MAXIMIN INFERENCE-TIME ALIGNMENT WITH DIVERSE REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Inference-Time Alignment (ITA) improves a learned model's output through sampling and search heuristics without modifying its underlying weights. While effective in guiding the model response, in reality a single model often serves a diverse group of users or is evaluated over multiple categories (e.g. accuracy, safety) with different reward models. This raises the question of how a shared and finite sampling budget should be allocated to ensure response quality across multiple groups of users or different evaluation criteria. We formulate this problem as a *maximin budget allocation* problem and study ITA under exact-reward and proxy reward settings. In the exact-reward setting, we upper bound the maximin regret of a greedy variant of Best-of-N algorithm (Greedy-BoN) and show that the regret goes to zero as the budget goes to infinity. However, in the proxy reward setting, the above approach fails even as the sampling budget goes to infinity. To overcome this, we develop Batched Greedy Pessimistic (BGP) algorithm that allocates each batch to the group using pessimistic response selection, and bound its maximin regret with respect to the proxy reward floor. Empirically, we evaluate Greedy-BoN and BGP on a dataset with five different domain (groups) and show that Greedy-BoN performs best with respect to the proxy reward, whereas BGP performs best when evaluated against an external judge, highlighting the importance of pessimism in inference-time alignment.

1 INTRODUCTION

Inference-Time Alignment (ITA) is a family of search-based techniques that align Large Language Models (LLMs) with human preferences during the generation process. The base policy stays fixed, and a learned reward model guides the selection of its outputs, without updating model parameters. ITA techniques such as Best-of-N (Beirami et al., 2025; Gui et al., 2024) have gained traction due to their low computation cost and high versatility, as it allows flexible test-time alignment that is responsive to a user's dynamic preferences and on-demand goals, using the same base model.

In practice, however, a single model serves many users whose preferences vary depending on the context and task, as well as due to their idiosyncratic differences. For example, a user writing code may prioritize correctness, whereas another group seeking advice on a high-stakes decision may prioritize safety or pluralistic values. Similarly, response quality could have several dimensions such as helpfulness, accuracy, verbosity, and safety (Wang et al., 2024), and ensuring that all users or evaluation categories are treated in a balanced way (Chakraborty et al., 2024; Ramesh et al., 2024). Training-based approaches to pluralistic or social-choice theoretic alignment (Sorensen et al., 2024; Conitzer et al., 2024) consider nuanced diversity of preferences, but they commit to a single trained policy and require expensive training. ITA provides a unique opportunity because the reward model acts only at generation time, and allows responses to be evaluated by different reward models.

We formulate this setting as a *budget allocation problem*: a single base model serves multiple groups, each with its own reward model, and a shared sampling budget B that must be distributed among the groups. When different groups have their own preference model, the system must decide how to allocate a fixed sampling budget of B so that no group receives a low-quality response. Groups may differ in the quality of the base model's responses, how much they benefit from additional samples, and the accuracy of their reward models.

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.

A uniform approach that allocates every group the same budget, therefore, may not adhere to differences between the groups. On the other hand, allocating budgets to maximize total rewards can favor groups with the largest marginal gains and leave the worst-off groups behind. Recent work has focused on budget allocation problem across queries at inference time (Damani et al., 2025; Zuo & Zhu, 2026). However, these methods optimize aggregate performance and do not explicitly account for group preferences.

Drawing from welfare economics, we formalize this budget allocation problem under the Rawlsian maximin principle, and ask the following research question: *How should a shared and finite sampling budget be allocated across groups with heterogeneous reward models so as to maximize the reward of the worst-off group?*

1.1 Our Contributions

We formulate the *regret* of an algorithm as the gap between the best worst-group reward achievable by the optimal fixed allocation and the expected reward of the worst group under the algorithm. Since optimizing a proxy reward can significantly degrade the response quality through *reward hacking* (Gao et al., 2023), we consider two settings, one in which the true rewards are known, and one in which only proxy rewards are observed. Our contributions are as follows.

1. **Greedy-BoN algorithm under exact rewards.** Under the exact reward setting, we propose a Greedy-BoN algorithm (Algorithm 1), which iteratively allocates a sample to the group with the lowest current reward. We show that for every realization of the responses, this algorithm returns an optimal allocation (Lemma 1). Furthermore, under a mild polynomial tail assumption, we prove that the regret of the Greedy-BoN algorithm is bounded by $O(B^{-\gamma_{\min}})$ (Theorem 1), where $\gamma_{\min} \in (0, 1]$ measures the rate of the "hardest" group.
2. **Robust allocation under proxy rewards.** When reward models are imperfect, we show how Greedy-BoN fails due to the mismatch between proxy and true rewards, even when the budget goes to infinity (Theorem 2). Inspired by the Inference-Time Pessimism (ITP) (Huang et al., 2025), we develop an algorithm called Batched Greedy Pessimistic (BGP) algorithm (Algorithm 2), and show that its regret is bounded by a floor determined by reward model error and coverage, plus an explicit rate that decays with total budget B (Theorem 3).
3. **Empirical validation.** We evaluate both methods using different domains as distinct preference groups, with proxy reward models for inference-time alignment and independent Judge models for evaluation. The experiments show that Greedy-BoN achieves strong worst-group proxy rewards, while BGP attains the best worst-group judge score, demonstrating the significance of pessimism when reward models are imperfect.

1.2 Related Work

ITA and reward hacking. ITA improves model outputs without changing base models' parameters, e.g., through parallel sampling, decoding-time search, or iterative textual feedback (Liu et al., 2024; Xu et al., 2025; Li et al., 2025). The canonical example is Best-of-N, which samples N responses in parallel and returns the one with the highest proxy reward (Beirami et al., 2025). Since learned reward models are imperfect proxies for human preferences, BoN is vulnerable to reward hacking (Gao et al., 2023). This problem has been addressed through reward regularization or uncertainty-aware selection (Jinnai et al., 2025; Khalaf et al., 2025; Huang et al., 2025). Our work is closest to the ITP algorithm in Huang et al. (2025), which solves a χ^2-regularized reward objective.

Alignment with general or heterogeneous preferences. Standard RLHF and ITA assume a single reward model, which can fail to represent general population preferences. Thus, one line of work relaxes the reward model assumption and aligns a general preference model by solving a two-player zero-sum game (Munos et al., 2024; Ye et al., 2024; Zhang et al., 2024; Zhou et al., 2025; Hosseini et al., 2026). Another line models heterogeneous preferences among different groups explicitly. Among post-training methods, Chakraborty et al. (2024); Ramesh et al. (2024) consider max-min or group-robust objectives, (Jang et al., 2023; Poddar et al., 2024) study personalized LLM alignment, and Rame et al. (2023); Zhou et al. (2024); Yang et al. (2024) study multi-objective alignment across different rewards. At inference time, Shi et al. (2024); Lin et al. (2025); Chen et al. (2025) study alignment with multiple objectives.

Allocation of inference-time computation. A related direction of work studies the allocation of test-time computation. Prior results show that the value of additional computation varies across inputs (Snell et al., 2025). Based on this observation, adaptive methods allocate more computation to inputs with larger predicted marginal gains (Damani et al., 2025) or learn the allocation online through a bandit formulation (Zuo & Zhu, 2026). These methods optimize aggregate performance across inputs and do not account for group-dependent preferences and fairness.

2 PRELIMINARIES

Problem setup. Denote $\mathcal{X}$ as the prompt space and $\mathcal{Y}$ as the response space. The fair allocation problem consists of a central designer and k groups. The central designer has access to a base model $\pi_{\text{ref}} : \mathcal{X} \to \Delta(\mathcal{Y})$, where $\pi_{\text{ref}}(y \mid x)$ denotes the probability that the base model outputs a response y given the prompt x. Group $i \in [k]$ is specified by a prompt $x_i \in \mathcal{X}$ along with the distribution $\pi_{i,\text{ref}}(\cdot) := \pi_{\text{ref}}(\cdot \mid x_i)$ and an unknown ground-truth reward $r_i^\star : \mathcal{X} \times \mathcal{Y} \to [0, R_{\max}]$. The reward $r_i^\star(x_i, y)$ evaluates the quality of a response y given the prompt x_i. As a proxy, the central designer has access to a proxy reward model $\hat{r}_i : \mathcal{X} \times \mathcal{Y} \to [0, R_{\max}]$. The designer interacts with the groups over B rounds. At round $t \in [B]$ it selects a group $i_t \in [k]$, and then group i_t draws a fresh response $Y_t \sim \pi_{i_t,\text{ref}}(\cdot)$, and observes its proxy score. We write $\hat{n}_i := \sum_{t=1}^{B} \mathbf{1}\{i_t = i\}$ for the number of draws group i receives, such that $\sum_{i=1}^{k} \hat{n}_i = B$. After the B rounds, for each group i, the designer picks a response $\hat{y}_i$ from the sampled pool. We denote the distribution on responses of $\hat{n}_i$ samples in group i as $\hat{\pi}_i$.

Reward model error and coverage. The difficulty of individual-group inference-time alignment is specified by two notions: reward model error and coverage (Huang et al., 2025). For prompt x_i, we measure the quality of the reward model $\hat{r}_i$ via the square error with respect to $r_i^\star$, with responses drawn from the base policy:

$$\varepsilon_i^2 := \mathbb{E}_{y \sim \pi_{i,\text{ref}}}[(\hat{r}_i(x_i, y) - r_i^\star(x_i, y))^2]. \tag{1}$$

On the other hand, coverage measures how much a policy π concentrates probability mass relative to the base policy $\pi_{i,\text{ref}}$. Formally, we define the coverage of policy π for group i:

$$\mathcal{C}_i^\pi := \mathbb{E}_{y \sim \pi(\cdot)} \left[\frac{\pi(y)}{\pi_{i,\text{ref}}(y)} \right].$$

Regret notions. In the following, we introduce a deterministic benchmark, and the regret as the reward difference between the benchmark and the worst-off group.[1] The reward value of an output distribution $\hat{\pi}_i$ in group i is $J_i(\hat{\pi}_i) := \mathbb{E}_{y \sim \hat{\pi}_i}[r_i^\star(x_i, y)]$. When group i samples n responses: $Y_i^n = \{Y_{i,1}, Y_{i,2}, \ldots, Y_{i,n}\} \sim \pi_{i,\text{ref}}$, we denote the best realized value from the set Y_i^n as

$$\widetilde{V}_i(n) := \max_{s \le n} r_i^\star(x_i, Y_{i,s})$$

with initialization $\widetilde{V}_i(0) = -\infty$, and the expected best-of-n value

$$V_i(n) := \mathbb{E}_{Y_i^n}[\widetilde{V}_i(n)].$$

Then we denote the deterministic maximin benchmark as

$$V^\star := \max_{\sum_i n_i = B} \min_{i \in [k]} V_i(n_i), \qquad \mathbf{n}^\star \in \arg \max_{\sum_i n_i = B} \min_{i \in [k]} V_i(n_i).$$

In words, $V^\star$ is the highest worst-group expected reward that any optimal fixed allocation of the B samples can achieve, given full knowledge of each group's (best-of-n) value V_i.

Definition 1. *Suppose an algorithm returns responses $\{\hat{y}_1, \ldots, \hat{y}_k\}$ for each of the k groups. Then we define the regret as*

$$\mathcal{R}_B = V^\star - \min_{i \in [k]} \mathbb{E}_{\hat{y}_i} \left[r_i^\star(x_i, \hat{y}_i) \right].$$

Intuitively, the regret is the gap between $V^\star$ and the value of the worst-off group under the algorithm.

[1] Since the benchmark is deterministic, regrets can be negative. We will introduce an adaptive benchmark and discuss a new regret notion in Section A.2.

Algorithm 1 Greedy BoN Algorithm

1: Input : budget B; samplers $\pi_{i,\text{ref}}$, reward oracles $r_i^\star, \forall i \in [k]$.
2: Initialize $n_i \leftarrow 0$, $\widetilde{V}_i \leftarrow -\infty$, and $\hat{y}_i \leftarrow \emptyset$, $\forall i \in [k]$.
3: **for** $t = 1, 2, \ldots, B$ **do**
4: Calculate the index $i_t \in \arg\min_{i\in[k]} \widetilde{V}_i$ (ties broken arbitrarily).
5: Draw a response $y \sim \pi_{i_t,\text{ref}}$ and set $n_{i_t} \leftarrow n_{i_t} + 1$.
6: **if** $r_{i_t}^\star(y) > \widetilde{V}_{i_t}$ **then**
7: $\widetilde{V}_{i_t} \leftarrow r_{i_t}^\star(y)$ and $\hat{y}_{i_t} \leftarrow y$. ▷ update both the value and its maximizer
8: **end if**
9: **end for**
10: Output : responses $(\hat{y}_1, \ldots, \hat{y}_k)$.

3 Exact Reward Setting

To warm up, we start by focusing on a simpler case where we have access to the ground-truth reward model (r_i^*) of all groups. Thus, all reward estimates are accurate, i.e. $\hat{r}_i = r_i^\star, \forall i \in [k]$. We propose a simple Greedy algorithm based on the best-of-N sampling algorithm (Algorithm 1). Formally, at each step it selects the group whose current best reward $\widetilde{V}$ is lowest (Line 4). Then it draws one response for that group from the base model, and updates the group's maximum reward. In order to upper bound the regret of Algorithm 1 we will introduce an assumption regarding the tail distribution of the rewards for the groups.

Let us write F_i for the CDF of $r_i^\star(x_i, y)$ under $y \sim \pi_{i,\text{ref}}$, i.e., $F_i(t) := \mathbb{P}_{y\sim\pi_{i,\text{ref}}}\big(r_i^\star(x_i, y) \le t\big)$. We denote $\overline{R}_i := \sup\{t : F_i(t) < 1\} \le R_{\max}$ as the essential supremum. In the following, we assume that the reward distribution places polynomial mass near its essential supremum.

Assumption 1 (Polynomial Tail Probability.)**.** *There exists $u_0 \in (0, R_{\max}]$ and, for each $i \in [k]$, constants $c_i > 0$ and $\alpha_i \ge 1$ with $c_i u_0^{\alpha_i} \le 1$, such that*

$$\mathbb{P}_{y\sim\pi_{i,\text{ref}}}\big[r_i^\star(x_i, y) \ge \overline{R}_i - u\big] \ge c_i u^{\alpha_i} \qquad \text{for all } u \in (0, u_0].$$

We write $\gamma_i := 1/\alpha_i \in (0, 1]$.

The value γ_i controls how fast the near-optimal mass vanishes as the tolerance u shrinks. When $\gamma_i = 1$, the mass is at least linear in u; whereas a small γ_i means near-optimal responses are rare and the group needs many draws to approach $\overline{R}_i$. In fact, if we assume response space $\mathcal{Y}$ is finite, then this assumption trivially holds with $\gamma_i = 1$.

The key structural fact is that the realized value is monotone *pathwise*: for any sample path $Y_{i,1}, Y_{i,2}, \ldots$, we have $\widetilde{V}_i(n) \ge \widetilde{V}_i(n-1), \forall n \ge 1$. As a consequence, the greedy algorithm is optimal on every sample path (i.e., after responses are sampled).

Lemma 1 (Ex-post optimality of Algorithm 1)**.** *Let $\widetilde{V}_1, \ldots, \widetilde{V}_k : \mathbb{N}^+ \to \mathbb{R}^+$ be arbitrary non-decreasing sequences and let $\hat{\mathbf{n}} = (\hat{n}_1, \ldots, \hat{n}_k)$ with $\sum_{i=1}^k \hat{n}_i = B$ be the counts produced by Algorithm 1. Then*

$$\min_{i\in[k]} \widetilde{V}_i(\hat{n}_i) = \max_{\sum_i m_i = B} \min_{i\in[k]} \widetilde{V}_i(m_i).$$

The following theorem provides a finite-budget regret bound of Algorithm 1.

Theorem 1. *Under the exact reward setting $\hat{r}_i = r_i^\star, \forall i \in [k]$ and Assumption 1, we have*

$$\mathcal{R}_B \le 2 \max_i \left(\frac{1 + \log k}{c_i \lfloor B/k \rfloor}\right)^{\gamma_i} + R_{\max} \sum_i \exp\left(-c_i \lfloor B/k \rfloor u_0^{\alpha_i}\right),$$

assuming that $\left\lfloor \frac{B}{k} \right\rfloor \ge \frac{1+\log k}{c_i u_0^{\alpha_i}}, \forall i \in [k]$.

Remark 1. *Since the second term of the regret goes to zero at a rate of $\exp(-B/k)$, the regret bound essentially decays at the rate $O((k \log k/B)^{\gamma_{\min}})$ with $\gamma_{\min} := \min_i \gamma_i$. Definition 1 applies the minimum over groups after taking expectations, and here we define a stronger **ex-post** regret that takes expectation after applying minumum over groups,*

$$\mathcal{R}_B^{\text{ep}} := V^\star - \mathbb{E}_{\hat{y}_1,\ldots,\hat{y}_k}[\min_{i\in[k]} r_i^\star(x_i, \hat{y}_i)].$$

We have $\mathcal{R}_B \le \mathcal{R}_B^{\text{ep}}$ by Jensen inequality. Our proof essentially rests on the ex-post optimality in Lemma 1: on every sample path, Algorithm 1 attains a floor at least as high as that of any fixed allocation, in particular the uniform allocation, and therefore it suffices to bound the uniform allocation. Assumption 1 controls the tail probability at a polynomial rate, and Lemma 3 turns it, via a union bound over the k groups, into the ex-post regret.

Remark 2. *Assumption 1 is used primarily to derive a finite-budget regret upper bound for a large class of reward distributions. It is not required to our algorithm in order to achieve a vanishing regret bound. In Section A.3, we derive a distribution-dependent regret bound of Algorithm 1 and show it still converges to 0 as $B \to \infty$ without any assumption on the tail probability.*

4 Proxy Reward Setting

Thus far, we assumed that the true rewards are exactly known. However, the ground-truth reward models are unavailable in practice, and they are estimated from responses available from each group. Thus, the proxy reward model $\hat{r}_i$ is often *imperfect* (i.e. $\hat{r}_i \neq r_i^\star$). This setting accurately captures the *reward-modeling* phase of alignment and will be the focus for the rest of the section.

Recall the definition of reward model error ε_i between the proxy reward model $\hat{r}_i$ and true reward model $r_i^\star$. When $\varepsilon_i > 0$, a reader might believe that the guarantees of Greedy-BoN can be extended perhaps with an additional dependence on ε_i. However, in a stark contrast to the result for a single group (Huang et al., 2025), we show that this is generally impossible. In the following theorem, we construct an instance with two groups ($k = 2$) where Greedy-BoN can have non-vanishing regret. Group 1 has a good response and an over-scored response, with the base model sampling them with probability $1 - p$ and p respectively. Both true reward and proxy reward models score the good response $1/2$, while the true reward scores the other response 0 and the proxy reward scores it 1. Group 2 has a single deterministic response, and both reward models score $3/4$. We show that Algorithm 1 searches for the over-scored response, and allocates it to group 1 with probability tending to 1 as $B \to \infty$.

Theorem 2. *There exists an instance with two groups, $\varepsilon_1^2 = p$ and $\varepsilon_2^2 = 0$, in which Algorithm 1 has regret*

$$\mathcal{R}_B = \frac{1}{2}\Big[1 - p^{B-1} - (1-p)^{B-1}\Big] \longrightarrow \frac{1}{2} \text{ as } B \to \infty.$$

4.1 Within-Group Pessimism

In the proxy reward setting, the best-of-N reward is not a good proxy as two groups can have similar proxy rewards despite having very different ground truth rewards. Instead, we consider a χ^2-regularized reward objective for each group, and construct a pessimistic surrogate of the true value of the regularized policy. We make the following assumption about the errors ε_i.

Assumption 2. *For each group $i \in [k]$, there is a known reward error bound $\bar{\varepsilon}_i$ such that $\varepsilon_i \le \bar{\varepsilon}_i$.*

Let $D_{\chi^2}(\pi \,\|\, \pi_{i,\text{ref}}) := \frac{1}{2}\,\mathbb{E}_{y\sim\pi_{i,\text{ref}}}\left[\left(\frac{\pi(y)}{\pi_{i,\text{ref}}(y)} - 1\right)^2\right]$ denote the χ^2-divergence between policy π and $\pi_{i,\text{ref}}$, so that $D_{\chi^2}(\pi \,\|\, \pi_{i,\text{ref}}) = (\mathcal{C}_i^\pi - 1)/2$. Now fix group index i and a parameter $\beta > 0$, the goal is the χ^2-regularized objective,

$$\max_{\pi\in\Delta(\mathcal{Y})} \{\mathbb{E}_{y\sim\pi}[\hat{r}_i(x_i, y)] - \beta \cdot D_{\chi^2}(\pi||\pi_{i,\text{ref}})\}. \tag{2}$$

The solution to the optimization problem defined in eq. (2) is

$$\pi_{i,\beta}(y) = \pi_{i,\text{ref}}(y) \cdot \left(\beta^{-1}\left(\hat{r}_i(y) - \lambda_{i,\beta}\right)\right)_+ \tag{3}$$

where we define $z_+ = \max\{0, z\}$ and $\lambda_{i,\beta}$ is a normalization constant defined as $\mathbb{E}_{y\sim\pi_{i,\text{ref}}}\left[\left(\beta^{-1}\left(\hat{r}_i(y) - \lambda_{i,\beta}\right)\right)_+\right] = 1$.

Pessimistic surrogate. Let us write $\text{Obj}_i(\beta)$ to be the optimal value of the objective defined in eq. (2). Since $\pi_{i,\beta}$ is the optimal solution we have,

$$\text{Obj}_i(\beta) := \mathbb{E}_{y\sim\pi_{i,\beta}}[\hat{r}_i(y)] - \beta\, D_{\chi^2}(\pi_{i,\beta} \,\|\, \pi_{i,\text{ref}}).$$

Since $\mathrm{Obj}_i(\beta)$ is computed from the proxy reward, it may overestimate the true value $J_i(\pi_{i,\beta})$. We therefore construct a pessimistic surrogate

$$G_i(\beta) := \mathrm{Obj}_i(\beta) - \bar{\varepsilon}_i - \bar{\varepsilon}_i^2/(2\beta).$$

An important observation is that $G_i(\beta)$ never over-estimates the true value of the pessimistic policy, i.e., $G_i(\beta) \le J_i(\pi_{i,\beta})$ (Lemma 7). Our new algorithm will use this pessimistic surrogate.

Empirical estimates. We estimate the pessimistic surrogate $G_i(\beta)$ from samples. Assume that group i has access to n i.i.d. draws $y_{i,1}, \dots, y_{i,n} \sim \pi_{i,\mathrm{ref}}$. Denote $\pi_{i,n} := \frac{1}{n}\sum_{j=1}^n \delta_{y_{i,j}}$ as the empirical distribution of these draws. Replacing $\pi_{i,\mathrm{ref}}$ by $\pi_{i,n}$ gives estimate of the optimal policy $\pi_{i,\beta}$ from finite samples.

$$\widehat{\pi}_{i,\beta,n}(y_{i,j}) = \frac{1}{n} \cdot \left(\beta^{-1}\left(\hat{r}_i(y_{i,j}) - \widehat{\lambda}_{i,\beta,n}\right)\right)_+ \tag{4}$$

where $\widehat{\lambda}_{i,\beta,n}$ is a normalization constant such that $\frac{1}{n}\sum_{j=1}^n \left(\beta^{-1}\left(\hat{r}_i(y_{i,j}) - \widehat{\lambda}_{i,\beta,n}\right)\right)_+ = 1$. We can now define the empirical objective value and the empirical pessimistic surrogate as

$$\widehat{\mathrm{Obj}}_i(\beta; n) := \mathbb{E}_{y\sim\widehat{\pi}_{i,\beta,n}}\left[\hat{r}_i(y)\right] - \beta\, D_{\chi^2}(\widehat{\pi}_{i,\beta,n} \,\|\, \pi_{i,n}), \qquad \widehat{G}_i(\beta; n) := \widehat{\mathrm{Obj}}_i(\beta; n) - \bar{\varepsilon}_i - \frac{\bar{\varepsilon}_i^2}{2\beta}.$$

The policy (4) and the surrogate $\widehat{G}_i(\beta; n)$ depend on β, which must be selected from data. In order to choose β, we introduce the notion of a finite geometric grid.

Definition 2 (Regularization Grid.). *Fix $\beta_{\min} > 0$, $\rho > 1$ and grid size $N_\beta \in \mathbb{Z}^+$. The geometric regularization grid is $\mathcal{G} := \{\beta_{\min}\rho^j : j = 0, \dots, N_\beta - 1\}$, with largest point $\beta_{\max} := \beta_{\min}\rho^{N_\beta - 1}$.*

When β is chosen from the grid $\mathcal{G}$, we show that the empirical pessimistic surrogate $\widehat{G}_i(\beta; n)$ is close to the true pessimistic surrogate $G_i(\beta)$. Indeed for a given $\beta > 0$ and $\delta \in (0, 1)$ let us define

$$\Delta^{\mathrm{obj}}(n, \beta, \delta) := 5R_{\max}\left[\sqrt{\frac{M_\beta l}{n}} + \frac{M_\beta l}{n}\right], \quad \Delta^{\mathrm{ITP}}(n, \beta, \delta) := R_{\max}\sqrt{\frac{M_\beta l}{n}}, \quad \Delta := \Delta^{\mathrm{obj}} + \Delta^{\mathrm{ITP}}.$$

where $M_\beta = 1 + R_{\max}/\beta$, and $l = \log\left(16kN_\beta B(B^2+2)/\delta\right)$. Then Proposition 3 shows that with a high probability, Δ^{obj} bounds the deviation of the empirical pessimistic surrogate ($\widehat{G}_i(\beta; n)$) from its population counterpart ($G_i(\beta)$), while Δ^{ITP} bounds the amount by which the empirical surrogate ($\widehat{G}_i(\beta; n)$) can exceed the true value of the empirical policy $\widehat{\pi}_{i,\beta,n}$.

4.2 Allocation Across Groups: A Novel BGP Algorithm

In this section, we introduce the Batched Greedy Pessimistic Algorithm (Algorithm 2) that uses the pessimistic surrogates to allocate samples across groups. In order to control the statistical error $\Delta(n, \beta, \delta)$ in estimating the pessimistic surrogate, we introduce a statistical tolerance η and the following admissible regularization set.

Definition 3 (Admissible Regularization Set.). *Fix a **statistical tolerance** $\eta > 0$ and a confidence level $\delta \in (0, 1)$. For $n \in [B]$, define the admissible set*

$$\mathcal{B}(n) := \{\beta \in \mathcal{G} : \Delta(n, \beta, \delta) \le \eta\}.$$

Assume $\mathcal{B}(B) \neq \varnothing$, and let $n_0 := \min\{n \in [B] : \mathcal{B}(n) \neq \varnothing\}$.

Since $\Delta(n, \beta, \delta)$ decreases with β, the set $\mathcal{B}(n)$ is nonempty if and only if $\beta_{\max}$ is admissible, so $n_0 = \min\{n \in [B] : \Delta(n, \beta_{\max}, \delta) \le \eta\}$. The tolerance η controls how many samples are required before a regularization level becomes admissible. Algorithm 2 begins by picking at least n_0 samples per group and computing the pessimistic surrogate. For $n \ge n_0$, define the best admissible population surrogate by

$$P_i(n) := \max_{\beta\in\mathcal{B}(n)} G_i(\beta).$$

The finite-sample counterpart of $P_i(n)$ is the *pessimistic index*

$$L_i(n) := \max_{\beta\in\mathcal{B}(n)} \widehat{G}_i(\beta; n), \qquad \widehat{\beta}_i(n) \in \arg\max_{\beta\in\mathcal{B}(n)} \widehat{G}_i(\beta; n),$$

Algorithm 2 Batched Greedy Pessimistic (BGP)

Require: budget B; error bounds $(\overline{\varepsilon}_i)_{i\in[k]}$; tolerance $\eta > 0$; confidence $\delta \in (0,1)$; grid $\mathcal{G}$; samplers $(\pi_{i,\mathrm{ref}})_{i\in[k]}$; reward models $(\hat{r}_i)_{i\in[k]}$.

Compute n_0 (Definition 3).
for $i = 1, \dots, k$ **do**
  Draw $y_{i,1}, \dots, y_{i,n_0} \sim \pi_{i,\mathrm{ref}}$.
  Set $n_i \leftarrow n_0$ and compute $L_i(n_i)$ by INDEX(i, n_i).
end for
while $\sum_j n_j < B$ **do**
  Select $i_t \in \arg\min_{i\in[k]} L_i(n_i)$. and $b \leftarrow \min\{n_{i_t},\ B - \sum_j n_j\}$.
  Draw fresh responses $y_{i_t,n_{i_t}+1}, \dots, y_{i_t,n_{i_t}+b} \sim \pi_{i_t,\mathrm{ref}}$.
  Set $n_{i_t} \leftarrow n_{i_t} + b$ and recompute $L_{i_t}(n_{i_t})$ by INDEX(i_t, n_{i_t}).
end while
for $i = 1, \dots, k$ **do**
  Set $\hat{n}_i \leftarrow n_i$, $\widehat{\beta}_i \leftarrow \widehat{\beta}_i(\hat{n}_i)$, and draw $\hat{y}_i \leftarrow$ SAMPLE$(i, \hat{n}_i, \widehat{\beta}_i)$. ▷ $\hat{y}_i \sim \widehat{\pi}_{i,\widehat{\beta}_i,\hat{n}_i}$
end for
return $(\hat{y}_1, \dots, \hat{y}_k)$.

where $\widehat{\beta}_i(n)$ is the admissible regularization that maximize the empirical pessimistic surrogate $\widehat{G}_i$.

The BGP algorithm. Algorithm 2 first draws n_0 responses for every group and computes its pessimistic index $L_i(n_0)$ and maximizer $\widehat{\beta}_i(n_0)$ for each group. It then repeatedly selects a group with the smallest pessimistic index and draws a new batch for that group. The batch size equals the group's current sample count, or the remaining budget if smaller. This effectively doubles the samples per group, and therefore, the number of index computations per group is at most $O(\log B)$. After each batch, the algorithm uses subroutine INDEX to recompute the selected group's pessimistic index using all its samples. After the budget B is exhausted, group i returns a response drawn from $\widehat{\pi}_{i,\widehat{\beta}_i,\widehat{n}_i}$, where $\widehat{n}_i$ is its final sample count and $\widehat{\beta}_i$ is the selected regularization. Algorithm 3 in Section C provides details of the INDEX method to compute the pessimistic index $L_i(n)$ and maximizer $\widehat{\beta}_i(n)$, SAMPLE method to sample a response from the pool.

4.3 Theoretical Guarantees

In this section, We provide the regret upper bound for Algorithm 2. Naturally, we choose the comparator policy for group i as $\pi_i^\star$, the law of the true-reward best-of-$n_i^\star$ response under the maximin allocation $\mathbf{n}^\star$. For simplicity we write the coverage $C_i := \mathcal{C}_i^{\pi_i^\star}$ for group i.

Theorem 3 (Regret Upper Bound). *Under Assumption 2, fix any $\beta^0 = (\beta_i^0)_{i\in[k]} \in (0,\infty)^k$ and run Algorithm 2 with confidence $\delta \in (0,1)$ and tolerance $\eta = \eta_B(\beta^0) := 24R_{\max}\sqrt{l\sum_{i\in[k]} M_{\beta_i^0}/B}$. Then, for every profile $(\beta_i)_{i\in[k]} \in \mathcal{G}^k$ with $\beta_i \geq \beta_i^0$ for all i,*

$$\mathcal{R}_B \;\leq\; \max_{i\in[k]} \Bigg(\underbrace{2\sqrt{C_i}\,\overline{\varepsilon}_i}_{\textit{floor}} + \underbrace{\frac{\beta_i C_i}{2} + \frac{\overline{\varepsilon}_i^2}{2\beta_i}}_{\textit{bias}(\beta_i)} \Bigg) + \underbrace{96R_{\max}\sqrt{\frac{l\sum_i \left(1 + R_{\max}/\beta_i^0\right)}{B}}}_{\textit{variance}(\beta^0,B)} + \delta R_{\max}.$$

The floor is the price of reward-model error that does not vanish with B, and it matches the single-group lower bound $\Omega(\sqrt{\mathcal{C}(x)}\varepsilon(x))$ (Huang et al., 2025). Note that in Theorem 3, β_i and β_i^0 are free regularization levels, and are not chosen by the Algorithm 2. We defer the formal discussion on how to choose regularization to minimize the regret to Corollary 1 in Section D.3.

We introduce two choices of β_i, β_i^0 in Corollary 1, and both have the same floor term $O(\max_{i\in[k]}\sqrt{C_i}\overline{\varepsilon}_i)$. The difference lies in the additional term that decays with B. That term in the first choice decays as $O((\frac{C_{\max}}{B})^{\frac{1}{3}})$, where $C_{\max} = \max_i C_i$, while the additional term in the second choice decays as $O((\frac{\sum_{i=1}^k \sqrt{C_i}/\overline{\varepsilon}_i}{B})^{\frac{1}{2}})$. In general $C_i, C_{\max}$ can grow with B, and we can bound it by a constant $C_i \leq 1/p_i$ (Lemma 9), where $p_i = \pi_{i,\mathrm{ref}}(\{y : r_i^\star(y) = \overline{R}_i\})$ is the probability mass on the optimal responses. At the first glance, the rate of the second choice seems better, but

Table 1: BGP statistics at total budget $B = 255$. Each row reports a group's reward-error estimate, budget allocation, and expected rewards under the 3B proxy and 7B Judge. Allocations and rewards are recorded after the full budget has been allocated in each trial and then averaged over 200 trials. BGP allocates budgets based on both the group error $\bar{\varepsilon}_i$ and the proxy reward.

Group	Error $\bar{\varepsilon}_i$	Budget allocation	Avg. proxy reward	Avg. eval reward
academia	0.1705	90.600	0.5098	0.4668
change my view	0.1466	81.750	0.5269	0.4592
engineers	0.1006	56.775	0.5064	0.3331
culinary	0.1127	14.600	0.5786	0.3300
science fiction	0.1465	11.275	0.5893	0.3282

its coefficient grows with $\sqrt{1/\bar{\varepsilon}_i}$, which can be large if $\bar{\varepsilon}_i$ is small. Thus, we conclude that first rate $O((\frac{C_{\max}}{B})^{\frac{1}{3}})$ is smaller when we have moderate size B, while $O((\frac{\sum_{i=1}^{k}\sqrt{C_i/\bar{\varepsilon}_i}}{B})^{\frac{1}{2}})$ is better when B is large.

5 Experiments

We empirically evaluate Greedy-BoN and BGP algorihtms, using different domains to represent distinct preference groups. By comparing our algorithms with uniform-BoN and BGP variants, we assess the worst-group performance with both proxy and Judge evaluation rewards. We further investigate the advantage of BGP algorithm beyond the worst group and analyze budget allocation to understand how group rewards and uncertainty shape BGP's sampling decisions. We provide the source code[2] and further details of our experimental settings and results in Appendix E and F.

5.1 Experimental setup

Dataset. We use five groups (domains) from `stanfordnlp/SHP` (Ethayarajh et al., 2022). Each group provides 25 tuning prompts and 25 disjoint evaluation prompts from the test split. We pre-generate a response pool for each prompt using `Qwen2.5-3B-Instruct` (Qwen et al., 2025). For each experiment, we run $T = 200$ paired evaluation trials and report 95% confidence interval. In each trial, we independently sample one prompt from each group and the algorithm allocates the total sampling budget $B \in \{65, 90, 125, 180, 255\}$ across all groups.

Language Models. For each group, we train a proxy reward model initialized from `Qwen2.5-3B-Instruct` to guide the algorithms. For independent evaluation, we train a corresponding Judge reward model from `Qwen2.5-7B-Instruct` using the same procedure and score all candidate responses. The error bounds $\bar{\varepsilon}_i$ are estimated from discrepancies between proxy and Judge rewards and reported in Table 1.

Methods and baselines. To examine the performance of *Greedy-BoN* (Algorithm 1), we compare it with *Uniform-BoN*, which allocates the budget equally across groups. For *BGP* (Algorithm 2), we examine automatic regularization and adaptive allocation through two controlled variants. *BGP-Fixed* replaces automatic β selection with a global value tuned using proxy scores on the tuning split. *Uniform-Pess-Fixed* retains the same fixed-β pessimistic selection but allocates the budget uniformly.

5.2 Main results

How do proposed algorithms perform in proxy rewards and Judge evaluation? Figure 1 (a) compares the worst-group expected rewards across sampling budgets, with error bars indicating pointwise 95% confidence intervals. Under proxy evaluation, Greedy-BoN consistently outperforms Uniform-BoN, demonstrating the benefit of adapting the allocation to group-specific rewards. Under the independent judge models, BGP achieves the highest worst-group evaluation scores across all budgets, demonstrating its effectiveness under independent evaluation..

[2]The source code is available at anonymized repository link.

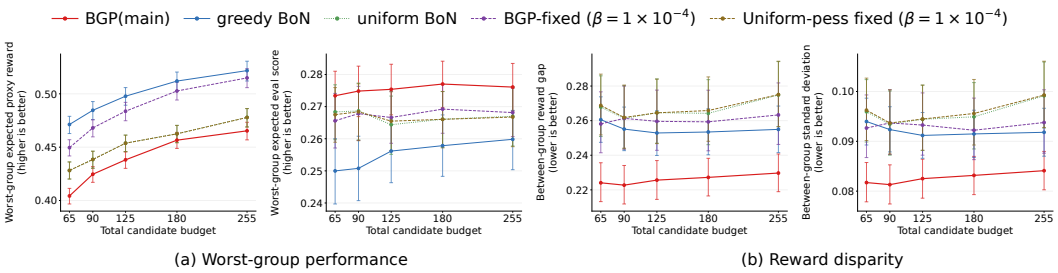

Figure 1: Performance across total sampling budgets B. **(a)** Worst-group expected rewards under the 3B proxy (left) and 7B Judge (right). Greedy-BoN consistently achieves the highest worst-group proxy reward, while BGP performs best in terms of worst-group evaluation reward. **(b)** Between-group reward disparity, measured by the gap between the highest and lowest group rewards (left) and the standard deviation of group rewards (right). BGP achieves the smallest reward disparity across groups among all algorithms.

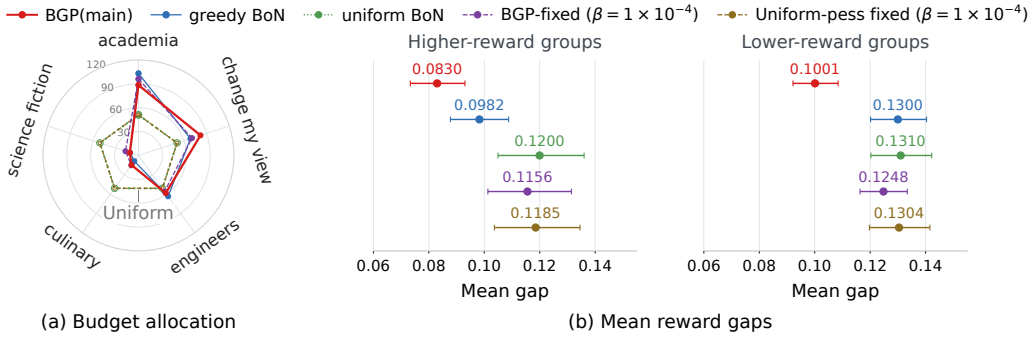

Figure 2: **(a)** Budget allocated to each group at $B = 255$. **(b)** Reward gaps within the higher- and lower-reward group subsets, averaged over 200 paired trials at $B = 255$. `askacademia` and `changemyview` form the higher-reward subset, and the remaining three domains form the lower-reward subset, determined by Judge. BGP achieves lowest mean group gaps in both subsets.

Do worst-group gains extend to fairness across groups? Figure 2(b) shows that BGP achieves the lowest mean between-group reward gap and standard deviation, extending its advantage beyond worst-group performance to overall reward balance. Since Judge-oracle BoN reveals substantial differences in attainable reward levels across groups, as shown in Figure 3. We therefore divide the domains into higher- and lower-reward subsets and examine disparities within each subset. As shown in Figure 2(b), BGP achieves the smallest mean within-trial reward gap in both subsets, indicating more balanced performance among domains with similar attainable reward levels.

How does BGP adapt its budget allocation across groups? Compared with Greedy-BoN, BGP adjusts sampling priorities using pessimistic reward estimates that account for group-specific reward errors $\bar{\varepsilon}_i$ (Figure 2(a)). This is important because groups with similar proxy rewards can have substantially different error estimates, as shown in Table 1. BGP's allocation criterion therefore considers both reward levels and their reliability when directing additional samples.

6 Conclusion

We studied the inference-time alignment for multiple groups sharing a finite sampling budget, and formulated it as a maximin allocation problem. Under exact rewards, the regret of Greedy-BoN vanishes to 0 as budget increases, but it can have non-vanishing regrets when proxy reward models are imperfect. Hence we develop a novel BGP algorithm that greedily allocates batches by a pessimistic index. Our experiments on five groups support our theoretical claims. We end this section with a few future directions. First, BGP requires an upper bound value on each reward model's error, so one interesting direction would be to design an algorithm that estimates the error online. Other directions include extending the allocation problem to distributions over prompts per group, or to other fairness objectives such as Nash welfare.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing, find related works, implement our algorithms, and generate proof ideas. We have not used generative AI tools for generating synthetic data. We verified LLM-generated codes and wrote formal math proofs on our own. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide anonymized source code for our proposed algorithms, baselines, and reward-model training pipeline at anonymized repository link. The accompanying README files include all of essential execution commands. The proof details of the omitted analysis are in the appendix, from Section A to Section D. The experimental setup, model configurations, hyperparameter settings, random seeds, and evaluation protocols are described in Section 5.1 and Section E.

REFERENCES

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3580–3602. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/beirami25a.html.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. 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. 6116–6135. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chakraborty24b.html.

Quan Ze Chen, Kevin Feng, Chan Young Park, and Amy X Zhang. SPICA: Retrieving scenarios for pluralistic in-context alignment. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 748–765, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.41. URL https://aclanthology.org/2025.findings-acl.41/.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.

Mehul Damani, Idan Shenfeld, Andi Peng, Andreea Bobu, and Jacob Andreas. Learning how hard to think: Input-adaptive allocation of LM computation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6qUUgw9bAZ.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Lin Gui, Cristina Gârbacea, and Victor Veitch. Bonbon alignment for large language models and the sweetness of best-of-n sampling. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 2851–2885. Curran Associates, Inc., 2024. doi: 10.52202/079017-0094. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/056521a35eacd9d2127b66a7d3c499c5-Paper-Conference.pdf.

Hadi Hosseini, Debmalya Mandal, and Duohan Zhang. Inference-time nash alignment. *arXiv preprint arXiv:2609.08082*, 2026.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. *arXiv preprint arXiv:2503.21878*, 2025.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

Yuu Jinnai, Tetsuro Morimura, Kaito Ariu, and Kenshi Abe. Regularized best-of-n sampling with minimum Bayes risk objective for language model alignment. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9321–9347, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.472. URL https://aclanthology.org/2025.naacl-long.472/.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio Calmon. Inference-time reward hacking in large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 61720–61760. Curran Associates, Inc., 2025. doi: 10.52202/085713-2064. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/590a0cc0306c1c63e2d66a51a407718f-Paper-Conference.pdf.

Yafu Li, Xuyang Hu, Xiaoye Qu, Linjie Li, and Yu Cheng. Test-time preference optimization: On-the-fly alignment via iterative textual feedback. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34630–34673. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25ac.html.

Baijiong Lin, Weisen Jiang, Yuancheng Xu, Hao Chen, and Ying-Cong Chen. PARM: Multi-objective test-time alignment via preference-aware autoregressive reward model. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37874–37888. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/lin25h.html.

Tianlin Liu, Shangmin Guo, Leonardo Bianco, Daniele Calandriello, Quentin Berthet, Felipe Llinares-López, Jessica Hoffmann, Lucas Dixon, Michal Valko, and Mathieu Blondel. Decoding-time realignment of language models. 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. 31015–31031. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/liu24r.html.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, et al. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2024.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems*, 37:52516–52544, 2024.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36:71095–71134, 2023.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free rlhf. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 37100–37137. Curran Associates, Inc., 2024. doi: 10.52202/079017-1171. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4147dfaa46cd7e20a2aecb91097ae8cc-Paper-Conference.pdf.

Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hannaneh Hajishirzi, Noah A. Smith, and Simon S. Du. Decoding-time language model alignment with multiple objectives. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 48875–48920. Curran Associates, Inc., 2024. doi: 10.52202/079017-1549. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/57c89126d60c209f48d0e6395c766bb3-Paper-Conference.pdf.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 10131–10165, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1b623663fd9b874366f3ce019fdfdd44-Paper-Conference.pdf.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic 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. 46280–46302. PMLR, 21–27 Jul 2024.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. Helpsteer2-preference: Complementing ratings with preferences. *arXiv preprint arXiv:2410.01257*, 2024.

Yuancheng Xu, Udari Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumitra Ganesh. Genarm: Reward guided generation with autoregressive reward model for test-time alignment. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 73970–73990, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b7a55b1fa8f2500299e1af3562a1191d-Paper-Conference.pdf.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*, 2024.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Nan Jiang, and Tong Zhang. A theoretical analysis of nash learning from human feedback under general kl-regularized preference. *arXiv preprint arXiv:2402.07314*, 4(5):10, 2024.

Yuheng Zhang, Dian Yu, Baolin Peng, Linfeng Song, Ye Tian, Mingyue Huo, Nan Jiang, Haitao Mi, and Dong Yu. Iterative nash policy optimization: Aligning llms with general preferences via no-regret learning. *arXiv preprint arXiv:2407.00617*, 2024.

Runlong Zhou, Maryam Fazel, and Simon S Du. Extragradient preference optimization (egpo): Beyond last-iterate convergence for nash learning from human feedback. *arXiv preprint arXiv:2503.08942*, 2025.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10586–10613, 2024.

Bowen Zuo and Yinglun Zhu. Strategic scaling of test-time compute: A bandit learning approach. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0mNnINd2z5.

Appendix

CONTENTS

A Omitted Analysis under the Exact Reward Setting **14**

A.1 Omitted Proofs in Section 3 14

A.2 Adaptive Benchmark 16

A.3 An Assumption-free Regret Bound 17

B Omitted Proof of Theorem 2 **18**

C Missing Algorithms **18**

D Omitted Analysis under the Reward Mismatch Setting **19**

D.1 Some Technical Lemmas 19

D.2 Omitted Analysis of Theorem 3 24

D.3 Choosing the Regularization 30

E Further Details of Experimental Setting **31**

F Additional Experimental Results **31**

A Omitted Analysis under the Exact Reward Setting

Throughout the appendices, we suppress the dependence on the prompt x_i and write $r_i^\star(y)$ and $\hat{r}_i(y)$ for $r_i^\star(x_i, y)$ and $\hat{r}_i(x_i, y)$.

A.1 Omitted Proofs in Section 3

In this section, we introduce a few technical lemmas and show the formal proof for the regret bound of Algorithm 1. The first lemma is the direct consequence of Assumption 1, and it links to the tail probability of $\widetilde{V}_i$.

Lemma 2. *Under Assumption 1, for every $i \in [k]$ and $n \geq 1$, we have*

$$\mathbb{P}[\widetilde{V}_i(n) < \overline{R}_i - u] \leq \begin{cases} e^{-c_i n u^{\alpha_i}}, & u \in (0, u_0], \\ e^{-c_i n u_0^{\alpha_i}}, & u \in (u_0, \overline{R}_i], \end{cases}$$

and

$$\overline{R}_i - V_i(n) \;\leq\; \Gamma(1+\gamma_i)\,(c_i n)^{-\gamma_i} + R_{\max}\, e^{-c_i n u_0^{\alpha_i}}.$$

Proof. For $u \in (0, u_0]$, using independence and $c_i u^{\alpha_i} \leq c_i u_0^{\alpha_i} \leq 1$,

$$\begin{aligned} \mathbb{P}[\widetilde{V}_i(n) < \overline{R}_i - u] &= \mathbb{P}_{y_1,\ldots,y_n \sim \pi_{i,\mathrm{ref}}}[\max_{j \leq n} r_i^\star(y_j) < \overline{R}_i - u] \\ &= (\mathbb{P}_{y \sim \pi_{i,\mathrm{ref}}}[r_i^\star(y) < \overline{R}_i - u])^n \\ &\leq (1 - c_i u^{\alpha_i})^n \\ &\leq e^{-c_i n u^{\alpha_i}}. \end{aligned}$$

For $u > u_0$, by monotonicity we have $\mathbb{P}[\widetilde{V}_i(n) < \overline{R}_i - u] \leq \mathbb{P}[\widetilde{V}_i(n) < \overline{R}_i - u_0] \leq e^{-c_i n u_0^{\alpha_i}}$. For the second bound, since $0 \leq \overline{R}_i - \widetilde{V}_i(n) \leq \overline{R}_i$ almost surely,

$$\begin{aligned} \overline{R}_i - V_i(n) &= \mathbb{E}[\overline{R}_i - \widetilde{V}_i(n)] \\ &= \int_0^{\overline{R}_i} \mathbb{P}[\overline{R}_i - \widetilde{V}_i(n) > u]\, du \\ &\leq \int_0^{\infty} e^{-c_i n u^{\alpha_i}}\, du + R_{\max} e^{-c_i n u_0^{\alpha_i}} \\ &= \Gamma(1+\gamma_i)(c_i n)^{-\gamma_i} + R_{\max} e^{-c_i n u_0^{\alpha_i}}. \end{aligned}$$

□

The second lemma combines Lemma 2 with a union bound over the k groups to bound the gap between essential supremum and rewards under a fixed allocation.

Lemma 3. *Under Assumption 1, let $m_1, \ldots, m_k$ be positive integers satisfying*

$$m_i \;\geq\; \frac{1 + \log k}{c_i\, u_0^{\alpha_i}} \qquad \textit{for all } i \in [k], \tag{5}$$

and write $D_i(m_i) := \overline{R}_i - \widetilde{V}_i(m_i)$. Then

$$\mathbb{E}\Big[\max_{i \in [k]} D_i(m_i)\Big] \leq 2 \max_{i \in [k]} \Big(\frac{1 + \log k}{c_i m_i}\Big)^{\gamma_i} + R_{\max} \sum_{i=1}^{k} e^{-c_i m_i u_0^{\alpha_i}}.$$

Proof. Write $D_i := D_i(m_i)$ and let $u^\star := \max_i \left(\frac{1+\log k}{c_i m_i}\right)^{\gamma_i}$; Equation (5) gives $u^\star \le u_0$. Since $0 \le \max_i D_i \le R_{\max}$, by Lemma 2 and a union bound,

$$\begin{aligned}\mathbb{E}[\max_i D_i] &= \int_0^{u^\star} \mathbb{P}[\max_i D_i > u]du + \int_{u^\star}^{u_0} \mathbb{P}[\max_i D_i > u]du + \int_{u_0}^{R_{\max}} \mathbb{P}[\max_i D_i > u]du \\ &\le u^\star + \sum_i \int_{u^\star}^{u_0} \mathbb{P}[D_i > u]du + \sum_i \int_{u_0}^{R_{\max}} \mathbb{P}[D_i > u]du \\ &\le u^\star + \sum_i \int_{u^\star}^{u_0} e^{-c_i m_i u^{\alpha_i}} du + R_{\max} \sum_i e^{-c_i m_i u_0^{\alpha_i}}.\end{aligned}$$

For the middle term, let $s := 1 + \log k$, $\lambda_i := c_i m_i$, and $s_i := \lambda_i (u^\star)^{\alpha_i}$, so that $s_i \ge s$ by the definition of $u^\star$. Substituting $v = \lambda_i u^{\alpha_i}$,

$$\begin{aligned}\int_{u^\star}^{u_0} e^{-\lambda_i u^{\alpha_i}} du &\le \gamma_i \lambda_i^{-\gamma_i} \int_{s_i}^{\infty} v^{\gamma_i - 1} e^{-v} dv \\ &\le \gamma_i \lambda_i^{-\gamma_i} s^{\gamma_i - 1} e^{-s},\end{aligned}$$

where the last line uses $\gamma_i \le 1$, so that $v^{\gamma_i - 1} \le s^{\gamma_i - 1}$ on $[s_i, \infty) \subseteq [s, \infty)$, and $\int_s^\infty e^{-v} dv = e^{-s}$. Summing over i and using $e^{-s} = 1/(ek)$,

$$\sum_{i=1}^k \gamma_i \lambda_i^{-\gamma_i} s^{\gamma_i - 1} e^{-s} \le k \cdot \max_i \left[\left(\frac{s}{\lambda_i}\right)^{\gamma_i} \cdot \frac{\gamma_i}{s} \right] \cdot \frac{1}{ek} \le \frac{1}{e} u^\star \le u^\star,$$

where the second inequality uses $\gamma_i \le 1 \le s$. Combining the displays proves the lemma. □

Lemma 1 (Ex-post optimality of Algorithm 1)**.** *Let $\widetilde{V}_1, \dots, \widetilde{V}_k : \mathbb{N}^+ \to \mathbb{R}^+$ be arbitrary non-decreasing sequences and let $\hat{\mathbf{n}} = (\hat{n}_1, \dots, \hat{n}_k)$ with $\sum_{i=1}^k \hat{n}_i = B$ be the counts produced by Algorithm 1. Then*

$$\min_{i \in [k]} \widetilde{V}_i(\hat{n}_i) = \max_{\sum_i m_i = B} \min_{i \in [k]} \widetilde{V}_i(m_i).$$

Proof. We say an allocation $\mathbf{o} = (o_1, \dots, o_k) \in (\mathbb{N}^+)^k$ feasible if $\sum_i o_i = B$. We prove our claim by contradiction. Suppose that some feasible $\mathbf{o}$ satisfies $\min_i \widetilde{V}_i(o_i) > \min_i \widetilde{V}_i(\hat{n}_i)$, and let $i_0 \in \arg\min_i \widetilde{V}_i(\hat{n}_i)$. Then $\widetilde{V}_i(o_i) > \widetilde{V}_{i_0}(\hat{n}_{i_0})$ for all i; in particular $\widetilde{V}_{i_0}(o_{i_0}) > \widetilde{V}_{i_0}(\hat{n}_{i_0})$, so $o_{i_0} > \hat{n}_{i_0}$ by monotonicity. Since $\sum_i \hat{n}_i = \sum_i o_i$, there exists a group j with $\hat{n}_j \ge o_j + 1 \ge 2$. Consider the round at which group j was incremented from $\hat{n}_j - 1$ to $\hat{n}_j$, and let $\hat{\mathbf{n}}^{\text{past}}$ be the counts at the start of that round, so that $\hat{n}_j^{\text{past}} = \hat{n}_j - 1$ and $\hat{n}_i^{\text{past}} \le \hat{n}_i$ for all i. By the greedy rule and monotonicity,

$$\widetilde{V}_j(\hat{n}_j - 1) = \widetilde{V}_j(\hat{n}_j^{\text{past}}) \le \widetilde{V}_i(\hat{n}_i^{\text{past}}) \le \widetilde{V}_i(\hat{n}_i), \quad \forall i \in [k].$$

Since $\hat{n}_j - 1 \ge o_j$, monotonicity also gives $\widetilde{V}_j(\hat{n}_j - 1) \ge \widetilde{V}_j(o_j)$. Taking $i = i_0$ above,

$$\widetilde{V}_{i_0}(\hat{n}_{i_0}) \ge \widetilde{V}_j(o_j) \ge \min_i \widetilde{V}_i(o_i) > \min_i \widetilde{V}_i(\hat{n}_i) = \widetilde{V}_{i_0}(\hat{n}_{i_0}),$$

a contradiction. □

Theorem 1. *Under the exact reward setting $\hat{r}_i = r_i^\star, \forall i \in [k]$ and Assumption 1, we have*

$$\mathcal{R}_B \le 2 \max_i \left(\frac{1 + \log k}{c_i \lfloor B/k \rfloor} \right)^{\gamma_i} + R_{\max} \sum_i \exp\left(-c_i \lfloor B/k \rfloor u_0^{\alpha_i}\right),$$

assuming that $\left\lfloor \frac{B}{k} \right\rfloor \ge \frac{1+\log k}{c_i u_0^{\alpha_i}}, \forall i \in [k]$.

Proof. By Remark 1, $\mathcal{R}_B \le \mathcal{R}_B^{\text{ep}}$, so it suffices to bound $\mathcal{R}_B^{\text{ep}}$. Since $\hat{y}_i$ is the best response among the draws of group i, we have $r_i^\star(\hat{y}_i) = \widetilde{V}_i(\hat{n}_i)$, and hence $\mathcal{R}_B^{\text{ep}} = V^\star - \mathbb{E}[\min_i \widetilde{V}_i(\hat{n}_i)]$. Fix a uniform allocation $\mathbf{m} = (m_1, \dots, m_k)$ with $m_i \in \{\lfloor B/k \rfloor, \lceil B/k \rceil\}$ and $\sum_i m_i = B$, and write

$R_{\min} := \min_i \overline{R}_i$ and $D_i(m_i) := \overline{R}_i - \widetilde{V}_i(m_i)$ as in Lemma 3. Since $\mathbf{m}$ is feasible, Lemma 1 gives, on every sample path,

$$\min_i \widetilde{V}_i(\hat{n}_i) \geq \min_i \widetilde{V}_i(m_i) = \min_i [\overline{R}_i - D_i(m_i)] \geq R_{\min} - \max_i D_i(m_i).$$

Since $V_i(n) \leq \overline{R}_i$ for all n, we also have $V^\star \leq R_{\min}$. Therefore,

$$\mathcal{R}_B^{\mathrm{ep}} = V^\star - \mathbb{E}[\min_i \widetilde{V}_i(\hat{n}_i)] \leq V^\star - R_{\min} + \mathbb{E}[\max_i D_i(m_i)] \leq \mathbb{E}[\max_i D_i(m_i)].$$

Since $m_i \geq \lfloor B/k \rfloor \geq (1+\log k)/(c_i u_0^{\alpha_i})$ for all i, Lemma 3 applies and gives

$$\begin{aligned}\mathcal{R}_B \leq \mathcal{R}_B^{\mathrm{ep}} &\leq 2 \max_i \left(\frac{1+\log k}{c_i m_i}\right)^{\gamma_i} + R_{\max} \sum_i e^{-c_i m_i u_0^{\alpha_i}} \\ &\leq 2 \max_i \left(\frac{1+\log k}{c_i \lfloor B/k \rfloor}\right)^{\gamma_i} + R_{\max} \sum_i e^{-c_i \lfloor B/k \rfloor u_0^{\alpha_i}},\end{aligned}$$

which completes the proof. □

A.2 Adaptive Benchmark

The benchmark $V^\star$ in Definition 1 is the value of the best *fixed* allocation, whereas Algorithm 1 allocates adaptively, so $\mathcal{R}_B$ can be negative. In this section, we define a benchmark given by the best adaptive policy and show that Theorem 1 holds verbatim against it.

An adaptive allocation policy π selects, at each round $t \in [B]$, a group i_t as a (possibly randomized) function of the past selections and observed rewards, and after B rounds returns for each group the best response among its draws. Let Π_B denote the class of adaptive allocation policies with budget B, and for $\pi \in \Pi_B$ let $\hat{n}_i^\pi$ be the number of draws group i receives, so that the returned response $\hat{y}_i^\pi$ satisfies $r_i^\star(\hat{y}_i^\pi) = \widetilde{V}_i(\hat{n}_i^\pi)$. We define the adaptive benchmark and the corresponding regret as

$$V_{\mathrm{ad}}^\star := \sup_{\pi \in \Pi_B} \min_{i \in [k]} \mathbb{E}\big[\widetilde{V}_i(\hat{n}_i^\pi)\big], \qquad \mathcal{R}_B^{\mathrm{ad}} := V_{\mathrm{ad}}^\star - \min_{i \in [k]} \mathbb{E}_{\hat{y}_i}\big[r_i^\star(\hat{y}_i)\big].$$

Since every fixed allocation is an adaptive policy, $V^\star \leq V_{\mathrm{ad}}^\star$ and hence $\mathcal{R}_B \leq \mathcal{R}_B^{\mathrm{ad}}$; moreover, $\mathcal{R}_B^{\mathrm{ad}} \geq 0$ for every policy in Π_B by definition. The following proposition shows that the Greedy BoN algorithm achieves the same adaptive regret bound.

Proposition 1. *Write $R_{\min} := \min_i \overline{R}_i$. Under Assumption 1,*

$$V^\star \leq V_{\mathrm{ad}}^\star \leq R_{\min} \leq V^\star + \max_{i \in [k]} \Big\{ \Gamma(1+\gamma_i) \big(c_i \lfloor B/k \rfloor\big)^{-\gamma_i} + R_{\max}\, e^{-c_i \lfloor B/k \rfloor u_0^{\alpha_i}} \Big\}.$$

Consequently, under the conditions of Theorem 1, Algorithm 1 satisfies

$$\mathcal{R}_B^{\mathrm{ad}} \leq 2 \max_i \left(\frac{1+\log k}{c_i \lfloor B/k \rfloor}\right)^{\gamma_i} + R_{\max} \sum_i e^{-c_i \lfloor B/k \rfloor u_0^{\alpha_i}}.$$

Proof. The first inequality was noted above. For the second, we have $\widetilde{V}_i(\hat{n}_i^\pi) \leq \overline{R}_i$ almost surely for every π and i, so $\min_i \mathbb{E}[\widetilde{V}_i(\hat{n}_i^\pi)] \leq R_{\min}$; taking the supremum over π gives the result. For the third, let $\mathbf{m}$ be a uniform allocation with $m_i \in \{\lfloor B/k \rfloor, \lceil B/k \rceil\}$ and $\sum_i m_i = B$. Since $\mathbf{m}$ is feasible, $V^\star \geq \min_i V_i(m_i)$, and therefore

$$R_{\min} - V^\star \leq \min_i \overline{R}_i - \min_i V_i(m_i) \leq \max_i \big(\overline{R}_i - V_i(m_i)\big) \leq \max_i \big(\overline{R}_i - V_i(\lfloor B/k \rfloor)\big),$$

where the last step uses that V_i is non-decreasing and $m_i \geq \lfloor B/k \rfloor$. Lemma 2 with $n = \lfloor B/k \rfloor$ completes the proof of the first statement.

Then we prove the regret upper bound. Recall that $r_i^\star(\hat{y}_i) = \widetilde{V}_i(\hat{n}_i)$ for Algorithm 1, and write $D_i(m_i) := \overline{R}_i - \widetilde{V}_i(m_i)$ for the uniform allocation $\mathbf{m}$ above. Since $\mathbf{m}$ is feasible, Lemma 1 gives, on every sample path,

$$\min_i \widetilde{V}_i(\hat{n}_i) \geq \min_i \widetilde{V}_i(m_i) = \min_i \big[\overline{R}_i - D_i(m_i)\big] \geq R_{\min} - \max_i D_i(m_i).$$

Combining this with $V^\star_{\mathrm{ad}} \le R_{\min}$ and Jensen's inequality,

$$\mathcal{R}^{\mathrm{ad}}_B = V^\star_{\mathrm{ad}} - \min_i \mathbb{E}[\widetilde{V}_i(\hat{n}_i)] \le V^\star_{\mathrm{ad}} - \mathbb{E}[\min_i \widetilde{V}_i(\hat{n}_i)] \le V^\star_{\mathrm{ad}} - R_{\min} + \mathbb{E}[\max_i D_i(m_i)] \le \mathbb{E}[\max_i D_i(m_i)].$$

Since $m_i \ge \lfloor B/k \rfloor \ge (1+\log k)/(c_i u_0^{\alpha_i})$ for all i, Lemma 3 applies and gives

$$\begin{aligned}\mathbb{E}[\max_i D_i(m_i)] &\le 2 \max_i \left(\frac{1+\log k}{c_i m_i}\right)^{\gamma_i} + R_{\max} \sum_i e^{-c_i m_i u_0^{\alpha_i}} \\ &\le 2 \max_i \left(\frac{1+\log k}{c_i \lfloor B/k \rfloor}\right)^{\gamma_i} + R_{\max} \sum_i e^{-c_i \lfloor B/k \rfloor u_0^{\alpha_i}},\end{aligned}$$

which completes the proof. □

A.3 An Assumption-free Regret Bound

In this section, we show that with no reward distribution assumption, the regret of Algorithm 1 converges to 0 as $B \to \infty$.

First, for $u > 0$, we define the tail probability

$$q_i(u) := \mathbb{P}_{y \sim \pi_{i,\mathrm{ref}}}\left[r^\star_i(y) \ge \overline{R}_i - u\right] = 1 - F_i(\overline{R}_i - u),$$

which is positive for every $u > 0$ by the definition of $\overline{R}_i$. Let $R_{\min} = \min_i \overline{R}_i$.

Proposition 2. *Let* $\mathbf{m}$ *be the uniform allocation with* $m_i \in \{\lfloor B/k \rfloor, \lceil B/k \rceil\}$ *and* $\sum_i m_i = B$*, and we define the value*

$$A_B := \inf_{u>0} \left\{ u + R_{\max} \sum_{i=1}^{k} e^{-m_i q_i(u)} \right\}.$$

The adaptive regret of Algorithm 1 satisfies

$$\mathcal{R}^{\mathrm{ad}}_B \le A_B,$$

and $A_B \to 0$ *as* $B \to \infty$*. Therefore,* $\mathcal{R}^{\mathrm{ad}}_B \to 0$ *as* $B \to \infty$*.*

Proof. For simplicity, we write $D_i := \overline{R}_i - \widetilde{V}_i(m_i)$. For every $u > 0$, we have

$$\mathbb{P}[D_i > u] = \mathbb{P}\left[\widetilde{V}_i(m_i) < \overline{R}_i - u\right] = \left(1 - q_i(u)\right)^{m_i} \le e^{-m_i q_i(u)}$$

where the last inequality follows since $(1-x)^n \le e^{-nx}$ for $x \in [0,1]$ and positive integer n. Since $0 \le \max_i D_i \le R_{\max}$, a union bound gives, for every $u > 0$,

$$\mathbb{E}[\max_i D_i] \le u + R_{\max}\, \mathbb{P}[\max_i D_i > u] \le u + R_{\max} \sum_{i=1}^{k} e^{-m_i q_i(u)}.$$

Now we take the inf over $u > 0$, yielding $\mathbb{E}[\max_i D_i] \le A_B$. Lemma 1 gives, on every sample path,

$$\min_i \widetilde{V}_i(\hat{n}_i) \ge \min_i \widetilde{V}_i(m_i) = \min_i(\overline{R}_i - D_i) \ge R_{\min} - \max_i D_i,$$

so $\mathbb{E}[\min_i \widetilde{V}_i(\hat{n}_i)] \ge R_{\min} - A_B$. On the other hand, since $V^\star_{\mathrm{ad}} \le R_{\min}$ and $\min_i \mathbb{E}[\widetilde{V}_i(\hat{n}_i)] \ge \mathbb{E}[\min_i \widetilde{V}_i(\hat{n}_i)]$ by Jensen's inequality, we have

$$\mathcal{R}^{\mathrm{ad}}_B = V^\star_{\mathrm{ad}} - \min_i \mathbb{E}[\widetilde{V}_i(\hat{n}_i)] \le R_{\min} - \mathbb{E}[\min_i \widetilde{V}_i(\hat{n}_i)] \le A_B.$$

It remains to show that $A_B \to 0$ as $B \to \infty$. Fix $\varepsilon > 0$ and take $u = \varepsilon$ in the definition of A_B. Since $q_i(\varepsilon) > 0$ for every i and $m_i \ge \lfloor B/k \rfloor \to \infty$ as $B \to \infty$,

$$\limsup_{B \to \infty} A_B \le \varepsilon + R_{\max} \lim_{B \to \infty} \sum_{i=1}^{k} e^{-m_i q_i(\varepsilon)} = \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we have $A_B \to 0$, and the proof is complete. □

B Omitted Proof of Theorem 2

Theorem 2. *There exists an instance with two groups, $\varepsilon_1^2 = p$ and $\varepsilon_2^2 = 0$, in which Algorithm 1 has regret*

$$\mathcal{R}_B = \frac{1}{2}\Big[1 - p^{B-1} - (1-p)^{B-1}\Big] \longrightarrow \frac{1}{2} \text{ as } B \to \infty.$$

Proof. Consider two groups and $p \in (0, 1)$:

Group	Response	Reference probability $\pi_{i,\text{ref}}$	True reward $r_i^\star$	Proxy reward $\hat{r}_i$
1	a	$1-p$	$1/2$	$1/2$
1	b	p	0	1
2	c	1	$3/4$	$3/4$

The reward-model errors satisfy

$$\varepsilon_1^2 = p, \qquad \varepsilon_2 = 0.$$

Now we run Algorithm 1 with the proxy scores:

- After initialization, if group 1 returned a, since $1/2 < 3/4$, the algorithm continues sampling a response for group 1.
- Once response b is sampled, it replaces a because it has a higher proxy score, and group 1's proxy value jumps to 1.
- The algorithm then allocates every remaining sample to group 2. Group 1 permanently returns response b with true reward zero.

According to the above procedure, group 1 returns a if and only if none of its first $B-1$ draws is b. Therefore,

$$\mathbb{E}\big[r_1^\star(\hat{y}_1)\big] = \tfrac{1}{2}(1-p)^{B-1}, \qquad \mathbb{E}\big[r_2^\star(\hat{y}_2)\big] = \tfrac{3}{4}.$$

Then we compute the deterministic benchmark, which allocates $B-1$ quota to group 1 and 1 quota to group 2:

$$V_1(n) = \frac{1}{2}(1-p^n), \quad V_2(n) = \frac{3}{4},$$

and

$$V^\star = \frac{1}{2}(1-p^{B-1}).$$

Consequently, the regret is computed as

$$\mathrm{R}_B = \frac{1}{2}\Big[1 - p^{B-1} - (1-p)^{B-1}\Big] \longrightarrow \frac{1}{2}.$$

□

C Missing Algorithms

The subroutines that are used in Algorithm 2 are formally presented in Algorithm 3.

Algorithm 3 Within-group subroutines

Require: the first n draws $y_{i,1}, \dots, y_{i,n}$ of group i with proxy scores $\hat{r}_{i,j} := \hat{r}_i(y_{i,j})$; error bound $\bar{\varepsilon}_i$; grid $\mathcal{G}$; tolerance η; confidence δ.

procedure SURROGATE(i, n, β) ▷ computes $\widehat{\lambda}_{i,\beta,n}$, $\widehat{\pi}_{i,\beta,n}$, and $\widehat{G}_i(\beta; n)$
  Sort the scores as $\hat{r}_{i,(1)} \geq \cdots \geq \hat{r}_{i,(n)}$ and let $S_m := \sum_{j \leq m} \hat{r}_{i,(j)}$ for $m \in [n]$.
  $m^\star \leftarrow \max\{m \in [n] : m\,\hat{r}_{i,(m)} > S_m - n\beta\}$ and $\widehat{\lambda} \leftarrow (S_{m^\star} - n\beta)/m^\star$. ▷ $\widehat{\lambda} = \widehat{\lambda}_{i,\beta,n}$
  $\widehat{w}_j \leftarrow (\hat{r}_{i,j} - \widehat{\lambda})_+/\beta$ for $j \in [n]$. ▷ $\frac{1}{n}\sum_j \widehat{w}_j = 1$
  $\widehat{G} \leftarrow \widehat{\lambda} + \frac{\beta}{2}\big(1 + \frac{1}{n}\sum_{j \leq n} \widehat{w}_j^2\big) - \bar{\varepsilon}_i - \frac{\bar{\varepsilon}_i^2}{2\beta}$. ▷ $\widehat{G} = \widehat{G}_i(\beta; n)$
  return $\big((\widehat{w}_j)_{j \leq n}, \widehat{G}\big)$.
end procedure
procedure INDEX(i, n) ▷ computes $L_i(n)$ and $\widehat{\beta}_i(n)$
  Obtain $\widehat{G}_i(\beta; n)$ from SURROGATE(i, n, β) for each $\beta \in \mathcal{B}(n)$.
  return $L_i(n) \leftarrow \max_{\beta \in \mathcal{B}(n)} \widehat{G}_i(\beta; n)$ and a maximizer $\widehat{\beta}_i(n)$.
end procedure
procedure SAMPLE(i, n, β) ▷ draws $\hat{y} \sim \widehat{\pi}_{i,\beta,n}$
  Obtain $(\widehat{w}_j)_{j \leq n}$ from SURROGATE(i, n, β) and draw $J \in [n]$ with $\mathbb{P}(J = j) = \widehat{w}_j/n$.
  return $y_{i,J}$.
end procedure

D OMITTED ANALYSIS UNDER THE REWARD MISMATCH SETTING

Before diving into technical analysis, we first introduce some notations that are heavily used in the analysis. We abbreviate $\pi_i := \pi_{i,\mathrm{ref}}$, and for a measurable function f we write $\pi_i f := \mathbb{E}_{y \sim \pi_{i,\mathrm{ref}}}[f(y)]$ and, given draws $y_{i,1}, \dots, y_{i,n}$, we write $\pi_{i,n} f := \frac{1}{n}\sum_{j \leq n} f(y_{i,j})$, where $\pi_{i,n}$ is the empirical distribution of the draws as in Section 4. Recall $M_\beta = 1 + R_{\max}/\beta$.

Weights and dual objectives. For $\beta > 0$ and $\lambda \in \mathbb{R}$, define the weight of a response y and its quadratic loss

$$w_{\beta,\lambda}(y) := \frac{(\hat{r}_i(y) - \lambda)_+}{\beta}, \qquad Z_{\beta,\lambda}(y) := \frac{\beta}{2}\, w_{\beta,\lambda}(y)^2.$$

Then the pessimistic policy (3) reads $\pi_{i,\beta} = \pi_{i,\mathrm{ref}} \cdot w_{\beta,\lambda_{i,\beta}}$, where $\lambda_{i,\beta}$ is the unique root of $\pi_i w_{\beta,\lambda} = 1$ (Lemma 4), and the empirical policy (4) puts mass $w_{\beta,\widehat{\lambda}_{i,\beta,n}}(y_{i,j})/n$ on the draw $y_{i,j}$, where $\widehat{\lambda}_{i,\beta,n}$ is the unique root of $\pi_{i,n} w_{\beta,\lambda} = 1$. Define the population and empirical dual objectives

$$F_{i,\beta}(\lambda) := \lambda + \frac{\beta}{2} + \pi_i Z_{\beta,\lambda}, \qquad \widehat{F}_{i,\beta,n}(\lambda) := \lambda + \frac{\beta}{2} + \pi_{i,n} Z_{\beta,\lambda}.$$

Both are convex, with $F'_{i,\beta}(\lambda) = 1 - \pi_i w_{\beta,\lambda}$ and $\widehat{F}'_{i,\beta,n}(\lambda) = 1 - \pi_{i,n} w_{\beta,\lambda}$, so $\lambda_{i,\beta} = \arg\min_\lambda F_{i,\beta}(\lambda)$ and $\widehat{\lambda}_{i,\beta,n} = \arg\min_\lambda \widehat{F}_{i,\beta,n}(\lambda)$ (Lemma 4). By Lemma 5(i), the objectives of Section 4 equal the minimal dual values,

$$\mathrm{Obj}_i(\beta) = F_{i,\beta}(\lambda_{i,\beta}), \qquad \widehat{\mathrm{Obj}}_i(\beta; n) = \widehat{F}_{i,\beta,n}(\widehat{\lambda}_{i,\beta,n}).$$

When i, β, or n are fixed and clear from context, we drop them from the notation.

D.1 SOME TECHNICAL LEMMAS

In the following, we provide a few important lemmas that will be heavily used in the analysis.

Lemma 4. *Fix i, $\beta > 0$, $n \geq 1$, and draws $y_{i,1}, \dots, y_{i,n}$. The following statements hold for $Q \in \{\pi_i, \pi_{i,n}\}$.*

(i) The map $\lambda \mapsto Q w_{\beta,\lambda}$ is continuous, nonincreasing, equal to 0 for $\lambda \geq R_{\max}$, tends to $+\infty$ as $\lambda \to -\infty$, and is strictly decreasing on $\{\lambda : Q w_{\beta,\lambda} > 0\}$. Hence $Q w_{\beta,\lambda} = 1$ has a unique root λ_Q, and λ_Q is the unique minimizer of the convex function $F_Q(\lambda) := \lambda + \beta/2 + Q Z_{\beta,\lambda}$.

(ii) $-\beta \le \lambda_Q < R_{\max}$.
(iii) For every $\lambda \ge -\beta$ *and* $y \in \mathcal{Y}$: $0 \le w_{\beta,\lambda}(y) \le M_\beta$ *and* $0 \le Z_{\beta,\lambda}(y) \le \beta M_\beta^2/2$.
(iv) For every y, $\lambda \mapsto w_{\beta,\lambda}(y)$ *is* $1/\beta$*-Lipschitz on* $\mathbb{R}$, *and* $\lambda \mapsto Z_{\beta,\lambda}(y)$ *is* M_β*-Lipschitz on* $[-\beta, R_{\max}]$.

Proof. (i) By definition $Qw_{\beta,\lambda} = \mathbb{E}_{y\sim Q}[\frac{(\hat{r}_i(y)-\lambda)_+}{\beta}]$. $(\hat{r}_i(y)-\lambda)_+$ is continuous and nonincreasing in λ; it equals 0 for $\lambda \ge R_{\max} \ge \hat{r}_i(y)$, and equals $\hat{r}_i(y) - \lambda \ge -\lambda$ for $\lambda \le 0$, so $Qw_{\beta,\lambda} \ge -\lambda/\beta \to \infty$ when $\lambda \to -\infty$. If $\lambda < \lambda'$ and $Qw_{\beta,\lambda'} > 0$ then $\mathbb{P}_{y\sim Q}(\hat{r}_i(y) > \lambda') > 0$ and

$$\begin{aligned} Qw_{\beta,\lambda} - Qw_{\beta,\lambda'} &= \mathbb{E}_{y\sim Q}\big[\frac{(\hat{r}_i(y)-\lambda)_+ - (\hat{r}_i(y)-\lambda')_+}{\beta}\big] \\ &\ge \mathbb{E}_{y\sim Q}\Big[\mathbf{1}\{\hat{r}_i(y) > \lambda'\}\,\frac{(\hat{r}_i(y)-\lambda) - (\hat{r}_i(y)-\lambda')}{\beta}\Big] \\ &\ge \frac{\lambda'-\lambda}{\beta}\mathbb{P}_{y\sim Q}[\hat{r}_i(y) > \lambda'] > 0 \end{aligned}$$

By continuity and the two limits, $Qw_{\beta,\lambda} = 1$ has a root, and by strict decrease on $\{\lambda : Qw_{\beta,\lambda} > 0\}$ it has exactly one; we call it λ_Q. Since $\frac{d}{d\lambda}(\hat{r}_i(y)-\lambda)_+^2 = -2(\hat{r}_i(y)-\lambda)_+$ and $\frac{d}{d\lambda}Z_{\beta,\lambda}(y) = -\frac{(\hat{r}_i(y)-\lambda)_+}{\beta} = -w_{\beta,\lambda}(y)$, so we compute $F_Q'(\lambda) = 1 - Qw_{\beta,\lambda}$, which is continuous, nondecreasing, negative to the left of the root and positive to the right; so F_Q is convex with the root as its unique minimizer.

(ii) We first prove the lower bound. If $\lambda_Q < 0$ then $(\hat{r}_i(y)-\lambda_Q)_+ = \hat{r}_i(y) - \lambda_Q$ pointwise, so $1 = Qw_{\beta,\lambda_Q} = (Q\hat{r}_i - \lambda_Q)/\beta$, i.e. $\lambda_Q = Q\hat{r}_i - \beta \ge -\beta$. Then we prove the upper bound. If $\lambda_Q \ge R_{\max}$, then $Qw_{\beta,\lambda_Q} = 0 \ne 1$ by (i); hence $\lambda_Q < R_{\max}$.

(iii) For $\lambda \ge -\beta$ and $\hat{r}_i(y) \le R_{\max}$, we have $(\hat{r}_i(y)-\lambda)_+ \le R_{\max} + \beta$. Hence $w_{\beta,\lambda} \le (R_{\max} + \beta)/\beta = M_\beta$ and $Z_{\beta,\lambda} = \frac{\beta}{2}w_{\beta,\lambda}^2 \le \beta M_\beta^2/2$.

(iv) $|\frac{(\hat{r}_i(y)-\lambda)_+}{\beta} - \frac{(\hat{r}_i(y)-\lambda')_+}{\beta}| \le \frac{|\lambda-\lambda'|}{\beta}$ gives the first claim. For the second claim, $\lambda \mapsto Z_{\beta,\lambda}(y)$ is continuously differentiable with $|\frac{d}{d\lambda}Z_{\beta,\lambda}(y)| = w_{\beta,\lambda}(y) \le M_\beta$ on $[-\beta, R_{\max}]$ by (iii), so the mean value theorem gives the claim. □

Lemma 5. *Fix* i, $\beta > 0$, $n \ge 1$, *draws* $y_{i,1}, \dots, y_{i,n}$, *and* $Q \in \{\pi_i, \pi_{i,n}\}$, *with* λ_Q *and* F_Q *as in Lemma 4; thus* $\lambda_Q = \lambda_{i,\beta}$ *and* $F_Q = F_{i,\beta}$ *for* $Q = \pi_i$, *and* $\lambda_Q = \widehat{\lambda}_{i,\beta,n}$ *and* $F_Q = \widehat{F}_{i,\beta,n}$ *for* $Q = \pi_{i,n}$. *For simplicity, we write* $w := w_{\beta,\lambda_Q}$ *and* $S := Qw^2$.

(i) (Pointwise identity) $w(y)\,\hat{r}_i(y) = \beta w(y)^2 + \lambda_Q w(y)$ *for every* y. *Consequently* $Q(w\hat{r}_i) = \beta S + \lambda_Q$ *and*

$$F_Q(\lambda_Q) = Q(w\hat{r}_i) - \frac{\beta}{2}(S-1).$$

In particular, $F_{i,\beta}(\lambda_{i,\beta}) = \mathbb{E}_{\pi_{i,\beta}}[\hat{r}_i] - \beta D_{\chi^2}(\pi_{i,\beta}\|\pi_i)$ *and* $\widehat{F}_{i,\beta,n}(\widehat{\lambda}_{i,\beta,n}) = \mathbb{E}_{\widehat{\pi}_{i,\beta,n}}[\hat{r}_i] - \beta D_{\chi^2}(\widehat{\pi}_{i,\beta,n}\|\pi_{i,n})$.

(ii) (Variational inequality) For every measurable $q : \mathcal{Y} \to [0,\infty)$ *with* $Qq = 1$ *and* $Qq^2 < \infty$,

$$Q(q\hat{r}_i) - \frac{\beta}{2}(Qq^2 - 1) \;\le\; F_Q(\lambda_Q),$$

with equality at $q = w$. *In particular,* $F_{i,\beta}(\lambda_{i,\beta}) \ge \mathbb{E}_\pi[\hat{r}_i] - \beta D_{\chi^2}(\pi\|\pi_i)$ *for every* $\pi \ll \pi_i$ *with* $\mathcal{C}_i^\pi < \infty$, *with equality at* $\pi = \pi_{i,\beta}$.
(iii) (Range) $1 \le S \le M_\beta$ *and* $0 \le F_Q(\lambda_Q) \le R_{\max}$.

Proof. (i) Fix $y \in \mathcal{Y}$. If $\hat{r}_i(y) > \lambda_Q$, then by definition $w(y) = (\hat{r}_i(y) - \lambda_Q)/\beta$, i.e., $\hat{r}_i(y) - \lambda_Q = \beta w(y)$. Hence

$$w(y)\,\hat{r}_i(y) = w(y)\big(\hat{r}_i(y) - \lambda_Q\big) + \lambda_Q w(y) = \beta w(y)^2 + \lambda_Q w(y).$$

If $\hat{r}_i(y) \le \lambda_Q$, then $w(y) = 0$ and both sides vanish. Applying Q to both sides and using $Qw = 1$ from Lemma 4(i),

$$Q(w\hat{r}_i) = \beta\, Qw^2 + \lambda_Q\, Qw = \beta S + \lambda_Q.$$

Finally, $Z_{\beta,\lambda_Q} = \frac{\beta}{2}w^2$ gives $QZ_{\beta,\lambda_Q} = \frac{\beta}{2}S$, so

$$F_Q(\lambda_Q) = \lambda_Q + \frac{\beta}{2} + \frac{\beta}{2}S = (\beta S + \lambda_Q) - \frac{\beta}{2}(S-1) = Q(w\hat{r}_i) - \frac{\beta}{2}(S-1).$$

In particular, take $Q = \pi_i$, so that $\lambda_Q = \lambda_{i,\beta}$, $F_Q(\lambda_Q) = F_{i,\beta}(\lambda_{i,\beta})$, and $w = w_{\beta,\lambda_{i,\beta}}$. By definition, $\pi_{i,\beta}(y) = \pi_{i,\mathrm{ref}}(y)\, w(y)$, i.e., $w = d\pi_{i,\beta}/d\pi_{i,\mathrm{ref}}$. For the first term, changing measure gives

$$\pi_i(w\hat{r}_i) = \mathbb{E}_{y\sim\pi_{i,\mathrm{ref}}}[w(y)\hat{r}_i(y)] = \mathbb{E}_{\pi_{i,\beta}}[\hat{r}_i].$$

For the second term, by the definition of D_{χ^2} and $\pi_i w = 1$,

$$\beta D_{\chi^2}(\pi_{i,\beta}\|\pi_i) = \frac{\beta}{2}\,\pi_i\big[(w-1)^2\big] = \frac{\beta}{2}\big(\pi_i w^2 - 2\,\pi_i w + 1\big) = \frac{\beta}{2}(S-1).$$

Substituting both into $F_Q(\lambda_Q) = Q(w\hat{r}_i) - \frac{\beta}{2}(S-1)$ gives $F_{i,\beta}(\lambda_{i,\beta}) = \mathbb{E}_{\pi_{i,\beta}}[\hat{r}_i] - \beta D_{\chi^2}(\pi_{i,\beta}\|\pi_i)$. The case $Q = \pi_{i,n}$ is identical: $\widehat{\pi}_{i,\beta,n}$ puts mass $w(y_{i,j})/n$ on $y_{i,j}$, so $w = d\widehat{\pi}_{i,\beta,n}/d\pi_{i,n}$, hence $\pi_{i,n}(w\hat{r}_i) = \mathbb{E}_{\widehat{\pi}_{i,\beta,n}}[\hat{r}_i]$ and $\beta D_{\chi^2}(\widehat{\pi}_{i,\beta,n}\|\pi_{i,n}) = \frac{\beta}{2}(S-1)$ by $\pi_{i,n}w = 1$.

(ii) We first show that for all $a \in \mathbb{R}$ and $t \geq 0$,

$$ta - \frac{\beta}{2}t^2 \leq \frac{a_+^2}{2\beta}.$$

If $a \leq 0$, the left side is ≤ 0 while the right side equals 0. If $a > 0$, $ta - \frac{\beta}{2}t^2 = -\frac{\beta}{2}(t - a/\beta)^2 + \frac{a^2}{2\beta} \leq \frac{a^2}{2\beta} = \frac{a_+^2}{2\beta}$. Fix y and apply this with $a = \hat{r}_i(y) - \lambda_Q$ and $t = q(y)$. Since $\frac{(\hat{r}_i(y)-\lambda_Q)_+^2}{2\beta} = \frac{\beta}{2}w(y)^2 = Z_{\beta,\lambda_Q}(y)$,

$$q(y)\hat{r}_i(y) - \lambda_Q q(y) - \frac{\beta}{2}q(y)^2 \leq Z_{\beta,\lambda_Q}(y).$$

Integrating under Q and using $Qq = 1$ gives $Q(q\hat{r}_i) - \lambda_Q - \frac{\beta}{2}Qq^2 \leq QZ_{\beta,\lambda_Q} = \frac{\beta}{2}S$. Adding $\lambda_Q + \frac{\beta}{2}$ to both sides,

$$Q(q\hat{r}_i) - \frac{\beta}{2}(Qq^2 - 1) \leq \lambda_Q + \frac{\beta}{2} + \frac{\beta}{2}S = F_Q(\lambda_Q),$$

where the last equality is the definition of F_Q in Lemma 4(i), since $QZ_{\beta,\lambda_Q} = \frac{\beta}{2}S$. The choice $q = w$ is admissible: $Qw = 1$ by Lemma 4(i), and $Qw^2 \leq M_\beta^2 < \infty$ since Lemma 4(ii)–(iii). For this choice the left side equals $F_Q(\lambda_Q)$ by (i), so equality holds. In particular, take $Q = \pi_{i,\mathrm{ref}}$ and, for $\pi \ll \pi_{i,\mathrm{ref}}$, let $q = d\pi/d\pi_{i,\mathrm{ref}}$. Then $Qq = 1$ since π is a probability measure, $Q(q\hat{r}_i) = \mathbb{E}_\pi[\hat{r}_i]$, and $Qq^2 = \mathbb{E}_\pi[q] = \mathcal{C}_i^\pi < \infty$, so q is admissible. Moreover $\frac{\beta}{2}(Qq^2 - 1) = \beta D_{\chi^2}(\pi\|\pi_{i,\mathrm{ref}})$, so the inequality reads $\mathbb{E}_\pi[\hat{r}_i] - \beta D_{\chi^2}(\pi\|\pi_{i,\mathrm{ref}}) \leq F_{i,\beta}(\lambda_{i,\beta})$. Equality holds at $\pi = \pi_{i,\beta}$, since $d\pi_{i,\beta}/d\pi_{i,\mathrm{ref}} = w_{\beta,\lambda_{i,\beta}}$ is exactly the case $q = w$.

(iii) We first prove the range of S. By Lemma 4(i), $Qw = 1$, so Jensen's inequality gives $S = Qw^2 \geq (Qw)^2 = 1$. By Lemma 4(ii), $\lambda_Q \in [-\beta, R_{\max})$, and Lemma 4(iii) gives $0 \leq w \leq M_\beta$; hence $w^2 \leq M_\beta w$ pointwise, and $S \leq M_\beta\, Qw = M_\beta$.

Then we prove the range of $F_Q(\lambda_Q)$. For the lower bound, the definition of F_Q and $QZ_{\beta,\lambda_Q} = \frac{\beta}{2}S$ give $F_Q(\lambda_Q) = \lambda_Q + \frac{\beta}{2}(1+S) \geq \lambda_Q + \beta \geq 0$, using $S \geq 1$ and $\lambda_Q \geq -\beta$. For the upper bound, (i) and $S \geq 1$ give $F_Q(\lambda_Q) = Q(w\hat{r}_i) - \frac{\beta}{2}(S-1) \leq Q(w\hat{r}_i)$, and since $w \geq 0$ and $\hat{r}_i \leq R_{\max}$, $Q(w\hat{r}_i) \leq R_{\max}\, Qw = R_{\max}$. □

Lemma 6 (Bernstein and Hoeffding (Boucheron et al., 2013)). *Let $X_1, \ldots, X_n$ be i.i.d. with values in $[0, b]$, mean μ and variance σ^2, and let $\bar{X}_n$ be their average. For every $t > 0$, each of the following holds with probability at least $1 - 2e^{-t}$:*

$$\textit{(Bernstein)} \quad |\bar{X}_n - \mu| \leq \sqrt{\frac{2\sigma^2 t}{n}} + \frac{2bt}{3n}, \qquad \textit{(Hoeffding)} \quad |\bar{X}_n - \mu| \leq b\sqrt{\frac{t}{2n}}.$$

Lemma 7. *For every i and $\beta > 0$, we have*

$$J_i(\pi_{i,\beta}) \geq G_i(\beta).$$

Furthermore, for every $\pi \ll \pi_{i,\mathrm{ref}}$ with $\mathcal{C}_i^\pi < \infty$, we have

$$G_i(\beta) \geq J_i(\pi) - 2\sqrt{\mathcal{C}_i^\pi}\,\bar{\varepsilon}_i - \frac{\beta \mathcal{C}_i^\pi}{2} - \frac{\bar{\varepsilon}_i^2}{2\beta}.$$

Proof. Write $w := w_{\beta,\lambda_{i,\beta}}$, so that $\pi_{i,\beta}(y) = \pi_{i,\mathrm{ref}}(y)\, w(y)$, and let $D := D_{\chi^2}(\pi_{i,\beta} \| \pi_{i,\mathrm{ref}})$. The definition of D_{χ^2} gives

$$\mathbb{E}_{y\sim\pi_{i,\mathrm{ref}}}[w(y)^2] = 1 + 2D.$$

By a change of measure, Cauchy–Schwarz, and $\sqrt{1+x} \leq 1 + \sqrt{x}$ for $x \geq 0$,

$$\begin{aligned}\mathbb{E}_{y\sim\pi_{i,\beta}}\left[|\hat{r}_i(y) - r_i^\star(y)|\right] &= \mathbb{E}_{y\sim\pi_{i,\mathrm{ref}}}\left[w(y)\,|\hat{r}_i(y) - r_i^\star(y)|\right] \\ &\leq \sqrt{\mathbb{E}_{y\sim\pi_{i,\mathrm{ref}}}[w(y)^2]}\;\varepsilon_i = \sqrt{1+2D}\;\varepsilon_i \leq \varepsilon_i + \varepsilon_i\sqrt{2D}.\end{aligned}$$

By AM–GM, $\varepsilon_i\sqrt{2D} = 2 \cdot \frac{\varepsilon_i}{\sqrt{2\beta}} \cdot \sqrt{\beta D} \leq \frac{\varepsilon_i^2}{2\beta} + \beta D$, hence

$$\mathbb{E}_{y\sim\pi_{i,\beta}}\left[|\hat{r}_i(y) - r_i^\star(y)|\right] \leq \varepsilon_i + \beta D + \frac{\varepsilon_i^2}{2\beta}.$$

Since $r_i^\star \geq \hat{r}_i - |\hat{r}_i - r_i^\star|$ pointwise,

$$\begin{aligned}J_i(\pi_{i,\beta}) &= \mathbb{E}_{\pi_{i,\beta}}[r_i^\star] \\ &\geq \mathbb{E}_{\pi_{i,\beta}}[\hat{r}_i - |\hat{r}_i - r_i^\star|] \\ &\geq \mathbb{E}_{\pi_{i,\beta}}[\hat{r}_i] - \beta D - \varepsilon_i - \frac{\varepsilon_i^2}{2\beta} \\ &= \mathrm{Obj}_i(\beta) - \varepsilon_i - \frac{\varepsilon_i^2}{2\beta} \\ &\geq G_i(\beta).\end{aligned}$$

The equality in the fourth line is Lemma 5(i). The last inequality holds because $\varepsilon_i \leq \bar{\varepsilon}_i$ by Assumption 2.

We now prove the second statement. Let $q := d\pi / d\pi_{i,\mathrm{ref}}$ and $C := \mathcal{C}_i^\pi$. By a change of measure, $C = \mathbb{E}_{y\sim\pi}[q(y)] = \mathbb{E}_{y\sim\pi_{i,\mathrm{ref}}}[q(y)^2]$, and by Jensen's inequality $C \geq (\mathbb{E}_{y\sim\pi_{i,\mathrm{ref}}}[q(y)])^2 = 1$. By Lemma 5(ii), and since $\beta D_{\chi^2}(\pi \| \pi_{i,\mathrm{ref}}) = \frac{\beta}{2}(C-1)$,

$$\mathrm{Obj}_i(\beta) \geq \mathbb{E}_{y\sim\pi}[\hat{r}_i(y)] - \frac{\beta}{2}(C-1).$$

By a change of measure and Cauchy–Schwarz,

$$\left|\mathbb{E}_{y\sim\pi}[\hat{r}_i(y) - r_i^\star(y)]\right| = \left|\mathbb{E}_{y\sim\pi_{i,\mathrm{ref}}}[q(y)(\hat{r}_i(y) - r_i^\star(y))]\right| \leq \sqrt{C}\,\varepsilon_i,$$

so $\mathbb{E}_{y\sim\pi}[\hat{r}_i(y)] \geq J_i(\pi) - \sqrt{C}\,\varepsilon_i \geq J_i(\pi) - \sqrt{C}\,\bar{\varepsilon}_i$ by Assumption 2. Therefore

$$\begin{aligned}G_i(\beta) &= \mathrm{Obj}_i(\beta) - \bar{\varepsilon}_i - \frac{\bar{\varepsilon}_i^2}{2\beta} \\ &\geq \mathbb{E}_{y\sim\pi}[\hat{r}_i(y)] - \frac{\beta}{2}(C-1) - \bar{\varepsilon}_i - \frac{\bar{\varepsilon}_i^2}{2\beta} \\ &\geq J_i(\pi) - \sqrt{C}\,\bar{\varepsilon}_i - \bar{\varepsilon}_i - \frac{\beta C}{2} + \frac{\beta}{2} - \frac{\bar{\varepsilon}_i^2}{2\beta} \\ &\geq J_i(\pi) - 2\sqrt{C}\,\bar{\varepsilon}_i - \frac{\beta C}{2} - \frac{\bar{\varepsilon}_i^2}{2\beta},\end{aligned}$$

where the last step uses $\bar{\varepsilon}_i \leq \sqrt{C}\,\bar{\varepsilon}_i$ (as $C \geq 1$) and drops $\frac{\beta}{2} \geq 0$. □

The next lemma is the exact analogue of Lemma 7 under the empirical measure.

Lemma 8. *Fix i, $\beta > 0$, $n \geq 1$ and the draws $y_{i,1}, \ldots, y_{i,n}$. Let*

$$\widehat{\varepsilon}_{i,n}^2 := \frac{1}{n} \sum_{j=1}^{n} (\hat{r}_i(y_{i,j}) - r_i^\star(y_{i,j}))^2$$

be the empirical L_2 error. Then, deterministically,

$$J_i(\widehat{\pi}_{i,\beta,n}) \geq \widehat{G}_i(\beta; n) - \sqrt{1 + \frac{R_{\max}}{\beta}}\, (\widehat{\varepsilon}_{i,n} - \overline{\varepsilon}_i)_+.$$

In particular, $J_i(\widehat{\pi}_{i,\beta,n}) \geq \widehat{G}_i(\beta; n)$ whenever $\widehat{\varepsilon}_{i,n} \leq \overline{\varepsilon}_i$.

Proof. Take $Q = \pi_{i,n}$ and write $\widehat{\lambda} := \widehat{\lambda}_{i,\beta,n}$, $\widehat{w} := w_{\beta,\widehat{\lambda}}$, and $S := Q\widehat{w}^2$. By Lemma 4(i), $Q\widehat{w} = 1$, and by Lemma 5(iii), $1 \leq S \leq M_\beta$. Since $\widehat{\pi}_{i,\beta,n}$ puts mass $\widehat{w}(y_{i,j})/n$ on the draw $y_{i,j}$, Cauchy–Schwarz under Q gives

$$\begin{aligned} J_i(\widehat{\pi}_{i,\beta,n}) &= \mathbb{E}_{y \sim \widehat{\pi}_{i,\beta,n}}[r_i^\star(y)] \\ &= \sum_{j=1}^{n} \frac{\widehat{w}(y_{i,j})}{n} r_i^\star(y_{i,j}) \\ &= Q(\widehat{w}\, r_i^\star) \\ &\geq Q(\widehat{w}\, \hat{r}_i) - \sqrt{Q\widehat{w}^2}\, \sqrt{Q(\hat{r}_i - r_i^\star)^2} \\ &= Q(\widehat{w}\, \hat{r}_i) - \sqrt{S}\, \widehat{\varepsilon}_{i,n}. \end{aligned}$$

By Lemma 5(i), $Q(\widehat{w}\, \hat{r}_i) = \widehat{\mathrm{Obj}}_i(\beta; n) + \frac{\beta}{2}(S-1)$. Since $\widehat{\varepsilon}_{i,n} \leq \overline{\varepsilon}_i + (\widehat{\varepsilon}_{i,n} - \overline{\varepsilon}_i)_+$ and $\sqrt{S} \geq 0$,

$$\begin{aligned} J_i(\widehat{\pi}_{i,\beta,n}) - \widehat{\mathrm{Obj}}_i(\beta; n) &\geq \frac{\beta}{2}(S-1) - \sqrt{S}\, \widehat{\varepsilon}_{i,n} \\ &\geq \left[\frac{\beta}{2}(S-1) - \sqrt{S}\, \overline{\varepsilon}_i\right] - \sqrt{S}\, (\widehat{\varepsilon}_{i,n} - \overline{\varepsilon}_i)_+. \end{aligned}$$

For the first term, $\sqrt{S} \leq 1 + \sqrt{S-1}$ and AM–GM give

$$\begin{aligned} \sqrt{S}\, \overline{\varepsilon}_i &\leq \overline{\varepsilon}_i + \sqrt{\beta(S-1)} \cdot \frac{\overline{\varepsilon}_i}{\sqrt{\beta}} \\ &\leq \overline{\varepsilon}_i + \frac{\beta}{2}(S-1) + \frac{\overline{\varepsilon}_i^2}{2\beta}, \end{aligned}$$

so the first term is at least $-\overline{\varepsilon}_i - \frac{\overline{\varepsilon}_i^2}{2\beta}$. Using $\sqrt{S} \leq \sqrt{M_\beta} = \sqrt{1 + R_{\max}/\beta}$ on the second term and the definition of $\widehat{G}_i$,

$$\begin{aligned} J_i(\widehat{\pi}_{i,\beta,n}) &\geq \widehat{\mathrm{Obj}}_i(\beta; n) - \overline{\varepsilon}_i - \frac{\overline{\varepsilon}_i^2}{2\beta} - \sqrt{M_\beta}\, (\widehat{\varepsilon}_{i,n} - \overline{\varepsilon}_i)_+ \\ &= \widehat{G}_i(\beta; n) - \sqrt{1 + \frac{R_{\max}}{\beta}}\, (\widehat{\varepsilon}_{i,n} - \overline{\varepsilon}_i)_+. \end{aligned}$$

□

The following lemma bounds the coverage $\mathcal{C}_i^{\pi_i^\star}$ with a constant.

Lemma 9 (Coverage of the best-of-n response)**.** *Fix a group i and write $\pi := \pi_{i,\mathrm{ref}}$, $r := r_i^\star(x_i, \cdot)$, $T := \{y : r(y) = \overline{R}_i\}$ and $p := \pi(T)$. Let $Y_1, \ldots, Y_n \sim \pi$ be i.i.d., let S be the smallest index attaining $\max_{s \leq n} r(Y_s)$, and let π_n be the law of Y_S. Then $\pi_n \ll \pi$, the density ratio $q := d\pi_n/d\pi$ satisfies $q \leq \min\{n, 1/p\}$ π-almost surely, and*

$$1 \leq \mathcal{C}_i^{\pi_n} \leq \min\{n, 1/p\}.$$

Proof. We show $\pi_n(A) \le c\,\pi(A)$ for every measurable $A \subseteq \mathcal{Y}$, with $c = n$ and, when $p > 0$, with $c = 1/p$; this gives $\pi_n \ll \pi$ and $q \le c\,\pi$-almost surely.

The bound $c = n$. Since Y_S is one of the draws, a union bound gives $\pi_n(A) = \mathbb{P}(Y_S \in A) \le \mathbb{P}(\exists\, s \le n :\ Y_s \in A) \le \sum_{s=1}^{n} \mathbb{P}(Y_s \in A) = n\,\pi(A)$.

The bound $c = 1/p$ when $p > 0$. Any draw in T attains the largest possible reward, so if some draw lies in T then S is the smallest index s with $Y_s \in T$. Since $\pi_n(A) = \pi_n(A \cap T) + \pi_n(A \setminus T)$, it suffices to treat $A \subseteq T$ and $A \subseteq \mathcal{Y} \setminus T$. For $A \subseteq T$, the event $\{Y_S \in A\}$ is the disjoint union over $s \le n$ of $\{Y_1, \dots, Y_{s-1} \notin T,\ Y_s \in A\}$, so by independence

$$\pi_n(A) = \sum_{s=1}^{n} (1-p)^{s-1}\,\pi(A) = \frac{1-(1-p)^n}{p}\,\pi(A) \le \frac{\pi(A)}{p}.$$

For $A \subseteq \mathcal{Y} \setminus T$, the event $\{Y_S \in A\}$ requires that no draw lies in T and that some draw lies in A, so

$$\pi_n(A) \le \sum_{s=1}^{n} \mathbb{P}\big(Y_s \in A,\ Y_{s'} \notin T\ \ \forall s' \ne s\big) = n(1-p)^{n-1}\,\pi(A) \le \frac{\pi(A)}{p},$$

where the last step uses $n(1-p)^{n-1} \le ne^{-p(n-1)} \le e^{p-1}/p \le 1/p$; the middle inequality is the maximum of $x \mapsto xe^{-p(x-1)}$ over $x > 0$, attained at $x = 1/p$.

Coverage. By definition, $\mathcal{C}_i^{\pi_n} = \mathbb{E}_{y\sim\pi_n}[q(y)] \le \sup q \le \min\{n, 1/p\}$, and $\mathcal{C}_i^{\pi_n} = \mathbb{E}_{y\sim\pi}[q(y)^2] \ge \big(\mathbb{E}_{y\sim\pi}[q(y)]\big)^2 = 1$ by Jensen's inequality. □

D.2 Omitted Analysis of Theorem 3

Theorem 4. *There is an event $\mathcal{E}_{\mathrm{obj}}$ with $\mathbb{P}(\mathcal{E}_{\mathrm{obj}}) \ge 1 - \delta/2$ on which, simultaneously for every $i \in [k], n \in [B]$, and $\beta \in \mathcal{G}$,*

$$|\widehat{G}_i(\beta; n) - G_i(\beta)| \le \Delta^{\mathrm{obj}}(n, \beta, \delta).$$

Proof. In the proof, we fix i, β, n, and omit the subscript for simplicity. For every fixed $\lambda \in \mathbb{R}$, $F_\beta(\lambda) = \lambda + \frac{\beta}{2} + \pi Z_\lambda$ and $\widehat{F}_{\beta,n}(\lambda) = \lambda + \frac{\beta}{2} + \pi_n Z_\lambda$. Since $\widehat{G}(\beta; n)$ and $G(\beta)$ subtract the same constant $\bar{\varepsilon} + \bar{\varepsilon}^2/(2\beta)$, we have $\widehat{G}(\beta; n) - G(\beta) = \widehat{F}_{\beta,n}(\widehat{\lambda}) - F_\beta(\lambda_\beta)$. Because $\widehat{\lambda}$ minimizes $\widehat{F}_{\beta,n}$ and λ_β minimizes F_β, we have

$$\widehat{G}(\beta; n) - G(\beta) \le \widehat{F}_{\beta,n}(\lambda_\beta) - F_\beta(\lambda_\beta) = (\pi_n - \pi) Z_{\lambda_\beta},$$

and

$$G(\beta) - \widehat{G}(\beta; n) \le F_\beta(\widehat{\lambda}) - \widehat{F}_{\beta,n}(\widehat{\lambda}) = (\pi - \pi_n) Z_{\widehat{\lambda}},$$

where $\pi Z_{\widehat{\lambda}}$ denotes the map $\lambda \mapsto \pi Z_\lambda$ evaluated at $\lambda = \widehat{\lambda}$. Hence

$$|\widehat{G}(\beta; n) - G(\beta)| \le \max\big\{|(\pi_n - \pi) Z_{\lambda_\beta}|,\ |(\pi_n - \pi) Z_{\widehat{\lambda}}|\big\}.$$

The first term is evaluated at the deterministic point λ_β; the second at the data-dependent point $\widehat{\lambda}$, which we handle with a deterministic grid.

Since $k, N_\beta, B \ge 1$ and $\delta < 1$, the argument of the logarithm exceeds $48 > e$, so $l > 1$. If $c := M_\beta l / n > 1$, the claim holds deterministically: by Lemma 5(iii), both $\mathrm{Obj}(\beta)$ and $\widehat{\mathrm{Obj}}(\beta; n)$ lie in $[0, R_{\max}]$, so

$$|\widehat{G}(\beta; n) - G(\beta)| = |\widehat{\mathrm{Obj}}(\beta; n) - \mathrm{Obj}(\beta)| \le R_{\max} \le 5R_{\max}\big[\sqrt{c} + c\big].$$

Hence we assume $c \le 1$ from now on; together with $l > 1$, this gives $M_\beta \le n/l \le n$. We consider two cases.

Case 1: $\beta < R_{\max}$. We first bound the term at λ_β. By Lemma 4(ii), $\lambda_\beta \in [-\beta, R_{\max}]$, and Lemma 4(iii) gives $0 \le w_{\lambda_\beta} \le M_\beta$ and $0 \le Z_{\lambda_\beta} \le \beta M_\beta^2/2$. Using $w_{\lambda_\beta}^4 \le M_\beta^3 w_{\lambda_\beta}$ pointwise and $\pi w_{\lambda_\beta} = 1$ from Lemma 4(i),

$$\mathrm{Var}_\pi[Z_{\lambda_\beta}] \le \pi Z_{\lambda_\beta}^2 = \frac{\beta^2}{4}\,\pi w_{\lambda_\beta}^4 \le \frac{\beta^2 M_\beta^3}{4}.$$

Since λ_β does not depend on the draws, $Z_{\lambda_\beta}(y_1), \dots, Z_{\lambda_\beta}(y_n)$ are i.i.d. in $[0, \beta M_\beta^2/2]$. Bernstein's inequality (Lemma 6) with $t = l$ gives an event A_1 with $\mathbb{P}(A_1) \ge 1 - 2e^{-l}$ on which

$$\begin{aligned}
|(\pi_n - \pi)Z_{\lambda_\beta}| &\le \sqrt{\frac{2l}{n} \cdot \frac{\beta^2 M_\beta^3}{4}} + \frac{2l}{3n} \cdot \frac{\beta M_\beta^2}{2} \\
&= \frac{\beta M_\beta}{\sqrt{2}} \sqrt{\frac{M_\beta l}{n}} + \frac{\beta M_\beta}{3} \cdot \frac{M_\beta l}{n} \\
&\le \sqrt{2} R_{\max} \sqrt{\frac{M_\beta l}{n}} + \frac{2R_{\max}}{3} \cdot \frac{M_\beta l}{n},
\end{aligned}$$

where the last step uses $\beta M_\beta = \beta + R_{\max} < 2R_{\max}$ in Case 1.

Next we construct a deterministic grid for the term at $\widehat{\lambda}$. Let $\Lambda_{\beta,n} \subset [-\beta, R_{\max}]$ be a grid containing both endpoints with mesh at most $h := \beta/n$, so that every point of $[-\beta, R_{\max}]$ lies within h of $\Lambda_{\beta,n}$. It has at most $(R_{\max} + \beta)/h + 2 = M_\beta n + 2 \le B^2 + 2$ points, using $M_\beta \le n \le B$. Fix $\lambda \in \Lambda_{\beta,n}$. By Lemma 4(iii), $0 \le w_\lambda \le M_\beta$ and $0 \le Z_\lambda \le \beta M_\beta^2/2$, so pointwise $w_\lambda^2 \le M_\beta w_\lambda$ and $Z_\lambda^2 = \frac{\beta^2}{4} w_\lambda^4 \le \frac{\beta^2 M_\beta^3}{4} w_\lambda$. Hence

$$\mathrm{Var}_\pi[w_\lambda] \le M_\beta\, \pi w_\lambda, \qquad \mathrm{Var}_\pi[Z_\lambda] \le \frac{\beta^2 M_\beta^3}{4}\, \pi w_\lambda.$$

Unlike at λ_β, πw_λ need not equal 1 at a grid point, so we keep it in the bounds. Since λ does not depend on the draws, Bernstein's inequality (Lemma 6) with $t = l$ applies to both w_λ and Z_λ. A union bound over these $2|\Lambda_{\beta,n}| \le 2(B^2 + 2)$ applications gives an event A_2 with $\mathbb{P}(A_2) \ge 1 - 4(B^2 + 2)e^{-l}$ on which, for every $\lambda \in \Lambda_{\beta,n}$,

$$|(\pi_n - \pi)w_\lambda| \le \sqrt{\frac{2M_\beta\,(\pi w_\lambda)\,l}{n}} + \frac{2M_\beta l}{3n} \tag{6}$$

and

$$|(\pi_n - \pi)Z_\lambda| \le \sqrt{\frac{\beta^2 M_\beta^3\,(\pi w_\lambda)\,l}{2n}} + \frac{\beta M_\beta^2 l}{3n}. \tag{7}$$

By Lemma 4(ii) applied with $Q = \pi_n$, $\widehat{\lambda} \in [-\beta, R_{\max}]$, so there is $\lambda^\circ \in \Lambda_{\beta,n}$ with $|\lambda^\circ - \widehat{\lambda}| \le h$. Since $\lambda \mapsto w_\lambda(y)$ is $1/\beta$-Lipschitz (Lemma 4(iv)) and $\pi_n w_{\widehat{\lambda}} = 1$ (Lemma 4(i)),

$$\pi_n w_{\lambda^\circ} \le \pi_n w_{\widehat{\lambda}} + \frac{h}{\beta} = 1 + \frac{1}{n}.$$

Write $x := (\pi w_{\lambda^\circ})$. On A_2, Equation (6) at $\lambda = \lambda^\circ$ with $c = M_\beta l/n$ gives

$$x \le \pi_n w_{\lambda^\circ} + |(\pi_n - \pi)w_{\lambda^\circ}| \le 1 + \frac{1}{n} + \sqrt{2c\,x} + \frac{2c}{3} \le 1 + \frac{1}{n} + \frac{x}{2} + \frac{5c}{3},$$

where the last step is AM–GM, $\sqrt{2c\,x} \le \frac{x}{2} + c$. Rearranging and using $n \ge 1$ and $c \le 1$,

$$(\pi w_{\lambda^\circ}) = x \le 2\Big(1 + \frac{1}{n} + \frac{5c}{3}\Big) \le \frac{22}{3} < 8.$$

Since $(\pi w_{\lambda^\circ}) < 8$, Equation (7) at $\lambda = \lambda^\circ$ gives, on A_2,

$$\begin{aligned}|(\pi_n - \pi)Z_{\lambda^\circ}| &\le \sqrt{\frac{8\beta^2 M_\beta^3 l}{2n}} + \frac{\beta M_\beta^2 l}{3n} \\ &= 2\beta M_\beta \sqrt{\frac{M_\beta l}{n}} + \frac{\beta M_\beta}{3} \cdot \frac{M_\beta l}{n} \\ &\le 4R_{\max}\sqrt{\frac{M_\beta l}{n}} + \frac{2R_{\max}}{3} \cdot \frac{M_\beta l}{n},\end{aligned}$$

using $\beta M_\beta < 2R_{\max}$ in Case 1. Both $\widehat{\lambda}$ and λ° lie in $[-\beta, R_{\max}]$, where $\lambda \mapsto Z_\lambda(y)$ is M_β-Lipschitz (Lemma 4(iv)), so $|Z_{\widehat{\lambda}}(y) - Z_{\lambda^\circ}(y)| \le M_\beta h$ for every y. Hence

$$|(\pi_n - \pi)(Z_{\widehat{\lambda}} - Z_{\lambda^\circ})| \le 2M_\beta h = \frac{2\beta M_\beta}{n} < \frac{4R_{\max}}{n} \le 4R_{\max} \cdot \frac{M_\beta l}{n},$$

where the last step uses $M_\beta \ge 1$ and $l > 1$. By the triangle inequality, on A_2,

$$|(\pi_n - \pi)Z_{\widehat{\lambda}}| \le 4R_{\max}\sqrt{\frac{M_\beta l}{n}} + \frac{14R_{\max}}{3} \cdot \frac{M_\beta l}{n}.$$

On A_1, the term at λ_β satisfies the same kind of bound with coefficients $\sqrt{2}$ and $\frac{2}{3}$. So on $A_1 \cap A_2$,

$$|\widehat{G}(\beta; n) - G(\beta)| \le 5R_{\max}\Big[\sqrt{\frac{M_\beta l}{n}} + \frac{M_\beta l}{n}\Big],$$

with $\mathbb{P}(A_1 \cap A_2) \ge 1 - 2e^{-l} - 4(B^2+2)e^{-l} \ge 1 - 6(B^2+2)e^{-l}$.

Case 2: $\beta \ge R_{\max}$. Since $0 \le \hat{r} \le R_{\max} \le \beta$, the number $\lambda := \pi\hat{r} - \beta$ satisfies $\lambda \le 0 \le \hat{r}(y)$ for every y, so $(\hat{r} - \lambda)_+ = \hat{r} - \lambda$ and $\pi w_\lambda = (\pi\hat{r} - \lambda)/\beta = 1$. By uniqueness of the root (Lemma 4(i)), $\lambda_\beta = \pi\hat{r} - \beta$; likewise $\widehat{\lambda} = \pi_n\hat{r} - \beta$. Write $\mathrm{Var}_\pi[\hat{r}] := \pi\hat{r}^2 - (\pi\hat{r})^2$ and $\mathrm{Var}_{\pi_n}[\hat{r}] := \pi_n\hat{r}^2 - (\pi_n\hat{r})^2$. Since $\hat{r} - \lambda_\beta = (\hat{r} - \pi\hat{r}) + \beta$ and $\pi(\hat{r} - \pi\hat{r}) = 0$,

$$\begin{aligned}\mathrm{Obj}(\beta) &= \lambda_\beta + \frac{\beta}{2} + \frac{1}{2\beta}\,\pi\big[(\hat{r} - \pi\hat{r} + \beta)^2\big] \\ &= \pi\hat{r} - \frac{\beta}{2} + \frac{\mathrm{Var}_\pi[\hat{r}] + \beta^2}{2\beta} \\ &= \pi\hat{r} + \frac{\mathrm{Var}_\pi[\hat{r}]}{2\beta},\end{aligned}$$

and the same computation under π_n gives $\widehat{\mathrm{Obj}}(\beta; n) = \pi_n\hat{r} + \mathrm{Var}_{\pi_n}[\hat{r}]/(2\beta)$. The values $\hat{r}(y_j) \in [0, R_{\max}]$ and $\hat{r}(y_j)^2 \in [0, R_{\max}^2]$ are i.i.d., so Hoeffding's inequality (Lemma 6) with $t = l$, applied to each, gives an event A_3 with $\mathbb{P}(A_3) \ge 1 - 4e^{-l}$ on which

$$|(\pi_n - \pi)\hat{r}| \le R_{\max}\sqrt{\frac{l}{2n}}, \qquad |(\pi_n - \pi)\hat{r}^2| \le R_{\max}^2\sqrt{\frac{l}{2n}}.$$

On A_3, factoring $(\pi\hat{r})^2 - (\pi_n\hat{r})^2$ and using $\pi\hat{r} + \pi_n\hat{r} \le 2R_{\max}$,

$$\begin{aligned}\big|\mathrm{Var}_\pi[\hat{r}] - \mathrm{Var}_{\pi_n}[\hat{r}]\big| &\le |(\pi_n - \pi)\hat{r}^2| + (\pi\hat{r} + \pi_n\hat{r})\,|(\pi_n - \pi)\hat{r}| \\ &\le 3R_{\max}^2\sqrt{\frac{l}{2n}}.\end{aligned}$$

Therefore, using $R_{\max} \le \beta$, and $M_\beta \ge 1$, on A_3

$$\begin{aligned}|\widehat{G}(\beta; n) - G(\beta)| &\le |(\pi_n - \pi)\hat{r}| + \frac{|\mathrm{Var}_\pi[\hat{r}] - \mathrm{Var}_{\pi_n}[\hat{r}]|}{2\beta} \\ &\le \frac{5R_{\max}}{2}\sqrt{\frac{l}{2n}} \\ &\le 2R_{\max}\sqrt{\frac{M_\beta l}{n}} \\ &\le 5R_{\max}\Big[\sqrt{\frac{M_\beta l}{n}} + \frac{M_\beta l}{n}\Big],\end{aligned}$$

and the failure probability is at most $4e^{-l} \le 6(B^2+2)e^{-l}$.

Union bound. For each triple $(i, n, \beta) \in [k] \times [B] \times \mathcal{G}$, let $E_{i,n,\beta}$ be the event constructed above for that triple: the whole space if $M_\beta l/n > 1$, $A_1 \cap A_2$ in Case 1, and A_3 in Case 2. Each triple falls into exactly one of these cases, and in each, $\mathbb{P}(E^c_{i,n,\beta}) \le 6(B^2+2)e^{-l}$. Let $\mathcal{E}_{\text{obj}} := \bigcap_{i,n,\beta} E_{i,n,\beta}$; on it the claimed bound holds simultaneously for all kBN_β triples. Since $e^{-l} = \delta/\big(16kN_\beta B(B^2+2)\big)$, a union bound gives

$$\mathbb{P}(\mathcal{E}^c_{\text{obj}}) \le kBN_\beta \cdot 6(B^2+2)e^{-l} = \frac{3\delta}{8} \le \frac{\delta}{2}.$$

□

Lemma 10 (Concentration of the empirical reward model error)**.** *Under Assumption 2, there is an event $\mathcal{E}_{\text{err}}$ with $\mathbb{P}(\mathcal{E}_{\text{err}}) \ge 1 - \delta/2$ on which, simultaneously for every $i \in [k]$ and $n \in [B]$,*

$$\widehat{\varepsilon}_{i,n} \le \bar{\varepsilon}_i + R_{\max}\sqrt{\frac{l}{n}}.$$

Proof. Fix (i, n) and define $\xi(y) := (\hat{r}_i(y) - r^\star_i(y))^2$, so that $\widehat{\varepsilon}^2_{i,n} = \frac{1}{n}\sum_{j=1}^n \xi(y_{i,j})$ and $\mathbb{E}_{y\sim\pi_{i,\text{ref}}}[\xi(y)] = \varepsilon_i^2$. Since $\hat{r}_i$ and $r^\star_i$ take values in $[0, R_{\max}]$, we have $0 \le \xi(y) \le R^2_{\max}$, hence $\xi(y)^2 \le R^2_{\max}\,\xi(y)$ for every y, and

$$\text{Var}_{y\sim\pi_{i,\text{ref}}}[\xi(y)] \le \mathbb{E}_{y\sim\pi_{i,\text{ref}}}[\xi(y)^2] \le R^2_{\max}\,\mathbb{E}_{y\sim\pi_{i,\text{ref}}}[\xi(y)] = R^2_{\max}\varepsilon_i^2.$$

The values $\xi(y_{i,1}), \dots, \xi(y_{i,n})$ are i.i.d. in $[0, R^2_{\max}]$, so Bernstein's inequality (Lemma 6) with $t = l$ gives an event $E_{i,n}$ with $\mathbb{P}(E^c_{i,n}) \le 2e^{-l}$ on which

$$\begin{aligned}
\widehat{\varepsilon}^2_{i,n} &\le \varepsilon_i^2 + \sqrt{\frac{2R^2_{\max}\varepsilon_i^2\, l}{n}} + \frac{2R^2_{\max}\, l}{3n} \\
&\le \varepsilon_i^2 + 2\varepsilon_i R_{\max}\sqrt{\frac{l}{n}} + R^2_{\max}\frac{l}{n} \\
&= \left(\varepsilon_i + R_{\max}\sqrt{\frac{l}{n}}\right)^2.
\end{aligned}$$

Taking square roots and using $\varepsilon_i \le \bar{\varepsilon}_i$ (Assumption 2), $\widehat{\varepsilon}_{i,n} \le \bar{\varepsilon}_i + R_{\max}\sqrt{l/n}$ on $E_{i,n}$. Let $\mathcal{E}_{\text{err}} := \bigcap_{i,n} E_{i,n}$. Since $e^{-l} = \delta/\big(16kN_\beta B(B^2+2)\big)$, a union bound over the kB pairs gives

$$\mathbb{P}(\mathcal{E}^c_{\text{err}}) \le 2kB\, e^{-l} = \frac{\delta}{8N_\beta(B^2+2)} \le \frac{\delta}{2}.$$

□

Combining the above two concentration results yields the following proposition.

Proposition 3. *Under Assumption 2, let $\mathcal{E} := \mathcal{E}_{\text{obj}} \cap \mathcal{E}_{\text{err}}$, where $\mathcal{E}_{\text{obj}}$ and $\mathcal{E}_{\text{err}}$ are the events of Theorem 4 and Lemma 10. Then $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$, and on $\mathcal{E}$, simultaneously for every $i \in [k]$, $n \in [B]$ and $\beta \in \mathcal{G}$:*

(a) $\big|\widehat{G}_i(\beta; n) - G_i(\beta)\big| \le \Delta^{\text{obj}}(n, \beta, \delta)$;
(b) the empirical pessimistic policy $\widehat{\pi}_{i,\beta,n}$ satisfies

$$J_i(\widehat{\pi}_{i,\beta,n}) \ge \widehat{G}_i(\beta; n) - \Delta^{\text{ITP}}(n, \beta, \delta) \ge G_i(\beta) - \Delta(n, \beta, \delta).$$

Proof. By Theorem 4 and Lemma 10, $\mathbb{P}(\mathcal{E}^c) \le \mathbb{P}(\mathcal{E}^c_{\text{obj}}) + \mathbb{P}(\mathcal{E}^c_{\text{err}}) \le \delta/2 + \delta/2 = \delta$. Work on $\mathcal{E}$ and fix (i, n, β). Part (a) is Theorem 4, since $\mathcal{E} \subseteq \mathcal{E}_{\text{obj}}$. For (b), Lemma 10 gives $\widehat{\varepsilon}_{i,n} - \bar{\varepsilon}_i \le R_{\max}\sqrt{l/n}$; since the right side is nonnegative, also $(\widehat{\varepsilon}_{i,n} - \bar{\varepsilon}_i)_+ \le R_{\max}\sqrt{l/n}$. Combining this with Lemma 8,

$$\begin{aligned}
J_i(\widehat{\pi}_{i,\beta,n}) &\ge \widehat{G}_i(\beta; n) - \sqrt{M_\beta}\,(\widehat{\varepsilon}_{i,n} - \bar{\varepsilon}_i)_+ \\
&\ge \widehat{G}_i(\beta; n) - R_{\max}\sqrt{\frac{M_\beta l}{n}} \\
&= \widehat{G}_i(\beta; n) - \Delta^{\text{ITP}}(n, \beta, \delta).
\end{aligned}$$

Finally, (a) gives $\widehat{G}_i(\beta; n) \ge G_i(\beta) - \Delta^{\text{obj}}(n, \beta, \delta)$; subtracting $\Delta^{\text{ITP}}(n, \beta, \delta)$ from both sides gives the second inequality in (b). □

On the good event $\mathcal{E}$, the following lemma shows that the pessimistic index $L_i(n)$ is accurate to the truth $P_i(n)$ within the tolerance η, and the returned policy is worth at least the pessimistic index minus the tolerance.

Lemma 11. *Let $L_i(n)$ and $\widehat{\beta}_i(n)$ be as in Algorithm 2. On the event $\mathcal{E}$ of Proposition 3, for every $i \in [k]$ and all integers $n_0 \le n \le n' \le B$:*

(P1) $L_i(n) \le P_i(n) + \eta$;
(P2) $L_i(n) \ge P_i(n) - \eta$;
(P3) $J_i(\widehat{\pi}_{i,\widehat{\beta}_i(n),n}) \ge L_i(n) - \eta$;
(P4) $P_i(n) \le P_i(n')$.

Property (P4) holds deterministically.

Proof. Fix $n \ge n_0$, so that $\mathcal{B}(n)$ is finite and nonempty. For every $\beta \in \mathcal{B}(n)$, the quantities $\Delta^{\mathrm{obj}}(n,\beta,\delta)$ and $\Delta^{\mathrm{ITP}}(n,\beta,\delta)$ are nonnegative with sum $\Delta(n,\beta,\delta) \le \eta$, so each is at most η. Since Proposition 3 holds simultaneously for all $\beta \in \mathcal{G}$, it also applies at the data-dependent choice $\widehat{\beta}_i(n)$.

(P1) For every $\beta \in \mathcal{B}(n)$, Proposition 3(a) gives $\widehat{G}_i(\beta;n) \le G_i(\beta) + \eta \le P_i(n) + \eta$; maximize over $\beta \in \mathcal{B}(n)$.

(P2) Let $\beta^* \in \arg\max_{\beta\in\mathcal{B}(n)} G_i(\beta)$. By Proposition 3(a), $L_i(n) \ge \widehat{G}_i(\beta^*;n) \ge G_i(\beta^*) - \eta = P_i(n) - \eta$.

(P3) Since $\widehat{\beta}_i(n)$ maximizes $\widehat{G}_i(\cdot\,;n)$ over $\mathcal{B}(n)$, we have $\widehat{G}_i(\widehat{\beta}_i(n);n) = L_i(n)$. By Proposition 3(b) at $\beta = \widehat{\beta}_i(n)$,

$$J_i(\widehat{\pi}_{i,\widehat{\beta}_i(n),n}) \ge \widehat{G}_i(\widehat{\beta}_i(n);n) - \Delta^{\mathrm{ITP}}(n,\widehat{\beta}_i(n),\delta) \ge L_i(n) - \eta.$$

(P4) Since $\Delta(\cdot,\beta,\delta)$ is decreasing in its first argument, $\mathcal{B}(n) \subseteq \mathcal{B}(n')$, and a maximum over a larger set is at least as large. □

The next lemma quantifies the sample requirements to make a regularization level admissible.

Lemma 12 (Sample requirement for admissibility)**.** *If $\eta \le 14.4R_{\max}$, then $\Delta(n,\beta,\delta) \le \eta$ whenever $n \ge 144R_{\max}^2 M_\beta l/\eta^2$. Consequently*

$$n_0(\eta,\delta) \;\le\; \left\lceil \frac{144\,R_{\max}^2\,M_{\beta_{\max}}\,l}{\eta^2} \right\rceil \qquad \text{and} \qquad \beta \in \mathcal{B}(n) \;\text{ whenever }\; n \ge \frac{144R_{\max}^2 M_\beta l}{\eta^2}.$$

Proof. If $n \ge 144R_{\max}^2 M_\beta l/\eta^2$ then $\sqrt{M_\beta l/n} \le \eta/(12R_{\max})$, so $6R_{\max}\sqrt{M_\beta l/n} \le \eta/2$; and $M_\beta l/n \le \eta^2/(144R_{\max}^2)$, so $5R_{\max}M_\beta l/n \le 5\eta^2/(144R_{\max}) \le \eta/2$ as $\eta \le 14.4R_{\max}$. □

We can now prove the fixed-allocation guarantee for Algorithm 2: on the good event, its worst-group value is within 4η of the best admissible surrogate of any allocation that uses at most half the budget.

Theorem 5. *Under Assumption 2, let $\eta > 0$, $\delta \in (0,1)$, and suppose $B \ge 2kn_0(\eta,\delta)$. Let $\widehat{n}_i$ and $\widehat{\beta}_i$ be the final count and selected regularization of group i in Algorithm 2, let $\widehat{\pi}_i := \widehat{\pi}_{i,\widehat{\beta}_i,\widehat{n}_i}$ be the law of the returned response $\widehat{y}_i$ given the data, and let $J_i(\widehat{\pi}_i) = \mathbb{E}[r_i^\star(\widehat{y}_i) \mid \text{data}]$ be its conditional true value. On the event $\mathcal{E}$ in Proposition 3, simultaneously for every fixed allocation $m \in \mathbb{N}^k$ with $m_i \ge n_0$ for all i and $\sum_i m_i \le B/2$,*

$$\min_{i\in[k]} J_i(\widehat{\pi}_i) \;\ge\; \min_{i\in[k]} P_i(m_i) \;-\; 4\eta.$$

Consequently,

$$\min_{i\in[k]} \mathbb{E}\big[J_i(\widehat{\pi}_i)\big] \;\ge\; \mathbb{E}\Big[\min_{i\in[k]} J_i(\widehat{\pi}_i)\Big] \;\ge\; \max_{\substack{m:\ m_i\ge n_0,\\ \sum_i m_i \le B/2}} \min_{i\in[k]} P_i(m_i) \;-\; 4\eta \;-\; \delta R_{\max}.$$

Proof. Fix, for each group i, an i.i.d. stream $y_{i,1}, \dots, y_{i,B} \sim \pi_{i,\mathrm{ref}}$, independent across groups, and let the j-th draw of group i in Algorithm 2 be $y_{i,j}$; this does not change the law of the run. Every quantity the algorithm computes for group i at count n—$L_i(n)$, $\widehat{\beta}_i(n)$ and $\widehat{\pi}_{i,\widehat{\beta}_i(n),n}$—depends only on $y_{i,1}, \dots, y_{i,n}$, and the algorithm's internal randomness (tie-breaking and the final draw) is independent of the streams. Since Lemma 11 holds on $\mathcal{E}$ simultaneously for all $n \in \{n_0, \dots, B\}$, it applies at the random counts the algorithm reaches. Work on $\mathcal{E}$ and fix a feasible m, i.e., $m_i \geq n_0$ for all i and $\sum_i m_i \leq B/2$.

Since $B \geq kn_0$, the initialization uses $kn_0 \leq B$ draws, and each loop iteration draws b responses with $1 \leq b \leq B - \sum_j n_j$; hence the algorithm terminates with $\sum_i \widehat{n}_i = B \geq 2\sum_i m_i$. If $\widehat{n}_i < 2m_i$ held for every i, summing would give $B < 2\sum_i m_i \leq B$, a contradiction. Hence there is $g \in [k]$ with $\widehat{n}_g \geq 2m_g$.

Since $\widehat{n}_g \geq 2m_g > n_0$, group g received at least one batch after the initialization (line 2). Let t_g be the iteration of the while loop (line 6) in which g receives its last batch, and for each $i \in [k]$ let $n_i(t_g)$ be the count of group i at the start of iteration t_g; set $s_g := n_g(t_g)$, and note that $n_i(t_g) \geq n_0$ for all i. By line 8 this batch adds at most s_g responses, and g receives none afterwards, so $\widehat{n}_g \leq 2s_g$, i.e., $s_g \geq \widehat{n}_g/2 \geq m_g$. By line 7, $L_g(s_g) \leq L_i(n_i(t_g))$ for all i, and since counts never decrease, $n_i(t_g) \leq \widehat{n}_i$.

For every $i \in [k]$, applying (P4) of Lemma 11 with $n_i(t_g) \leq \widehat{n}_i$, (P1) at $n_i(t_g)$, the argmin inequality of the last step, (P2) at s_g, and (P4) with $m_g \leq s_g$,

$$\begin{aligned} P_i(\widehat{n}_i) &\geq P_i(n_i(t_g)) \\ &\geq L_i(n_i(t_g)) - \eta \\ &\geq L_g(s_g) - \eta \\ &\geq P_g(s_g) - 2\eta \\ &\geq P_g(m_g) - 2\eta \\ &\geq \min_j P_j(m_j) - 2\eta. \end{aligned}$$

By (P3) and (P2) at $\widehat{n}_i$, for every $i \in [k]$,

$$\begin{aligned} J_i(\widehat{\pi}_i) &= J_i\big(\widehat{\pi}_{i,\widehat{\beta}_i(\widehat{n}_i),\widehat{n}_i}\big) \\ &\geq L_i(\widehat{n}_i) - \eta \\ &\geq P_i(\widehat{n}_i) - 2\eta \\ &\geq \min_j P_j(m_j) - 4\eta. \end{aligned}$$

Taking the minimum over i gives the first claim. Since m is arbitrary, it holds for all feasible m simultaneously on $\mathcal{E}$.

Jensen's inequality gives $\min_i \mathbb{E}[J_i(\widehat{\pi}_i)] \geq \mathbb{E}[\min_i J_i(\widehat{\pi}_i)]$. Fix a feasible m and let $\Lambda_m := \min_j P_j(m_j) - 4\eta$. Since $G_j(\beta) \leq \mathrm{Obj}_j(\beta) \leq R_{\max}$ by Lemma 5(iii), $\Lambda_m \leq R_{\max}$; also $J_i \geq 0$ because $r_i^\star \geq 0$. If $\Lambda_m < 0$, the bound $\mathbb{E}[\min_i J_i(\widehat{\pi}_i)] \geq \Lambda_m - \delta R_{\max}$ is immediate. If $\Lambda_m \geq 0$, then by the last step and $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$,

$$\mathbb{E}\big[\min_i J_i(\widehat{\pi}_i)\big] \geq \mathbb{E}\big[\min_i J_i(\widehat{\pi}_i)\,\mathbf{1}_{\mathcal{E}}\big] \geq \Lambda_m\,\mathbb{P}(\mathcal{E}) \geq \Lambda_m(1-\delta) \geq \Lambda_m - \delta R_{\max}.$$

Maximizing over feasible m completes the proof. □

Now we are finally ready to prove Theorem 3.

Theorem 3 (Regret Upper Bound). *Under Assumption 2, fix any $\beta^0 = (\beta_i^0)_{i\in[k]} \in (0,\infty)^k$ and run Algorithm 2 with confidence $\delta \in (0,1)$ and tolerance $\eta = \eta_B(\beta^0) := 24R_{\max}\sqrt{l\sum_{i\in[k]} M_{\beta_i^0}/B}$. Then, for every profile $(\beta_i)_{i\in[k]} \in \mathcal{G}^k$ with $\beta_i \geq \beta_i^0$ for all i,*

$$\mathcal{R}_B \;\leq\; \max_{i\in[k]} \Big(\underbrace{2\sqrt{C_i}\,\bar{\varepsilon}_i}_{\textit{floor}} + \underbrace{\frac{\beta_i C_i}{2} + \frac{\bar{\varepsilon}_i^2}{2\beta_i}}_{\textit{bias}(\beta_i)} \Big) \;+\; \underbrace{96R_{\max}\sqrt{\frac{l\sum_i \big(1 + R_{\max}/\beta_i^0\big)}{B}}}_{\textit{variance}(\beta^0,B)} + \delta R_{\max}.$$

Proof. Fix a grid profile $(\beta_i)_{i\in[k]}$ as in the statement. By Lemma 9 with $n = n_i^\star$, the comparator satisfies $\pi_i^\star \ll \pi_{i,\mathrm{ref}}$ and $C_i \le n_i^\star < \infty$; moreover $J_i(\pi_i^\star) = \mathbb{E}\big[\max_{s\le n_i^\star} r_i^\star(x_i, Y_{i,s})\big] = V_i(n_i^\star)$, so $\min_i J_i(\pi_i^\star) = V^\star$. Write $\Sigma := \sum_i M_{\beta_i^0}$ and $b_i := 2\sqrt{C_i}\,\bar\varepsilon_i + \beta_i C_i/2 + \bar\varepsilon_i^2/(2\beta_i)$, so it suffices to prove $\mathcal{R}_B \le \max_i b_i + 4\eta_B(\beta^0) + \delta R_{\max}$. We have that $M_\beta \ge 1$, $\Sigma \ge k$. As $0 \le r_i^\star \le R_{\max}$, we have $V^\star \le R_{\max}$ and $\min_i \mathbb{E}[r_i^\star(x_i, \hat y_i)] \ge 0$, so $\mathcal{R}_B \le R_{\max}$ for any output, and the claim holds whenever $4\eta_B(\beta^0) = 96R_{\max}\sqrt{l\Sigma/B} \ge R_{\max}$. Assume from now on that $l\Sigma/B < 96^{-2}$. Then $\eta := \eta_B(\beta^0) < R_{\max}/4 \le 14R_{\max}$, and $k \le \Sigma < B/(96^2 l) < B/4$ since $l > 1$.

A feasible allocation. Set $m_i := \lceil 144R_{\max}^2 M_{\beta_i} l/\eta^2 \rceil$. Since $\beta_i \ge \beta_i^0$ gives $M_{\beta_i} \le M_{\beta_i^0}$, and $\eta^2 = 576R_{\max}^2 l\Sigma/B$,

$$m_i \le \left\lceil \frac{144R_{\max}^2 M_{\beta_i^0} l}{\eta^2} \right\rceil = \left\lceil \frac{BM_{\beta_i^0}}{4\Sigma} \right\rceil \le \frac{BM_{\beta_i^0}}{4\Sigma} + 1, \qquad \text{so} \qquad \sum_i m_i \le \frac{B}{4} + k \le \frac{B}{2}.$$

As $\eta \le 14R_{\max}$, Lemma 12 gives $\beta_i \in \mathcal{B}(m_i)$ for every i; in particular $\mathcal{B}(m_i) \ne \emptyset$, so $m_i \ge n_0$ and m is feasible in the sense of Theorem 5. Moreover $B \ge 2\sum_i m_i \ge 2kn_0$, so Theorem 5 applies.

Conclusion. Since $\beta_i \in \mathcal{B}(m_i)$, the definition of P_i gives $P_i(m_i) \ge G_i(\beta_i)$, and since $\pi_i^\star \ll \pi_{i,\mathrm{ref}}$ with $C_i < \infty$, the second part of Lemma 7 with $\pi = \pi_i^\star$ gives $G_i(\beta_i) \ge J_i(\pi_i^\star) - b_i$. As m is feasible, the second display of Theorem 5 yields

$$\begin{aligned}\min_i \mathbb{E}[J_i(\widehat\pi_i)] &\ge \min_i P_i(m_i) - 4\eta - \delta R_{\max} \\ &\ge \min_i \big(J_i(\pi_i^\star) - b_i\big) - 4\eta - \delta R_{\max} \\ &\ge V^\star - \max_i b_i - 4\eta - \delta R_{\max}.\end{aligned}$$

Since $\mathbb{E}[J_i(\widehat\pi_i)] = \mathbb{E}[r_i^\star(x_i, \hat y_i)]$ by the tower property, the left side equals $V^\star - \mathcal{R}_B$, and rearranging proves the bound. □

D.3 Choosing the Regularization

Corollary 1. *Assume $0 < \bar\varepsilon_i \le R_{\max}$ for all i, that the grid satisfies $\beta_{\min}/\rho \le \min_i \bar\varepsilon_i/\sqrt{C_i}$ and $\beta_{\max} \ge R_{\max}$, and let $C_{\max} := \max_i C_i$. Run Algorithm 2 with $\delta = 1/B$ and $\eta = \eta_B(\beta^0)$.*

(i) If $B \ge 27^3 kl/C_{\max}^2$ and $\beta_i^0 := 27R_{\max}(kl/B)^{1/3}C_{\max}^{-2/3}$ for all i, then

$$\mathcal{R}_B \;\le\; \frac{5+\rho}{2}\max_{i\in[k]}\sqrt{C_i}\,\bar\varepsilon_i \;+\; (14\rho + 27)\,R_{\max}\Big(\frac{k\,l\,C_{\max}}{B}\Big)^{1/3} \;+\; \frac{R_{\max}}{B}.$$

(ii) If $\beta_i^0 := \bar\varepsilon_i/\sqrt{C_i}$ for all i, then

$$\mathcal{R}_B \;\le\; \frac{5+\rho}{2}\max_{i\in[k]}\sqrt{C_i}\,\bar\varepsilon_i \;+\; 96R_{\max}\sqrt{\frac{l\sum_i\big(1 + R_{\max}\sqrt{C_i}/\bar\varepsilon_i\big)}{B}} \;+\; \frac{R_{\max}}{B}.$$

Proof. We apply Theorem 3 to a grid profile (β_i) with $\beta_i \ge \beta_i^0$ chosen below. By Definition 2, consecutive grid points differ by the factor ρ, so for every $a \in [\beta_{\min}/\rho, \beta_{\max}]$ the grid contains a point in $[a, \rho a]$. Recall also that $\bar\varepsilon_i/\sqrt{C_i} \le R_{\max} \le \beta_{\max}$, since $\bar\varepsilon_i \le R_{\max}$ and $C_i \ge 1$.

Part (i). The condition $B \ge 27^3 kl/C_{\max}^2$ is equivalent to $\beta^0 \le R_{\max}$. Set $a_i := \max\{\beta^0, \bar\varepsilon_i/\sqrt{C_i}\}$ and pick $\beta_i \in \mathcal{G} \cap [a_i, \rho a_i]$, which exists because $\beta_{\min}/\rho \le \bar\varepsilon_i/\sqrt{C_i} \le a_i \le R_{\max} \le \beta_{\max}$. Moreover $\beta_i \ge a_i \ge \beta^0$, so Theorem 3 applies. Since $\beta_i \ge \bar\varepsilon_i/\sqrt{C_i}$, the term $\bar\varepsilon_i^2/(2\beta_i)$ is at most $\frac12\sqrt{C_i}\,\bar\varepsilon_i$. Since $\beta_i \le \rho a_i \le \rho\beta^0 + \rho\bar\varepsilon_i/\sqrt{C_i}$, the term $\beta_i C_i/2$ is at most $\frac{\rho}{2}\sqrt{C_i}\,\bar\varepsilon_i + \rho\beta^0 C_{\max}/2$, and

$$\frac{\rho\beta^0 C_{\max}}{2} = \frac{27\rho}{2}R_{\max}\Big(\frac{klC_{\max}}{B}\Big)^{1/3} \le 14\rho R_{\max}\Big(\frac{klC_{\max}}{B}\Big)^{1/3}.$$

Hence the first maximum term in Theorem 3 is at most $\frac{5+\rho}{2}\max_i \sqrt{C_i}\,\bar{\varepsilon}_i + 14\rho R_{\max}(klC_{\max}/B)^{1/3}$. For the variance term, $\beta^0 \le R_{\max}$ gives $M_{\beta^0} = 1 + R_{\max}/\beta^0 \le 2R_{\max}/\beta^0$, so

$$96R_{\max}\sqrt{\frac{lkM_{\beta^0}}{B}} \le 96R_{\max}\sqrt{\frac{2klR_{\max}}{\beta^0 B}} = 96\sqrt{\tfrac{2}{27}}\,R_{\max}\Big(\frac{klC_{\max}}{B}\Big)^{1/3} \le 27R_{\max}\Big(\frac{klC_{\max}}{B}\Big)^{1/3},$$

where the equality term substitutes $\beta^0 = 27R_{\max}(kl/B)^{1/3}C_{\max}^{-2/3}$.

Part (ii). Pick $\beta_i \in \mathcal{G} \cap [\beta_i^0, \rho\beta_i^0]$, which exists because $\beta_{\min}/\rho \le \beta_i^0 = \bar{\varepsilon}_i/\sqrt{C_i} \le \beta_{\max}$. Then $\bar{\varepsilon}_i^2/(2\beta_i) \le \frac{1}{2}\sqrt{C_i}\,\bar{\varepsilon}_i$ and $\beta_i C_i/2 \le \frac{\rho}{2}\sqrt{C_i}\,\bar{\varepsilon}_i$, so the first maximum term in Theorem 3 is bouned by $\frac{5+\rho}{2}\max_i \sqrt{C_i}\,\bar{\varepsilon}_i$. The variance term equals $96R_{\max}\sqrt{l\sum_i(1 + R_{\max}\sqrt{C_i}/\bar{\varepsilon}_i)/B}$ because $M_{\beta_i^0} = 1 + R_{\max}\sqrt{C_i}/\bar{\varepsilon}_i$. □

E Further Details of Experimental Setting

Datasets. The five SHP groups (domains) are `askacademia`, `askculinary`, `askengineers`, `asksciencefiction`, and `changemyview`. Each response pool for a prompt consists 256 responses and defines a fixed sampling trajectory: allocating n draws to a prompt reveals and evaluates exactly the first n responses in its pool. All methods use the same response ordering, so differences in allocation determine how far each method proceeds along each trajectory. Each evaluation trial draws one prompt from the evaluation set of each group. We use the same random seed 2026 across methods, and all methods share the same prompt tuple and ordered response pools within each trial. We independently min–max normalize the proxy and judge scores to $[0, 1]$ within each domain, using their respective minimum and maximum scores across all responses to the 50 original prompts in that domain. Group reward error $\bar{\varepsilon}_i$ is estimated using the RMSE between proxy and Judge scores on the tuning split.

Details of model training. For both `Qwen2.5-3B-Instruct` and `Qwen2.5-7B-Instruct`, we train each model for 2 epochs on each domain with LoRA, using a learning rate of 5×10^{-6} and random seed 2026. For `Qwen2.5-3B-Instruct`, we use a batch size of 4 with 4 gradient accumulation steps, while for `Qwen2.5-7B-Instruct`, we use a batch size of 2 with 8 gradient accumulation steps. Thus, both the proxy and Judge reward models are trained with an effective batch size of 16.

Details of algorithms. Uniform-BoN and Greedy-BoN both select the highest-proxy-reward response within each group. Uniform-BoN splits the budget equally across groups, whereas Greedy-BoN allocates it adaptively. BGP selects β separately for each group from a grid of 12 geometrically spaced values in $[10^{-4}, 10^{-1}]$, with the range motivated by prior work on inference-time pessimism (Huang et al., 2025). BGP-Fixed retains BGP's greedy-pessimistic adaptive allocation and pessimistic response selection but uses a single global β tuned using proxy rewards on the tuning split. Uniform-Pess-Fixed retains the same fixed β and pessimistic selection rule while allocating the budget uniformly. Comparing BGP with BGP-Fixed assesses automatic group-specific regularization against a global value tuned with proxy scores, while comparing BGP-Fixed with Uniform-Pess-Fixed isolates the contribution of adaptive allocation under pessimistic selection. In F, we additionally compare with Uniform-Pess-Auto, which retains pessimistic selection and automatic β search while using uniform budget allocation across groups.

F Additional Experimental Results

Why do we divide groups into two subsets? Even after reward normalization to the range $[0, 1]$ for each group, Judge-oracle BoN reveals large differences in the maximum attainable rewards within the response pools, suggesting that the domains differ in prompt difficulty and score distribution. We therefore divide the five domains into a higher-reward subset (two domains) and a lower-reward subset (three domains) to examine whether BGP balances rewards among domains of

similar difficulty. And in the Section 5.2, it shows that BGP can extend its advantage from worst-group performance to low reward disparity among groups.

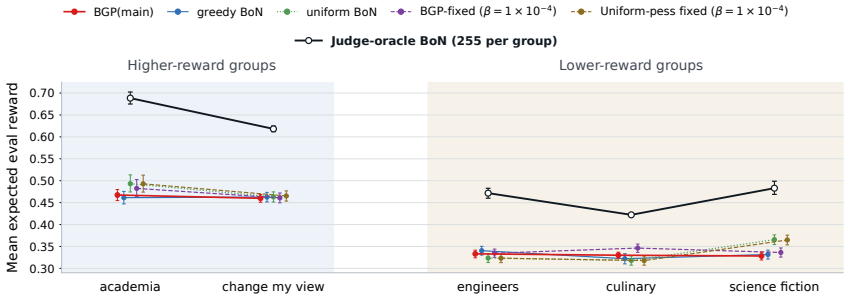

Figure 3: Mean expected Judge rewards for each method across groups at $B = 255$. Judge-oracle BoN provides an upper bound within the evaluated candidate pools by allocating the full budget to each group and selecting the response with the highest Judge score. BGP achieves stable performance within each subset.

Why does greedy pessimistic allocation yield limited gains? Figure 4(a) shows that BGP improves worst-group proxy rewards over Uniform-Pess-Auto, but does not yield consistent gains under the Judge. This occurs because the original five groups exhibit estimation errors in both directions, while the pessimistic index uses only the estimated error magnitude without distinguishing its sign. As a result, the index may not preserve the group ordering under evaluation, since an underestimated group can receive a low index and attract additional samples even when another group has a lower Judge reward. This limits the benefit of greedy pessimistic allocation. In contrast, when proxy rewards consistently overestimate Judge scores across groups, BGP achieves higher mean worst-group Judge rewards than Uniform-Pess-Auto in Figure 4(b), highlighting the importance of error direction for effective allocation.

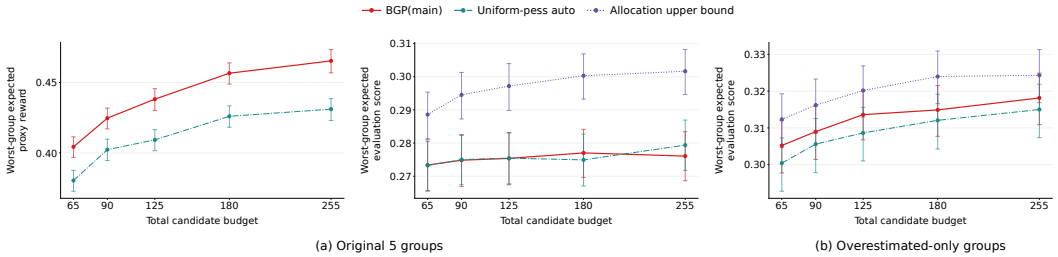

Figure 4: **(a)** Expected rewards under the proxy (left) and Judge (right) for the original five groups. BGP outperforms Uniform-Pess-Auto under the proxy reward, while the two methods achieve similar performance under the Judge evaluation. **(b)** Expected Judge rewards when all groups exhibit proxy reward overestimation. Dotted curves show the allocation upper bound determined by Judge scores. BGP consistently outperforms Uniform-Pess-Auto across budgets and approaches the allocation upper bound. Although BGP's error bars overlap with those of Uniform-Pess-Auto, they also overlap with the upper-bound estimates, indicating that BGP's mean performance is closer to the oracle allocation.

How effective is automatic β selection? As shown in Figure 5, BGP achieves worst-group Judge rewards comparable to BGP-Fixed with $\beta = 0.015$, selected using Judge scores on the tuning split. At larger budgets, BGP also attains higher proxy rewards. These results support the effectiveness of automatic selection without directly tuning β on Judge scores. Figure 6 further shows that the selected β vary across groups and budgets. The median β decreases with increasing budget in some groups and remains stable in others, illustrating how automatic selection accommodates different regularization strengths across groups and sampling budgets.

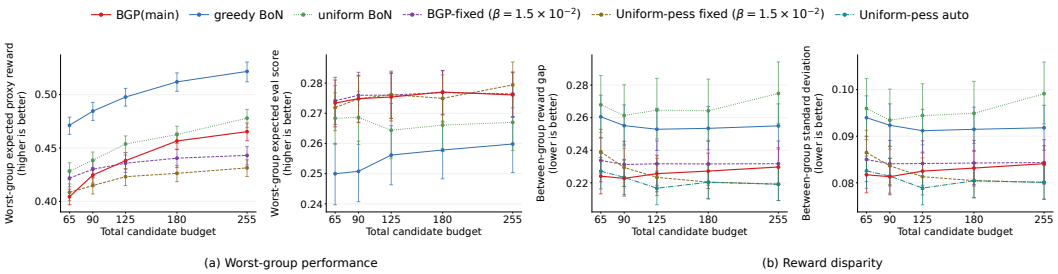

Figure 5: Performance comparing Judge-tuned β. **(a)** Worst-group expected rewards under the 3B proxy (left) and 7B Judge (right). **(b)** Between-group reward disparity, measured by the gap between the highest and lowest group rewards (left) and the standard deviation of group rewards (right).

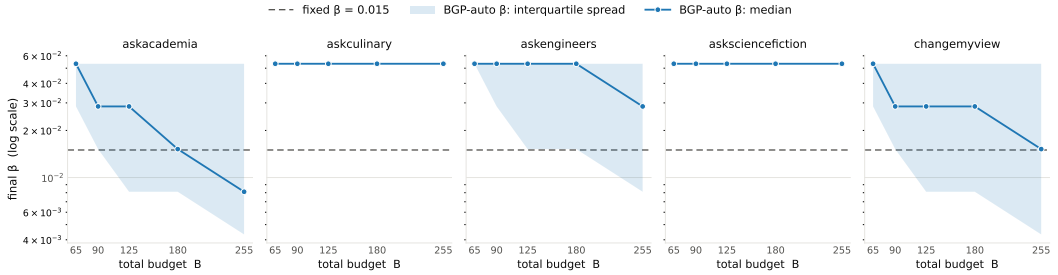

Figure 6: The converged regularization constant β by BGP with different total budget in each group. As the budget increases, the selected β values decrease more substantially in the heavily sampled groups of `askacademia`, `changemyview`, `askengineers`, and keep stable in other two groups.

Additional fairness metrics. We evaluate bottom-two reward and Nash welfare to examine whether BGP's worst-group gains extend to broader measures of group welfare. As shown in Figure 7, BGP achieves higher mean values than Greedy-BoN on both metrics across all evaluated budgets. Its bottom-two reward is close to that of BGP-Fixed and exceeds the uniform-allocation baselines at larger budgets, indicating benefits beyond the single worst group. However, uniform BoN and both fixed-β variants achieve higher Nash welfare, showing that BGP's improvements in lower-tail performance do not imply superiority under every welfare criterion.

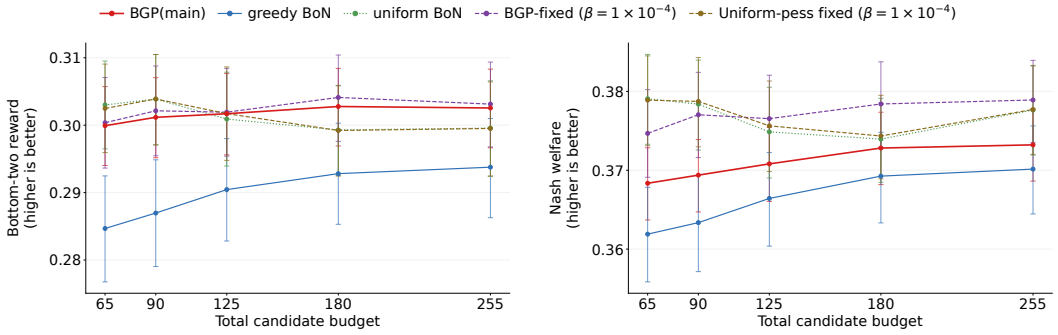

Figure 7: Additional fairness metrics under the 7B Judge across total sampling budgets B. **Left:** Bottom-two reward, the average expected reward of the two lowest-reward groups in each trial. **Right:** Nash welfare, the geometric mean of all five groups' expected rewards in each trial. Results are averaged over 200 paired trials, with error bars indicating 95% confidence intervals.