat{\pi} - \ln\pi_0, \text{ re-weighted log prob gap} \\ \ln\hat{\pi}, \text{ un-weighted log prob} \end{cases}$$

In the score gap of either readout's own scores, the contrastive loss exression by putting π_0^+ in $\mathcal{L}_{\mathrm{unw}}$ and $\mathcal{L}_{\mathrm{rw}}$ can be obtained by:

$$\pi_0^+(1 \mid s_0)\pi_0^-(2 \mid s_0)\,[\delta - g]_+ + \pi_0^+(2 \mid s_0)\pi_0^-(1 \mid s_0)\,[\delta + g]_+ + \text{const},$$

whose left and right derivatives at $g = \delta$ are $-\big(\pi_0^+(1 \mid s_0) - \pi_0^-(1 \mid s_0)\big)$ and $\pi_0^+(2 \mid s_0)\pi_0^-(1 \mid s_0)$. The SFT term is strictly convex in g with derivative $\lambda\big(\pi(1 \mid s_0) - \pi_0^+(1 \mid s_0)\big)$, of absolute value less than λ. By the choice of λ, the total left derivative at δ is negative and the total right derivative positive, so the unique minimizer of the (strictly convex) objective is $g = \delta$, for both readouts. In probabilities,

$$\pi_{\mathrm{unw}}(1 \mid s_0) = \frac{e^{\delta}}{1 + e^{\delta}}, \quad \pi_{\mathrm{rw}}(1 \mid s_0) = \frac{t\,e^{\delta}}{1 - t + t\,e^{\delta}},$$

and $\pi_{\mathrm{rw}}(1 \mid s_0) > \pi_{\mathrm{unw}}(1 \mid s_0)$ because $t > \frac{1}{2}$. Since both policies coincide at s_1,

$$V_{\pi_{\mathrm{unw}}}(s_0) - V_{\pi_{\mathrm{rw}}}(s_0) = \big(\pi_{\mathrm{rw}}(1 \mid s_0) - \pi_{\mathrm{unw}}(1 \mid s_0)\big)\big(V_{\pi_0^+}(s_1) - r\big) > 0.$$

Both optimizers are unique and attained at finite scores, and the conclusion holds for every $\delta \in (0, \log \frac{1}{q})$ and every λ below the displayed bound. □

C.1 THE COMBINED LOSS IS A TRUST-REGION STEP

Trust-region policy optimization (Schulman et al., 2017) maximizes a linear surrogate of the return under a divergence constraint to the current policy; in its practical form the constraint is the quadratic approximation, which for a tabular policy at a state is the χ^2-divergence $\chi^2(p\|\pi_0) = \sum_a \big(p(a) - \pi_0(a \mid s)\big)^2/\pi_0(a \mid s)$. Russo (2026) observed that SFT is exactly such a step: at every state, π_0^+ maximizes the bandit reward over a χ^2-ball around π_0. The point of this section is that adding the contrastive term does not change the nature of the step, only its size. Once the SFT weight λ is large enough for the hinges to be in their linear regime, both combined losses, reweighted and unweighted alike, produce the same policy, and that policy is again the maximizer of the bandit reward over a χ^2-ball around π_0, now with a radius that shrinks in λ toward the SFT radius. Contrastive training, in this regime, is SFT with a wider trust region.

Theorem C.1 (The combined policy is a Pearson trust-region step)**.** *Consider a contextual bandit with finitely many actions, a full-support base policy π_0, and a non-degenerate state s at which $Q(s,\cdot)$ is not constant, There are thresholds $\lambda_*, \delta_* > 0$, depending only on π_0 and $Q(s,\cdot)$, with the following property. For every $\lambda \geq \lambda_*$ and $\delta \geq \delta_*$ the the optimal solutions of re-weighted and the un-weighted hinge losses solve the following TRPO:*

$$\max_{p \in \Delta_N} \sum_a p(a)\, Q(s,a) \quad \textit{subject to} \qquad \chi^2(p\|\pi_0) \leq R_\lambda^2,$$

The radius never increases with λ and R_λ^2 converges to the SFT radius $\chi^2(\pi_0^+\|\pi_0)$ as $\lambda \to \infty$.

Proof. Throughout, λ^* is the threshold of Lemma B.3, and we take

$$\lambda_* = \max\{2\lambda^*, 2\}, \qquad \delta_* = \max\left\{\log \frac{\max_{a\in P} Q(s,a)}{\min_{a\in P} Q(s,a)},\ \log \frac{\max_{a\in P} \pi_0(a \mid s)Q(s,a)}{\min_{a\in P} \pi_0(a \mid s)Q(s,a)}\right\} + \log 3.$$

These are finite: P is nonempty because $V_{\pi_0} > 0$, and quality and base probability are positive on P.

The first thing to check is that the lemma can be applied uniformly in λ, that is, that its margin condition can be met with a δ_* that does not depend on λ. The policy π^λ is just as defined in Lemma B.3. Fix $\lambda \geq \lambda_*$, so in particular $\lambda > \lambda^*$. Dividing the lemma's formula for π^λ by π_0^+, for $a \in P$,

$$\frac{\pi^\lambda(a \mid s)}{\pi_0^+(a \mid s)} = 1 - \frac{1}{\lambda}\left(\frac{\pi_0^-(a \mid s)}{\pi_0^+(a \mid s)} - \pi_0^-(P \mid s)\right). \tag{1}$$

The bracket is at least $-\pi_0^-(P \mid s) \geq -1$ and at most λ^* (that is the definition of λ^*), so for $\lambda \geq \max\{2\lambda^*, 2\}$ the ratio in (1) lies between $\frac{1}{2}$ and $\frac{3}{2}$. Taking logarithms, the scores of π^λ on P, in either readout, differ from the corresponding SFT scores by a number between $\log\frac{1}{2}$ and $\log\frac{3}{2}$, and an oscillation can grow by at most $\log 3$ under such a perturbation. The SFT scores are explicit, since $\pi_0^+(a \mid s) = \pi_0(a \mid s)Q(s,a)/V_{\pi_0}$ on P: their oscillation over P is $\log\big(\max_P \pi_0 Q/\min_P \pi_0 Q\big)$ for the unweighted readout and $\log\big(\max_P Q/\min_P Q\big)$ for the reweighted one. So $D_{\mathrm{unw}}(\lambda) \leq \delta_*$ and $D_{\mathrm{rw}}(\lambda) \leq \delta_*$, and the lemma says that for all $\lambda \geq \lambda_*$ and $\delta \geq \delta_*$ both losses are optimized by π^λ — uniquely if P is the whole action set, and in the limiting sense of the statement otherwise.

The rest is a computation with the explicit policy. Summing the definition of π_0^- over P gives $\pi_0^-(P \mid s) = \big(\pi_0(P \mid s) - V_{\pi_0}\big)/(1 - V_{\pi_0})$, and then, for $a \in P$,

$$\begin{aligned}\frac{\pi_0^-(P \mid s)\pi_0^+(a \mid s) - \pi_0^-(a \mid s)}{\pi_0(a \mid s)} &= \frac{\big(\pi_0(P \mid s) - V_{\pi_0}\big)Q(s,a)/V_{\pi_0} - \big(1 - Q(s,a)\big)}{1 - V_{\pi_0}} \\ &= \frac{\pi_0(P \mid s)Q(s,a)/V_{\pi_0} - 1}{1 - V_{\pi_0}},\end{aligned}$$

so the lemma's formula reads

$$\frac{\pi^\lambda(a \mid s)}{\pi_0(a \mid s)} = \frac{1 + \frac{\pi_0(P|s)}{\lambda(1-V_{\pi_0})}}{V_{\pi_0}}\, Q(s,a) - \frac{1}{\lambda(1 - V_{\pi_0})} \qquad (a \in P). \tag{2}$$

It is convenient to center $Q(s,\cdot)$ at $V_{\pi_0}/\pi_0(P \mid s)$, its mean under π_0 restricted to P. Since $\frac{1+\frac{\pi_0(P|s)}{\lambda(1-V_{\pi_0})}}{V_{\pi_0}} \cdot \frac{V_{\pi_0}}{\pi_0(P|s)} = \frac{1}{\pi_0(P|s)} + \frac{1}{\lambda(1-V_{\pi_0})}$, subtracting 1 from (2) gives

$$\frac{\pi^\lambda(a \mid s)}{\pi_0(a \mid s)} - 1 = \frac{1+\frac{\pi_0(P|s)}{\lambda(1-V_{\pi_0})}}{V_{\pi_0}}\Big(Q(s,a) - \frac{V_{\pi_0}}{\pi_0(P \mid s)}\Big) + \frac{1-\pi_0(P \mid s)}{\pi_0(P \mid s)} \qquad (a \in P), \quad (3)$$

while off P, where π^λ vanishes, the left side of (3) is -1.

The radius now falls out. Square (3), weight by $\pi_0(a \mid s)$ and sum over P: the cross term is killed by $\sum_{a\in P}\pi_0(a \mid s)\big(Q(s,a) - V_{\pi_0}/\pi_0(P \mid s)\big) = 0$, the constant term contributes $(1-\pi_0(P \mid s))^2/\pi_0(P \mid s)$, and the actions off P contribute their base mass $1-\pi_0(P \mid s)$; the last two add up to $(1-\pi_0(P \mid s))/\pi_0(P \mid s)$. Thus

$$\chi^2(\pi^\lambda \| \pi_0) = R_\lambda^2. \quad (4)$$

For optimality no duality is needed: the divergence is a quadratic, so we expand it around π^λ. For $p \in \Delta_N$, by (4),

$$\chi^2(p\|\pi_0) - R_\lambda^2 = \sum_a \frac{\big(p(a) - \pi^\lambda(a \mid s)\big)^2}{\pi_0(a \mid s)} + 2\sum_a \big(p(a) - \pi^\lambda(a \mid s)\big)\Big(\frac{\pi^\lambda(a \mid s)}{\pi_0(a \mid s)} - 1\Big).$$

In the second sum substitute (3) on P and -1 off P. Both p and π^λ have total mass one, so $\sum_{a\in P}\big(p(a) - \pi^\lambda(a \mid s)\big) = -\sum_{a\notin P} p(a)$, and $Q(s,a) = \pi^\lambda(a \mid s) = 0$ off P. The pieces not involving Q therefore collect, with the common factor $-\sum_{a\notin P} p(a)$, to $-2\big(\frac{1}{\pi_0(P|s)} + \frac{1}{\lambda(1-V_{\pi_0})}\big) + \frac{2(1-\pi_0(P|s))}{\pi_0(P|s)} = -2 - \frac{2}{\lambda(1-V_{\pi_0})}$; together with the $-2\sum_{a\notin P} p(a)$ coming from off P this leaves

$$\begin{aligned} \chi^2(p\|\pi_0) - R_\lambda^2 &= \sum_a \frac{\big(p(a) - \pi^\lambda(a \mid s)\big)^2}{\pi_0(a \mid s)} \\ +2\,\frac{1+\frac{\pi_0(P|s)}{\lambda(1-V_{\pi_0})}}{V_{\pi_0}} &\sum_a \big(p(a) - \pi^\lambda(a \mid s)\big)Q(s,a) + \frac{2}{\lambda(1-V_{\pi_0})}\sum_{a\notin P} p(a). \end{aligned} \quad (5)$$

This is the whole argument. If p is feasible the left side is nonpositive; the first and last terms on the right are nonnegative and the coefficient of the middle one is positive, so $\sum_a p(a)Q(s,a) \le \sum_a \pi^\lambda(a \mid s)Q(s,a)$, and strictly so if $p \neq \pi^\lambda$, because full support makes the first term positive. As π^λ is feasible by (4), it is the unique maximizer. The last term in (5) is where the constraint $p \in \Delta_N$ earns its keep: it is the contribution of the nonnegativity constraints active off P, and without them the maximizer would carry negative mass there — which is also why the radius contains the extra $(1-\pi_0(P \mid s))/\pi_0(P \mid s)$.

It remains to look at the dependence on λ. In R_λ^2 only the squared coefficient $\big(1+\frac{\pi_0(P|s)}{\lambda(1-V_{\pi_0})}\big)^2/V_{\pi_0}^2$ depends on λ, and it decreases strictly to $1/V_{\pi_0}^2$. So R_λ is non-increasing, and it is strictly decreasing and strictly above its limit at every finite λ exactly when the sum in R_λ^2 is positive, i.e. when $Q(s,\cdot)$ is not constant on P (the sum vanishes iff $Q(s,a) = V_{\pi_0}/\pi_0(P \mid s)$ for every $a \in P$). Lastly, by (2), $\pi^\lambda(a \mid s) \to \pi_0(a \mid s)Q(s,a)/V_{\pi_0} = \pi_0^+(a \mid s)$ as $\lambda \to \infty$ for every action, and (4) together with continuity of $\chi^2(\cdot\|\pi_0)$ gives $R_\lambda^2 \to \chi^2(\pi_0^+\|\pi_0)$. □

Remark C.2. Let $P = \{a : Q(s,a) > 0\}$. Our proof actually shows tthe following closed form of the radius

$$R_\lambda^2 = \frac{1-\pi_0(P \mid s)}{\pi_0(P \mid s)} + \left(\frac{1+\frac{\pi_0(P|s)}{\lambda(1-V_{\pi_0})}}{V_{\pi_0}}\right)^2 \sum_{a\in P}\pi_0(a \mid s)\Big(Q(s,a) - \frac{V_{\pi_0}}{\pi_0(P \mid s)}\Big)^2.$$

C.2 Retrieval Guarantees for Contrastive Learning

Theorem 3.10. *Consider a contextual bandit with state s and actions $a_1, \cdots, a_n$ such that $Q(s, a_1) \geq \cdots \geq Q(s, a_n)$. Then the following holds:*

1. *Fix any $\lambda > 0$ and let π_{unw} and π_{rw} be the un-weighted and re-weighted contrastive policies respectively. Then, $\mathrm{Inv}(\pi_{\mathrm{unw}}) \subseteq \mathrm{Inv}(\pi_0^+) \subseteq \mathrm{Inv}(\pi_0)$ and $\mathrm{Inv}(\pi_{\mathrm{rw}}) \subseteq \mathrm{Inv}(\pi_0)$.*
2. *$\exists \lambda^* > 0, \delta^* > 0$ such that $\forall \lambda \geq \lambda^*$ and $\delta \geq \delta^*$, we have $\mathrm{Inv}(\pi_{\mathrm{rw}}) \subseteq \mathrm{Inv}(\pi_0^+)$*

Proof. Throughout, π_0 has full support and $V_{\pi_0} \in (0, 1)$, scores are log-probabilities for the unweighted loss and log-ratios to π_0 for the reweighted one, and a pair (i, j) with $Q(s, a_i) = Q(s, a_j)$ is never an inversion. When some action has zero quality the losses are not minimized at finite scores, and the contrastive policy is the limit of the policies induced by a minimizing score sequence, as in Theorem 3.1. The argument given there shows that every action omitted by the limit has zero quality and zero probability, and that the limit policy minimizes the loss restricted to its support. An omitted action has the smallest possible quality and the smallest possible probability, so it belongs to no inversion of the limit policy; it is therefore enough to prove the containments for a policy that minimizes its loss at finite scores, restricted if necessary to its support, and we assume this from now on.

We first record the fact used in the text: if $\mathrm{Inv}(\pi_1) \subseteq \mathrm{Inv}(\pi_2)$, then the list sorted by π_1 scores at least as well as the list sorted by π_2 under every swap-monotone metric. Sort by a policy breaking ties between equal probabilities in favor of the smaller index, so that the pairs a list orders against quality are exactly the inversions of its policy. Now take the first action of the π_1-list and bring it to the front of the π_2-list by adjacent swaps, then the second action of the π_1-list to the front of what remains, and so on until the π_2-list has become the π_1-list. Every swap moves an action being placed above an action that preceded it in the π_2-list but follows it in the π_1-list. If the two have equal quality, the sequence of qualities read along the list is unchanged, and so is every swap-monotone metric. If the action being placed has the larger quality, the swap moves a more relevant item above a less relevant one and no swap-monotone metric decreases. The third case does not occur: if the action being placed had the smaller quality, then, ties being broken by index, it would have strictly larger π_1-probability than the other action, the pair would be an inversion of π_1 and hence of π_2, and the action being placed would already precede the other in the π_2-list.

For item 1, the outer containment is one line. If $(i, j) \in \mathrm{Inv}(\pi_0^+)$ then $Q(s, a_i) > Q(s, a_j)$ and $\pi_0(a_i \mid s)Q(s, a_i) < \pi_0(a_j \mid s)Q(s, a_j)$, whence $\pi_0(a_i \mid s) < \pi_0(a_j \mid s)Q(s, a_j)/Q(s, a_i) < \pi_0(a_j \mid s)$, that is $(i, j) \in \mathrm{Inv}(\pi_0)$.

Next, $\mathrm{Inv}(\pi_{\mathrm{unw}}) \subseteq \mathrm{Inv}(\pi_0^+)$. Suppose not: there are i, j with $Q(s, a_i) > Q(s, a_j)$ and $\pi_0^+(a_i \mid s) \geq \pi_0^+(a_j \mid s)$ but $\pi_{\mathrm{unw}}(a_i \mid s) < \pi_{\mathrm{unw}}(a_j \mid s)$, so the score of a_i is smaller than that of a_j. Increase the score of a_i at rate $1/\pi_0^+(a_i \mid s)$ and decrease that of a_j at rate $1/\pi_0^+(a_j \mid s)$, keeping the other scores fixed; we show the right directional derivative of the unweighted loss is negative, contradicting minimality. Recall that a hinge has slope 1 at positive arguments and 0 at negative ones, so its right slope at a smaller argument never exceeds its left slope at a larger one, kinks included.

Consider the pairs in which some third action a is the success action. The hinge with failure action a_i has the smaller argument and it increases; the one with failure action a_j has the larger argument and it decreases; after the rates are absorbed, their weights are $\pi_0^+(a \mid s)\pi_0^-(a_i \mid s)/\pi_0^+(a_i \mid s)$ and $\pi_0^+(a \mid s)\pi_0^-(a_j \mid s)/\pi_0^+(a_j \mid s)$, and

$$\frac{\pi_0^-(a_i \mid s)}{\pi_0^+(a_i \mid s)} = \frac{V_{\pi_0}}{1 - V_{\pi_0}} \cdot \frac{1 - Q(s, a_i)}{Q(s, a_i)} < \frac{V_{\pi_0}}{1 - V_{\pi_0}} \cdot \frac{1 - Q(s, a_j)}{Q(s, a_j)} = \frac{\pi_0^-(a_j \mid s)}{\pi_0^+(a_j \mid s)},$$

so the smaller weight multiplies the smaller slope and the two contributions add to something nonpositive. In the pairs where the third action is the failure action, the hinge with success action a_i has the larger argument and it decreases, the one with success action a_j has the smaller argument and it increases, and after absorbing the rates both carry the same weight $\pi_0^-(a \mid s)$; again the sum is nonpositive. Pairs not involving a_i or a_j, and the two self-pairs, are constant. Of the two pairs formed by a_i and a_j, the hinge with success action a_i has argument δ plus the score gap, hence is

strictly active with slope 1 and its argument falls at rate $1/\pi_0^+(a_i \mid s) + 1/\pi_0^+(a_j \mid s)$; the reverse hinge rises at the same rate with slope at most 1. Together they contribute at most

$$\Big(\frac{1}{\pi_0^+(a_i \mid s)} + \frac{1}{\pi_0^+(a_j \mid s)}\Big)\big(\pi_0^+(a_j \mid s)\pi_0^-(a_i \mid s) - \pi_0^+(a_i \mid s)\pi_0^-(a_j \mid s)\big)$$
$$= \big(\pi_0^+(a_i \mid s) + \pi_0^+(a_j \mid s)\big)\Big(\frac{\pi_0^-(a_i \mid s)}{\pi_0^+(a_i \mid s)} - \frac{\pi_0^-(a_j \mid s)}{\pi_0^+(a_j \mid s)}\Big) < 0.$$

Finally the SFT term: differentiating $-\lambda \sum_a \pi_0^+(a \mid s) \log \pi_{\mathrm{unw}}(a \mid s)$ along the perturbation, the terms linear in the scores cancel and the log-partition contributes

$$\lambda\Big(\frac{\pi_{\mathrm{unw}}(a_i \mid s)}{\pi_0^+(a_i \mid s)} - \frac{\pi_{\mathrm{unw}}(a_j \mid s)}{\pi_0^+(a_j \mid s)}\Big) \le \lambda\Big(\frac{\pi_{\mathrm{unw}}(a_i \mid s)}{\pi_0^+(a_j \mid s)} - \frac{\pi_{\mathrm{unw}}(a_j \mid s)}{\pi_0^+(a_j \mid s)}\Big) < 0,$$

using $\pi_0^+(a_i \mid s) \ge \pi_0^+(a_j \mid s)$ and then $\pi_{\mathrm{unw}}(a_i \mid s) < \pi_{\mathrm{unw}}(a_j \mid s)$. The right directional derivative is therefore strictly negative, which is impossible at a minimizer.

For the reweighted policy the same perturbation, with rates $1/\pi_0(a_i \mid s)$ and $1/\pi_0(a_j \mid s)$, is exactly the computation in the proof of Theorem 3.1, which gives $\pi_{\mathrm{rw}}(a_i \mid s)/\pi_0(a_i \mid s) \ge \pi_{\mathrm{rw}}(a_j \mid s)/\pi_0(a_j \mid s)$ whenever $Q(s, a_i) > Q(s, a_j)$, for every $\lambda, \delta \ge 0$. If $(i, j) \notin \mathrm{Inv}(\pi_0)$, that is $\pi_0(a_i \mid s) \ge \pi_0(a_j \mid s)$, multiplying the two inequalities gives $\pi_{\mathrm{rw}}(a_i \mid s) \ge \pi_{\mathrm{rw}}(a_j \mid s)$, so $(i, j) \notin \mathrm{Inv}(\pi_{\mathrm{rw}})$. This proves item 1.

For item 2 take the thresholds λ_*, δ_* of Theorem C.1. For $\lambda \ge \lambda_*$ and $\delta \ge \delta_*$ both contrastive policies equal the policy π^λ of Lemma B.3, which vanishes on zero-quality actions and on the others equals

$$\pi^\lambda(a \mid s) = \pi_0^+(a \mid s)\Big(1 + \frac{1}{\lambda}\Big(\pi_0^-(P \mid s) - \frac{\pi_0^-(a \mid s)}{\pi_0^+(a \mid s)}\Big)\Big),$$

with a positive bracket because $\lambda > \lambda^*$. Let $Q(s, a_i) > Q(s, a_j)$ and $\pi_0^+(a_i \mid s) \ge \pi_0^+(a_j \mid s)$. If a_j has zero quality then $\pi^\lambda(a_j \mid s) = 0 \le \pi^\lambda(a_i \mid s)$. Otherwise both actions have positive quality, and since $\pi_0^-(a \mid s)/\pi_0^+(a \mid s)$ decreases with $Q(s, a)$, the bracket is strictly larger at a_i than at a_j; multiplied by $\pi_0^+(a_i \mid s) \ge \pi_0^+(a_j \mid s)$ this gives $\pi^\lambda(a_i \mid s) > \pi^\lambda(a_j \mid s)$. Either way $(i, j) \notin \mathrm{Inv}(\pi^\lambda)$, so $\mathrm{Inv}(\pi_{\mathrm{rw}}) = \mathrm{Inv}(\pi^\lambda) \subseteq \mathrm{Inv}(\pi_0^+)$, and the same holds for π_{unw}. □

D Additional Experiments and Details

Here we first validate our Theorems 3.2,3.6 via the following synthetic setup.

D.1 Empirical analysis in semi-synthetic bandits

We use a semi-synthetic contextual bandit to separate the effects of reference–reward alignment, finite feedback, and SFT regularization. The contexts are representations of MS MARCO queries; the actions, reward mechanism, and reference policies are constructed.

Experimental setup We want to design a contextual bandit setup where each context x is given by text passages and each context has a space of $|\mathcal{A}| = 32$ actions. We obtain query texts from MS MARCO dataset Bajaj et al. (2018) where each query is represented using a its feature extracted by a frozen T5-base encoder Raffel et al. (2023). In this synthetic setup, the policy selects an action rather than generating unrestricted text. Both the base π_0 and the leaarner π_θ are softmax policies of the form $\pi_\theta(\cdot \mid q) = \mathrm{softmax}(f_\theta(x_q))$, where f_θ in an MLP (π_0 is constructed at frozen weights of f_θ).

We specifically control the *alignment* between π_0 and Q, by assigning rewards to its action ranks using permutations with prescribed inversion counts. This gives exact per-query Kendall correlations $\tau \in \{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$ between the base and the reward. Positive and negative actions are sampled independently π_0 and we train f_θ using the contrastive losses $\mathcal{L}_{\mathrm{unw}}$, $\mathcal{L}_{\mathrm{rw}}$ on its induced policy π_θ.

Convex combination loss: Let $\mathcal{L}_{\text{ctr}}, \mathcal{L}_{\text{SFT}}$ be the contrastive (un-weighted or re-weighted) and SFT terms respectively. We optimize the convex combination $(1-\omega)\mathcal{L}_{\text{ctr}} + \omega\mathcal{L}_{\text{SFT}}$. It should be noted that the optimizer π_ω of this loss is equivalent to the optimizers of $\mathcal{L}_{\text{rw}}, \mathcal{L}_{\text{unw}}$ at $\lambda = \omega/(1-\omega)$. Taking the convex combination was an experimental choice to avoid numerical instabilities. Since $\lambda \to \infty$ as $\omega \to 1$, it implies that the pure SFT trained π_0^+ corresponds to the optimizer for the above loss when $\omega = 1$.

We evaluate ten $\omega \in \{0, 1/9, \ldots, 1\}$ and all five margins in log scale $\delta \in [0.1, 10]$ for both $\pi_{\text{rw}}, \pi_{\text{unw}}$ sharing one trained SFT endpoint at $\lambda = 1$, across all 5 reward functions with the correlations $\{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$. We show $V_{\pi_{\text{rw}}}, V_{\pi_{\text{unw}}}$ and $V_{\pi_{\text{rw}}} - V_{\pi_{\text{unw}}}$ across the grid in fig. 3

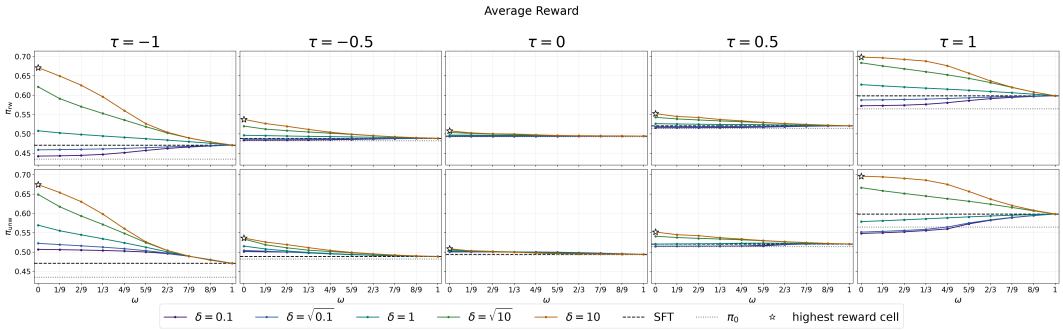


Figure 3: Caption

We make the following observations: **(1)** As $\tau : -1 \to +1$, we see that π_{rw} clearly improves, which is consistent with our intuition that as the alignment between π_0 and Q improves, the "effective margin" in the re-weighted loss increases, which makes the positive and negative actions even more well-separated, and the learnt policy ends up choosing the positives more. We see a consequence of that here. **(2)** the learnt SFT improves as the alignment τ improves **(3)** π_{unw} is mostly better than SFT at low τ, but as τ increases, SFT becomes better than π_{rw} in most of the cells. All of our observations are consistent with the intuition that if π_0 already ranks most of the actions correctly, the relative improvement of π_{unw} becomes less significant.

D.2 Generative Retrieval Experimental Details

Data and reference policies We use the MS MARCO passage collection (Bajaj et al., 2018) of 8,841,823 passages. The training source contains 808,731 queries and 532,751 qrel rows. Feedback construction uses a fixed subset of 10,000 training queries with at least one positive qrel. A fixed 500-query selection holdout is chosen from this subset, whch we used for validation. Final reporting uses all 6,980 dev-small queries and their 7,437 qrel rows.

The generator and tokenizer use `facebook/bart-large` (Lewis et al., 2020). A fixed compact trie and terminal-state mapping cover the complete passage collection. Each terminal pseudo-query identifier maps to one or more passage IDs and contains at most 31 content tokens followed by exactly one EOS token. All references come from one vanilla MINDER (Li et al., 2023b) training trajectory. Its upstream supervised query-to-pseudo-query recipe uses batch size 128, learning rate 5×10^{-5}, 1,000 warmup updates, and checkpoints every 1,000 updates.

We evaluate three reference blocks: *Early* at update 1,000, *Midpoint* at update 3,000, and *Final* at the final vanilla checkpoint. Midpoint is the saved checkpoint whose native beam-slot reward is closest to the arithmetic mean of Early and Final, among checkpoints strictly between these anchors; ties favor the earlier update. Each block uses its own frozen reference, candidate support, feedback pairs, and independent model selection.

Candidate support and reward construction For each reference and feedback query, we run trie-constrained beam search with width 100 and maximum identifier length 32, including EOS. The output contains exactly 100 terminal slots. The constrained beam score sums token log-probabilities, including EOS.

For every retained identifier, we store its full sequence log-probability

$$\ell_0(q, z) = \sum_{t=1}^{T(z)} \log \pi_0(z_t \mid q, z_{<t}). \tag{12}$$

EOS contributes to this sum; padding does not. These scores are neither length-normalized nor interchangeable with the constrained beam scores.

For training support, duplicate terminal states are collapsed, retaining the occurrence with the highest constrained score.For reward feedback, we have used `Qwen/Qwen3-Reranker-0.6B` (Zhang et al., 2025). For passage d, the soft reward is

$$\widehat{Q}(q, d) = \sigma(a_{\text{yes}}(q, d) - a_{\text{no}}(q, d)), \tag{13}$$

where a_{yes} and a_{no} are the corresponding output logits of the Qwen-encoder. If $M(z)$ is the set of unique passages associated with identifier z, its reward is the average:

$$\bar{Q}(q, z) = \frac{1}{|M(z)|} \sum_{d \in M(z)} \widehat{Q}(q, d). \tag{14}$$

Relevance label scoring Let $\mathcal{C}_q$ denote its finite set of unique candidate identifiers. Define

$$b_{\mathcal{C}}(z \mid q) = \frac{\exp \ell_0(q, z)}{\sum_{z' \in \mathcal{C}_q} \exp \ell_0(q, z')}, \quad V_{\mathcal{C}}(q) = \sum_{z \in \mathcal{C}_q} b_{\mathcal{C}}(z \mid q) \bar{Q}(q, z), \tag{15}$$

$$p_+(z \mid q) = \frac{b_{\mathcal{C}}(z \mid q) \bar{Q}(q, z)}{V_{\mathcal{C}}(q)}, \quad p_-(z \mid q) = \frac{b_{\mathcal{C}}(z \mid q)(1 - \bar{Q}(q, z))}{1 - V_{\mathcal{C}}(q)}. \tag{16}$$

We draw eight positives and eight negatives independently, with replacement, and pair them by draw index. Every positive draw is also an SFT target.

The trainable score $\ell_\theta(q, z)$ is the full-vocabulary sequence sum defined as in (12). The readouts are

$$s_{\text{UNW}}(q, z) = \ell_\theta(q, z), \qquad s_{\text{RW}}(q, z) = \ell_\theta(q, z) - \ell_0(q, z). \tag{17}$$

For a batch of B routed pairs and margin δ, we minimize

$$\mathcal{L}_{\text{ctr}} = \frac{1}{B} \sum_{i=1}^{B} [\delta - s(q_i, z_i^+) + s(q_i, z_i^-)]_+, \tag{18}$$

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{B} \sum_{i=1}^{B} \ell_\theta(q_i, z_i^+), \tag{19}$$

$$\mathcal{L}_\omega = (1 - \omega)\mathcal{L}_{\text{ctr}} + \omega \mathcal{L}_{\text{SFT}}. \tag{20}$$

Neither term is normalized by identifier length. We use ω for the experimental mixture weight to distinguish it from the additive SFT coefficient in the theory. For $\omega < 1$, the objectives have the same minimizers under $\lambda = \omega/(1 - \omega)$. This rescaling does not imply identical finite-training trajectories with a fixed optimizer configuration.

For each readout and reference, the grid is

$$\omega \in \{0, 1/9, \ldots, 8/9\}, \quad \delta \in \{0.1, 2.575, 5.05, 7.525, 10\}. \tag{21}$$

Thus each block has 45 coordinates per readout, including the pure hinge endpoint at $\omega = 0$. One shared SFT model represents $\omega = 1$.

Comparison methods We include the frozen reference, SFT, DPO (Rafailov et al., 2023), RankNet (Burges et al., 2005), and SimPO (Meng et al., 2024) in each block. Writing $\ell^\pm = \ell(q, z^\pm)$ and $\text{sp}(x) = \log(1 + e^x)$, the trainable comparison losses are

$$\mathcal{L}_{\text{DPO}} = \text{mean sp}\left(-0.1[(\ell_\theta^+ - \ell_\theta^-) - (\ell_0^+ - \ell_0^-)]\right), \tag{22}$$

$$\mathcal{L}_{\text{RankNet}} = \text{mean sp}[-(\ell_\theta^+ - \ell_\theta^-)], \tag{23}$$

$$\mathcal{L}_{\text{SimPO}} = \text{mean sp}\left(-\left[2\left(\frac{\ell_\theta^+}{T^+} - \frac{\ell_\theta^-}{T^-}\right) - 1\right]\right). \tag{24}$$

These three baselines use threshold-hard pairs. Positive identifiers satisfy $\bar{Q}(q, z^+) \geq 0.5$ and negative identifiers satisfy $\bar{Q}(q, z^-) < 0.5$. Eligible Cartesian pairs are ordered by decreasing reference-policy violation, then increasing reward gap, with a stable hash as the final tie-break; the first eight are retained. Their comparison therefore changes both the loss and the pair-construction rule. It does not isolate the effect of the loss alone.

Optimization and model selection Each trainable model starts from its block's reference. Optimization uses AdamW, batch size 16, one accumulation step, weight decay 0.01, gradient clipping at 1.0, and a linear schedule with 250 warmup updates. Maximum query and identifier lengths are 64 and 32. Runs use bf16, one GPU, seed 0, and deterministic PyTorch algorithms.

Native decoding and passage ranking Evaluation uses the same query prefix and constrained beam width 100. Generation reward retains all 100 raw terminal slots. Passage ranking first deduplicates decoded identifier text, retaining the occurrence with the largest constrained sequence score. For retained identifier z, let $c(z, d)$ be its mapping multiplicity for passage d. Native passage mass is

$$A(q, d) = \sum_z c(z, d) \exp \ell_{\text{beam}}(q, z). \tag{25}$$

Passages are ordered by decreasing mass; ties follow first appearance in the resolved identifier stream. The first 100 unique passage IDs form the native output. Beam search approximates sequence search, and the final passage ranking also aggregates across identifiers.

Reported metrics All metrics are macro-averaged over reporting queries. Let R_q be the set of positively judged passages, d_i the passage at native rank i, and $I_i = \mathbf{1}\{d_i \in R_q\}$. Missing positions contribute zero. We report the following at $k \in \{10, 100\}$:

$$\text{Precision } @k(q) = \frac{\sum_{i=1}^k I_i}{k}, \ \text{Recall } @k(q) = \frac{\sum_{i=1}^k I_i}{|R_q|}, \tag{26}$$

$$\text{AP } @k(q) = \frac{1}{|R_q|} \sum_{i=1}^k I_i \, \text{Precision } @i(q). \tag{27}$$

MAP is the mean AP; its denominator remains $|R_q|$, even when $k < |R_q|$. Reciprocal rank is $1/j$ when the first relevant passage occurs at rank $j \leq k$, and zero otherwise; MRR is its query mean. NDCG uses linear qrel gain,

$$\text{DCG } @k(q) = \sum_{i=1}^k \frac{g_q(d_i)}{\log_2(i+1)}, \quad \text{NDCG } @k(q) = \frac{\text{DCG } @k(q)}{\text{IDCG } @k(q)}, \tag{28}$$

where unjudged or nonrelevant passages have grade zero, and IDCG sorts the positive grades. This uses g, rather than $2^g - 1$, as gain.

For raw terminal slots $z_{q,1}, \ldots, z_{q,100}$, the primary reward criterion is

$$G(q) = \frac{1}{100} \sum_{j=1}^{100} \bar{Q}(q, z_{q,j}). \tag{29}$$

Each raw slot has equal weight; repeated slots retain multiplicity. Within each slot, passage IDs are deduplicated before applying (14). The secondary passage reward averages $\widehat{Q}(q, d)$ over the unique native top-100 passage set P_q, using denominator $|P_q|$. Neither summary is a likelihood-weighted policy expectation or an i.i.d. best-of-k value.

Additional ranking numbers Our goal is to have a detailed understanding of how good the trained policies $\pi_{\text{rw}}, \pi_{\text{unw}}$ are across different ranking metrics and across the (λ, δ) configurations of contrastive losses and how they vary across the checkpoints. Basically we want to check on 3 axes: alignment, contrastive loss and hyperparameters. We plot the variation of ranking performance for contrastive losses across their hyperparam grids for each of the checkpoints in the following figures.

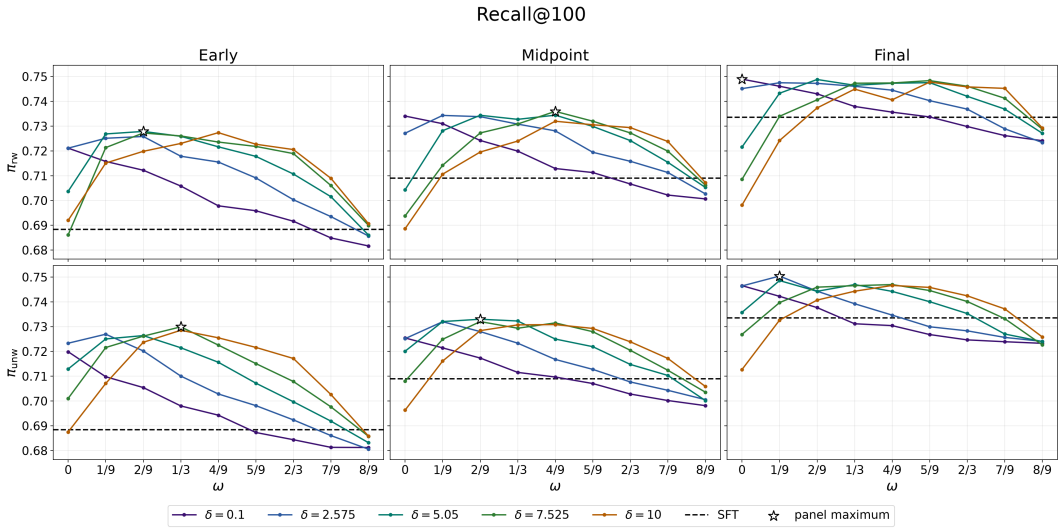


Figure 4: Variation of Recall with ω across different initial π_0. Each curve is one margin δ.

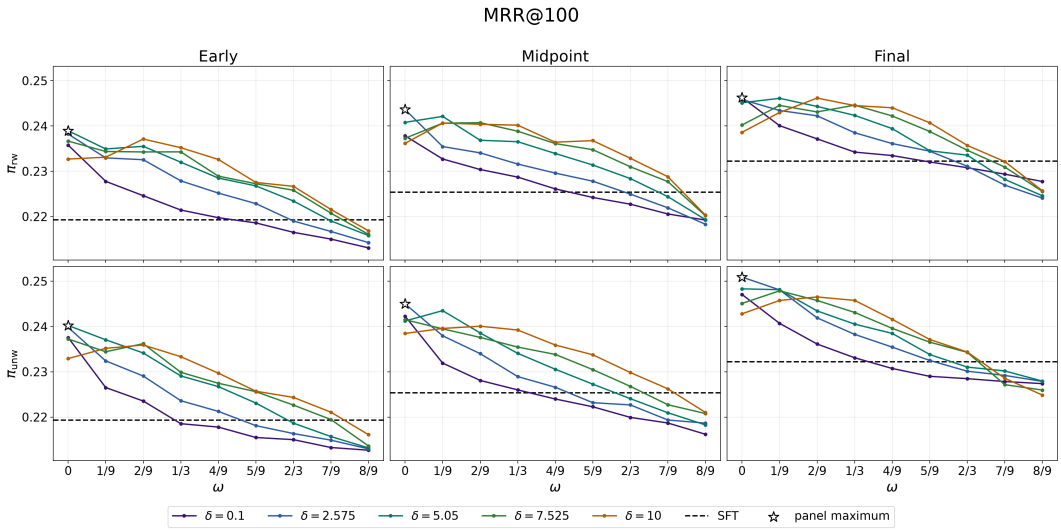


Figure 5: Variation of MRR with ω across different initial π_0. Each curve is one margin δ.

Across all ranking metrics in figs. 4 to 8 we observe that **(1)** Low margin curves degrage below π_0^+ fast, highlighting the importance of a large margin as prescribed by our theorems. **(2)** Both the contrastive losses are better than pure SFT for a significant number of hyperparameters.

For several of the ranking metrics we see that π_{unw} is the winner, but π_{unw} trails close, despite π_{rw} having higher reward for midpoint and final checkpoints. This is consistent with our theoretical intuition of π_{rw} being optimal for reward, but π_{unw} being designed more for ranking.

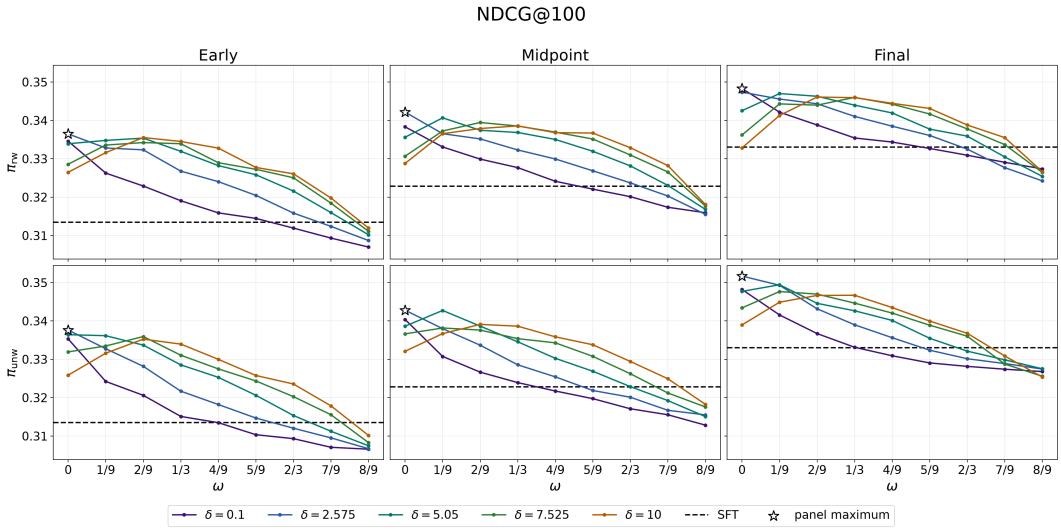


Figure 6: Variation of NDCG with ω across different initial π_0. Each curve is one margin δ.

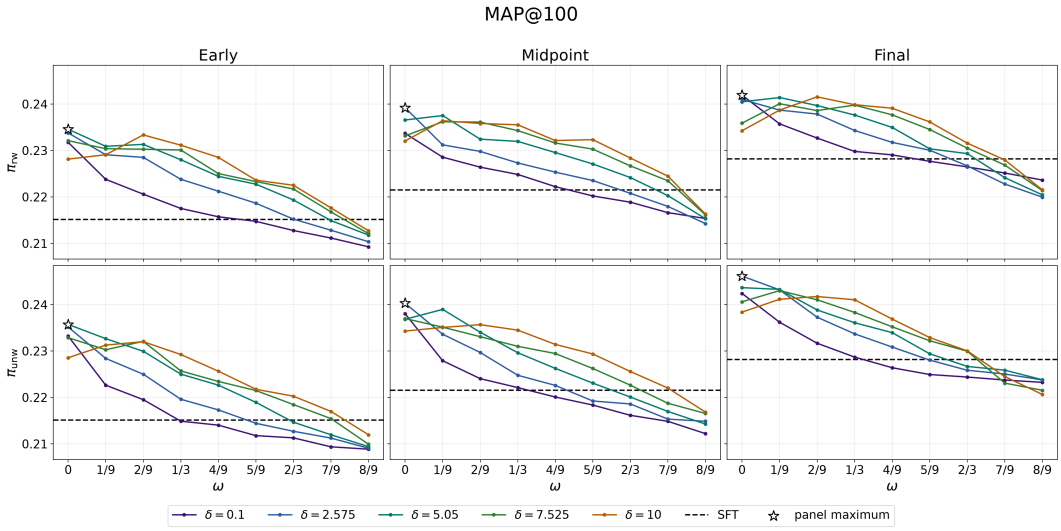


Figure 7: Variation of MAP with ω across different initial π_0. Each curve is one margin δ.

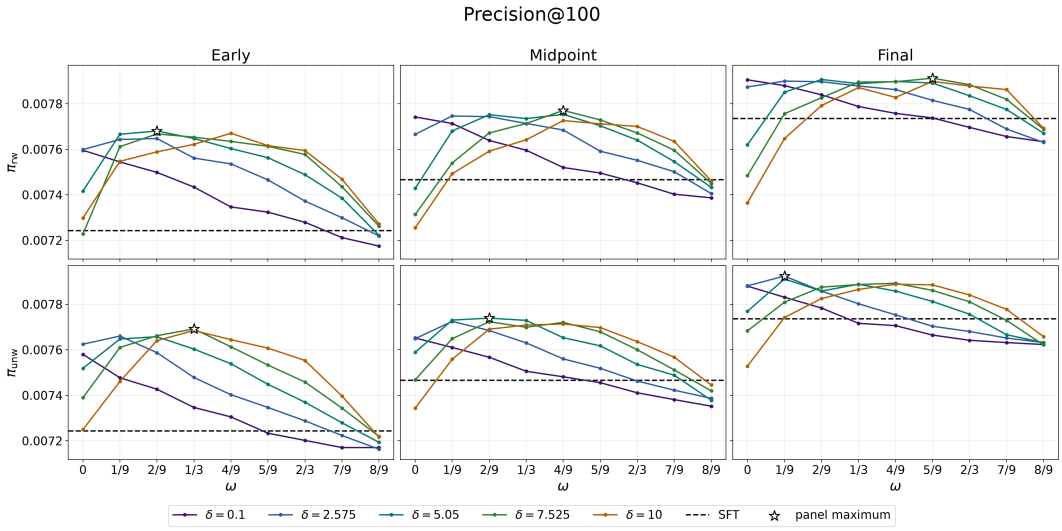


Figure 8: Variation of Precision with ω across different initial π_0. Each curve is one margin δ.