Best-of-N through the Smoothing Lens: KL Divergence and Regret Analysis

Gholamali Aminian [1] **Idan Shenfeld** [2] **Amir R. Asadi** [3] **Ahmad Beirami** **Youssef Mroueh** [4]

Abstract

A simple yet effective method for inference-time alignment of generative models is Best-of-N (BoN), where N outcomes are sampled from a reference policy, evaluated using a proxy reward model, and the highest-scoring one is selected. While prior work argues that BoN is almost optimal in reward vs KL tradeoffs, the effectiveness of BoN depends critically on the quality of the proxy reward model used for selection. For this purpose, we study BoN through a smooth version known as Soft Best-of-N (SBoN) and develop a theoretical framework to address this gap. We analyze the scaling behaviour of BoN by providing bounds on the KL divergence between the SBoN policy and the reference policy, offering insights into how performance varies with the number of samples. We also study the regret gap, i.e., the gap between the expected reward of the optimal (tilted) policy and the SBoN policy. Our theoretical and empirical findings show that smoothing helps SBoN mitigate reward overoptimization, especially when the quality of the proxy reward is low.

1. Introduction

Large language models (LLMs) have transformed machine learning, achieving state-of-the-art results on a variety of tasks. Despite all advancements, LLMs can still generate undesirable outputs, such as toxic or factually incorrect responses. This has made alignment a central goal in modern LLM development (Achiam et al., 2023; Team et al., 2023).

Several post-hoc alignment methods have been proposed to address this challenge, including Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022), SLiC (Zhao et al., 2022), Direct Preference Optimization (Rafailov et al., 2023), controlled decoding (Mudgal et al., 2024) and Best-of-N (BoN) sampling (Beirami et al., 2024). While these methods differ in their implementation—ranging from training-time optimization to test-time selection—they can be viewed, in principle, as approximating the solution to a KL-regularized reward maximization problem. The optimal solution to this problem is a tilted distribution over responses, which balances reward and proximity to the reference model (Yang et al., 2024).

In BoN as a test-time sampling strategy, given a prompt, N responses are sampled from the reference policy, and the one with the highest estimated reward is selected. Empirically, BoN has been shown to achieve competitive or superior performance in the reward-versus-KL divergence trade-off when compared to RLHF and other alignment methods (Gao et al., 2023a; Mudgal et al., 2024) under *true reward model*. Furthermore, under certain conditions, it asymptotically approximates the solution to the KL-regularized reward maximization objective (Yang et al., 2024). However, in practice, BoN relies on a learned proxy reward model—an approximation of the true reward function—to guide this selection. As such, their effectiveness critically depends on both the *proxy reward model* (estimation error of true reward) and the *quality of the reference policy*.

Understanding how these two components, the quality of the proxy reward model and the choice of reference policy, affect the alignment quality of test-time sampling algorithms is essential. There are different measures of alignment quality, including KL divergence [1] between aligned policy and reference policy and the *regret* defined as the gap between the expected reward under the optimal (tilted) policy and the alignment policy. Note that minimizing the regret gap is critical to ensuring high-quality outputs and close performance to the optimal policy. Recent work by Gao et al. (2023a) and Hilton et al. (2022) has investigated the scaling laws governing reward model optimization in both reinforcement learning (RL) and BoN settings as a function of KL divergence between aligned policy and reference policy. They empirically demonstrate that, under proxy reward models, the improvement in expected true reward, relative to a reference policy, scales proportionally for both RL and BoN

[1]The Alan Turing Institute, London, UK, [2]Massachusetts Institute of Technology, USA, [3]Statistical Laboratory, University of Cambridge, Cambridge, UK, [4]IBM Research, USA. Correspondence to: Gholamali Aminian <gaminian@turing.ac.uk>.

Workshop at the International Conference on Machine Learning (ICML) 2025.

[1]Unless stated otherwise, all KL divergences are understood to be measured between the aligned policy and the reference policy.

policies.

While recent work analyzes BoN under the idealized settings where there is no discrepancy between the proxy reward and the true reward (Yang et al., 2024; Beirami et al., 2024; Mroueh, 2024; Huang et al., 2025), our work relaxes this assumption to study the interplay between the reward discrepancy measured through regret and the KL-divergence. We present a theoretical study of **Soft Best-of-N (SBoN)**, a smoothed variant of BoN recently introduced by Claudio et al. (2025); Jinnai et al. (2024). Unlike BoN, SBoN draws the final response *probabilistically* from the N candidates, yielding a policy π_{SBoN} that is tunable with a temperature parameter. Our analysis centres on two metrics:

(a) the Kullback–Leibler divergence between SBoN policy (under the *true* reward) and the reference policy, and

(b) the *regret*, i.e. the expected reward gap between optimal policy and SBoN policy.

We show how these results specialize to the BoN (as a limit of SBoN for the temperature goes to infinity) and quantify the additional error incurred by using a *proxy* reward model instead of the true reward. Finally, we characterize regimes in which SBoN attains *lower* regret bound than BoN when we use the proxy reward model.

Our main contributions are:

- We derive finite-sample bounds for KL divergence between the SBoN policy and reference policy, and for the regret gap of the SBoN policy, and we extend these bounds to BoN. These bounds reveal how the number of responses N, proxy reward model quality and reference policy model affect performance.
- We quantify cases where SBoN performs better than BoN under the proxy reward model (a.k.a. reward model overoptimization) and quality of reference policy.
- We provide experimental validation using various proxy reward models to demonstrate SBoN's advantages in the overoptimization scenario.

2. Related Works

Best-of-N. Despite many recent advancements in alignment, a simple, popular, and well-performing method continues to be the BoN policy (Nakano et al., 2021; Stiennon et al., 2020; Beirami et al., 2024). In fact, Gao et al. (2023b); Mudgal et al. (2024); Eisenstein et al. (2023) show that BoN consistently achieves compelling win rate–KL tradeoff curves, often outperforming KL-regularized reinforcement learning and other more complex alignment strategies. LLaMA 2 (Touvron et al., 2023) leverages BoN outputs as teacher signals to further finetune the base model. Mudgal et al. (2024) extend BoN through Q-learning to block-wise BoN decoding. This empirical effectiveness has also inspired research into distilling BoN behaviour into standalone models (Amini et al., 2025; Sessa et al., 2024; Qiu et al., 2024). Hughes et al. (2024) utilize BoN as an effective method for jailbreaking, while BoN is also commonly used as a strong baseline for scaling inference-time compute (Brown et al., 2024; Snell et al., 2024). Given the broad success of BoN, we are motivated to theoretically investigate the BoN policies and the effect of the proxy reward model (reward hacking) and the quality of the reference policy.

Smoothing of Maximum. Approximating the maximum operator using a smoothed or softmax-based surrogate is a widely adopted technique in machine learning. This approach is particularly useful in settings where the hard maximum is non-differentiable or leads to unstable optimization. For instance, in robust regression, smooth approximations to the max operator are used in min-max formulations to achieve tractable optimization under distributional shifts (Wang et al., 2013; Li et al., 2023). In sequential decision-making, similar ideas appear in risk-sensitive control and Q-learning, where the softmax of Q-values leads to stochastic policies that balance exploration and exploitation (Howard and Matheson, 1972; Borkar, 2002). In convex and non-convex optimization, smoothing the maximum objective has been shown to improve convergence properties (Kort and Bertsekas, 1972; Pee and Royset, 2011; Liu and Theodorou, 2019). The Soft Best-of-N (SBoN) framework, (Claudio et al., 2025; Khanov et al., 2024; Jinnai et al., 2024), leverages this principle by replacing the hard selection of the highest-reward sample with a softmax-weighted sampling distribution. Regarding the SBoN, the empirical version of SBoN is introduced by (Khanov et al., 2024) as ARGS-stochastic, where a token from a probability distribution among the top-k candidate tokens is chosen. Then, the regularized version of BoN, which can be represented as SBoN, is discussed by (Jinnai et al., 2024). Given the broad success of SBoN, we are motivated to theoretically investigate the SBoN policies and the effect of the proxy reward model (reward hacking) and the quality of the reference policy.

Theoretical Foundation of (Soft) BoN: KL divergence of BoN is studied in (Beirami et al., 2024; Mroueh, 2024) via information theoretical tools where the KL divergence of BoN sampling from the reference distribution is bounded by $\log(N) - (N-1)/N$. Scaling laws governing reward as a function of KL divergence is empirically studied by Gao et al. (2023b) and theoretically formalized by Mroueh (2024). Furthermore, the asymptotic case and the equivalence of BoN to the KL-constrained reinforcement learning solution are studied by Yang et al. (2024) under the assumption of access to optimal reward. Gui et al. (2024) further characterized the win rate–KL gap in the asymptotic regime where a model assigns extremely low likelihoods to successful completions. Furthermore, Sun et al. (2024) accelerated BoN using speculative rejection sampling. The convergence rate of SBoN sampling to the optimal tilted distribution is studied by Claudio et al. (2025). The regret of BoN under some assumptions is studied in (Huang et al., 2025).

Theoretical Foundation of RLHF: Several works have studied the theoretical underpinnings of reverse KL-regularized RLHF, particularly in terms of sample complexity (Zhao et al., 2024; Xiong et al., 2024; Song et al., 2024; Zhan et al., 2023; Ye et al., 2024; Aminian et al., 2025). Note that, as the sampling distributions in BoN and SBoN are different, we can not apply RLHF analysis to these sampling strategies. Therefore, it is needed to develop new foundations for BoN and SBoN.

3. Problem Formulation

Notations: Upper-case letters denote random variables (e.g., Z), lower-case letters denote the realizations of random variables (e.g., z), and calligraphic letters denote sets (e.g., $\mathcal{Z}$). All logarithms are in the natural base. The set of probability distributions (measures) over a space $\mathcal{X}$ with finite variance is denoted by $\mathcal{P}(\mathcal{X})$. The KL divergence between two probability distributions on $\mathbb{R}^d$ with densities $p(x)$ and $q(x)$, such that $q(x) > 0$ when $p(x) > 0$, is $\mathrm{KL}(p\|q) := \int_{\mathbb{R}^d} p(x)\log(p(x)/q(x))\mathrm{d}x$ (with $0.\log 0 := 0$). The total-variation distance is defined as $\mathbb{TV}(p,q) = \frac{1}{2}\int_{\mathcal{X}} |p(x) - q(x)|$. Furthermore, we define chi-square divergence as $\chi^2(p(x)\|q(x)) = \int_{\mathcal{X}} \frac{p^2(x)}{q(x)} - 1$.

3.1. Preliminaries

Following Ye et al. (2024) and Zhao et al. (2024), we define inference-time alignment as an entropy-regularized optimization problem. Let the finite set of prompts be $\mathcal{X}$ and the discrete finite set of responses be $\mathcal{Y}$. Prompts are drawn from a distribution ρ over $\mathcal{X}$. A (stochastic) policy $\pi \in \Pi$ assigns, for every prompt $x \in \mathcal{X}$, a conditional distribution $\pi(\cdot \mid x)$ over $\mathcal{Y}$; drawing $y \sim \pi(\cdot \mid x)$ yields a response.

We treat the supervised-fine-tuned (SFT) model as a *reference policy*, denoted $\pi_{\mathrm{ref}}(\cdot \mid x)$. Given a true reward function $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+ \cup 0$ and temperature $\beta > 0$, we seek a policy that remains close to π_{ref} while maximizing expected reward, leading to the KL-regularized objective

$$\max_{\pi\in\Pi} \mathbb{E}_{y\sim\pi(\cdot|x)}\big[r^\star(x,y)\big] - \frac{1}{\beta}\,\mathrm{KL}\big(\pi(\cdot \mid x)\,\|\,\pi_{\mathrm{ref}}(\cdot \mid x)\big). \tag{1}$$

The unique solution is the *tilted* optimal policy (Korbak et al., 2022b;a; Yang et al., 2024)

$$\begin{aligned}\pi_{\beta,r^\star}(y|x) &= \frac{\pi_{\mathrm{ref}}(y \mid x)\,\exp\big(\beta\, r^\star(x,y)\big)}{Z_{r^\star,Y}(x,\beta)},\\ Z_{r^\star,Y}(x,\beta) &= \sum_{y\in\mathcal{Y}} \pi_{\mathrm{ref}}(y \mid x)\,\exp\big(\beta\, r^\star(x,y)\big),\end{aligned} \tag{2}$$

where $Z_{r^\star,Y}(x,\beta)$ is the normalising (*partition*) function.

Note that, in practice, we do not have access to the closed form of reference policy $\pi_{\mathrm{ref}}(y|x)$ and $r^\star(y,x)$. We can only first estimate the true reward function via proxy reward function $\hat{r}(y,x)$ using some datasets and then sample from $\pi_{\mathrm{ref}}(y|x)$ and compute $\hat{r}(y,x)$ for each individual sample. Then, for alignment, we can apply inference time algorithms, e.g., BoN or SBoN (Claudio et al., 2025), where N samples are generated from $\pi_{\mathrm{ref}}(y|x)$ and we choose the sample with the highest estimated reward (BoN) or sampled from a distribution (SBoN) using the estimated reward, $\hat{r}(y,x)$. When only a proxy reward $\hat{r}$ is available, we obtain the analogous partition function $Z_{\hat{r},Y}(x,\beta)$ and policy $\pi_{\beta,\hat{r}}(\cdot|x)$.

3.2. SBoN Algorithm

In SBoN, as introduced in (Claudio et al., 2025), the sampling strategy includes,

1. Choose N i.i.d. responses, $Y_1, \cdots, Y_n$, for a given prompt $x \in \mathcal{X}$ from $\pi_{\mathrm{ref}}(y|x)$,
2. Draw Z from $\{1, \cdots, N\}$ with softmax distribution with temperature β,
$$\Pr(Z = i) = \frac{\exp(\beta\hat{r}(Y_i,x))}{\sum_{j=1}^{N}\exp(\beta\hat{r}(Y_j,x))}, \tag{3}$$
3. Return $Y_{\mathrm{SBoN}} := Y_Z$.

We denote the final policy from SBoN via $\pi_{\hat{r}}^{(N,\beta)}(y|x)$. Note that for $\beta \to \infty$ and $\beta \to -\infty$, we recover BoN and worst-of-N (WoN) (Balashankar et al., 2025), respectively. Furthermore, for $\beta \to 0$, we recover uniform sampling among the N response samples, which is equivalent to sampling from the reference model $\pi_{\mathrm{ref}}(y|x)$. In (Claudio et al., 2025, Lemma 1), the closed form solution of SBoN is derived,

$$\pi_{\hat{r}}^{(N,\beta)}(y|x) = \frac{\pi_{\mathrm{ref}}(y|x)\exp(\beta\hat{r}(y,x))}{Z_{N,\beta}}, \tag{4}$$

where

$$Z_{N,\beta} = \mathbb{E}\Big[\Big(\frac{1}{N}\big(\exp(\beta\hat{r}(y,x)) + \sum_{i=1}^{N-1}\exp(\beta\hat{r}(Y_i,x))\big)\Big)^{-1}\Big]^{-1}. \tag{5}$$

For simplicity, we define BoN policies under true and proxy reward models as $\pi_{r^\star}^{(N,\infty)}(y|x)$ and $\pi_{\hat{r}}^{(N,\infty)}(y|x)$, respectively. In this work, we focus on $\beta \geq 0$. Similarly, we can define $\pi_{r^\star}^{(N,\beta)}(y|x)$ based on a true reward model. Another motivation for SBoN based on the Gumbel-Max trick is provided in Appendix B.

3.3. Assumptions

For our analysis, the following assumptions are needed.

Assumption 3.1 (Bounded True Reward)**.** We assume that the true and estimated reward functions, $r^\star(y,x)$ and $\hat{r}(y,x)$, are non-negative functions, bounded by $R_{\max}$.

Bounded reward assumption is also considered in previous work (Huang et al., 2025) for BoN. Furthermore, calibrated rewards, as introduced by (Beirami et al., 2024; Balashankar et al., 2025), are bounded. We can relax this assumption using a sub-gaussian or sub-exponential reward function as discussed in (Mroueh, 2024).

Let's define the tilted average of square estimation error of reward function for a given prompt x with parameter β, as follows,

$$\varepsilon_{\beta,r}(x) := \frac{1}{\beta}\log\left(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(y|x)}[e^{\beta(r^\star(Y,x)-\hat{r}(Y,x))^2}]\right). \tag{6}$$

A similar definition of estimation error is introduced in (Yang and Wibisono, 2022). When $\beta = 0$, the definition reduces to the mean-squared error, which is also introduced in (Huang et al., 2025). Letting $\beta \to \infty$ recovers the square of the supremum (infinity) norm of the estimation error between $r^\star(y,x)$ and $\hat{r}(y,x)$. The tilted average of the estimation error is monotonically increasing in β.

In addition to Assumption 3.1, for regret analysis the following assumptions are needed.

Assumption 3.2 (Bounded Estimation Error)**.** We assume that estimation error for reward function for all $x \in \mathcal{X}$ is bounded, i.e., $\varepsilon_{\beta,r}(x) < \infty$.

In practice, the estimated reward can be fitted to a human-preference labeled dataset or a dataset with samples from true rewards. As discussed in (Huang et al., 2025), for simplicity, we just assume that $\hat{r}(y,x)$ is given, where its estimation error with respect to the true reward function under reference policy is bounded.

Assumption 3.3 (Achievable maximum reward)**.** We assume that $\hat{r}(\hat{y}(x),x) = R_{\max}$ for all $\hat{y}(x) \in \hat{\mathcal{Y}}(x)$.

In many settings, the reward function attains its maximum at specific responses. In particular, since a large language model (LLM) generates outputs using a finite vocabulary and a bounded number of tokens, the space of possible generations is finite, and thus the assumption holds trivially. Next, we define overoptimization.

Definition 3.4 (Overoptimization)**.** The overoptimization happens whenever we have $\varepsilon_{\beta,r}(x) > 0$.

3.4. Optimal Policy and Coverage

We define the optimal policy under the true reward model as,

$$\pi_{r^\star}^\star(y|x) = \arg\max_{\pi} \mathbb{E}_{Y\sim\pi(\cdot|x)}[r^\star(Y,x)]. \tag{7}$$

Similarly, we can define $\pi_{\hat{r}}^\star(y|x)$ as the optimal policy under the proxy reward model.

Under bounded reward assumption, Assumption 3.1, we can interpret optimal policies as the limit of tilted optimal policies,

$$\begin{aligned}\pi_{\infty,r^\star}(\cdot|x) &:= \lim_{\beta\to\infty}\pi_{\beta,r^\star}(\cdot|x),\\ \pi_{\infty,\hat{r}}(\cdot|x) &:= \lim_{\beta\to\infty}\pi_{\beta,\hat{r}}(\cdot|x).\end{aligned} \tag{8}$$

where $\pi_{\infty,r^\star}(\cdot|x)$ and $\pi_{\infty,\hat{r}}(\cdot|x)$ place all their probability mass on the maximizers of $r^\star(y,x)$ and $\hat{r}(y,x)$, respectively. Therefore, under bounded reward assumption, we have $\pi_{r^\star}^\star(\,\cdot\,|x) = \pi_{\infty,r^\star}(\cdot|x)$ and $\pi_{\hat{r}}^\star(\,\cdot\,|x) = \pi_{\infty,\hat{r}}(\cdot|x)$. The connections between optimal, SBoN, BoN and tilted optimal policies under true or proxy reward models are shown in Figure 1.

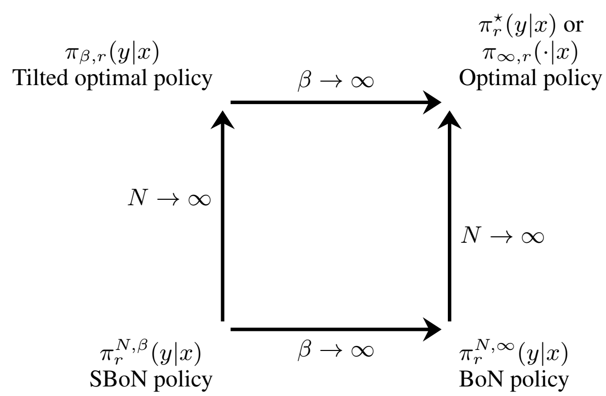

Figure 1: Connections of different policies under reward model $r \in \{\hat{r}(y,x), r^\star(y,x)\}$

Coverage: For a given reward function $r(x,y)$, we define the tilted policy (softmax policy):

$$\pi_{\beta,r}(y|x) \propto \pi_{\mathrm{ref}}(y|x)\exp(\beta r(x,y)).$$

Then, we introduce the coverage of tilted policy with respect to the reference policy as,

$$C_{\beta,r,\mathrm{ref}}(x) := \sum_{y\in\mathcal{Y}} \frac{\pi_{\beta,r}^2(y|x)}{\pi_{\mathrm{ref}}(y|x)}. \tag{9}$$

This measure $C_{\beta,r,\text{ref}}(x)$ can also be interpreted as a coverage constant, which is standard in KL-regularized policy learning. Furthermore, it can be defined as $\chi^2(\pi_{\beta,r}(y|x)\|\pi_{\text{ref}}(y|x))$. It ensures that the reference policy places sufficient probability mass on high-reward responses, thereby guaranteeing that the support of the optimal policy lies within the support of the reference. This prevents cases where optimal outputs are entirely excluded by the reference. Similar notions of coverage have been explored in Huang et al. (2025).

3.5. Optimal regret

For given policy $\pi(Y|x)$, we define expected reward with respect to the policy (a.k.a. value function[2]) as

$$J_{r^\star}(\pi(\,\cdot\,|x)) := \mathbb{E}_{Y\sim\pi(\,\cdot|x)}[r^\star(Y,x)]. \tag{10}$$

For two policies, $\pi_1(\cdot|x)$ and $\pi_2(\cdot|x)$, we define the gap between these two policies as follows,

$$\Delta_{J_{r^\star}}(\pi_1(\cdot|x),\pi_2(\cdot|x)) := J_{r^\star}(\pi_1(\cdot|x)) - J_{r^\star}(\pi_2(\cdot|x)). \tag{11}$$

We consider the following KL-Regularized objective function based on the true reward function for SBoN,

$$\begin{aligned} &J_{r^\star,\beta}(\pi_{\text{ref}}(y|x),\pi(\cdot|x)) := \\ &\quad \mathbb{E}_{Y\sim\pi(\cdot|x)}[r^\star(Y,x)] - \frac{1}{\beta}\mathrm{KL}(\pi(\cdot|x)\|\pi_{\text{ref}}(\cdot|x)). \end{aligned} \tag{12}$$

We provide an upper bound on the optimal gap of the SBoN solution, which is the gap between $\pi^\star_{r^\star}(\,\cdot\,|x)$ as the optimal policy and $\pi^{(N,\beta)}_{\hat r}(\cdot|x)$,

$$\begin{aligned} &\Delta_{J_{r^\star}}(\pi^\star_{r^\star}(\,\cdot\,|x),\pi^{(N,\beta)}_{\hat r}(\cdot|x)) \\ &= J_{r^\star}(\pi^\star_{r^\star}(\,\cdot\,|x)) - J_{r^\star}(\pi^{(N,\beta)}_{\hat r}(\cdot|x)). \end{aligned} \tag{13}$$

Regarding BoN, we consider $\pi^{(N,\infty)}_{\hat r}(\cdot|x)$ instead of $\pi^{(N,\beta)}_{\hat r}(\cdot|x)$.

Our results can be extended to the sub-optimal gap of the SBoN solution, which is the gap between $\pi_{\beta,r^\star}(\,\cdot\,|x)$ as the optimal solution to (12) and $\pi^{(N,\beta)}_{\hat r}(\cdot|x)$,

$$\begin{aligned} &\Delta_{J_{r^\star}}(\pi_{\beta,r^\star}(\,\cdot\,|x),\pi^{(N,\beta)}_{\hat r}(\cdot|x)) \\ &:= J_{r^\star}(\pi_{\beta,r^\star}(\cdot|x)) - J_{r^\star}(\pi^{(N,\beta)}_{\hat r}(\cdot|x)). \end{aligned} \tag{14}$$

4. Regret Analysis

In this section, we derive theoretical regret bounds for SBoN and BoN. First, we provide a helpful Lemma regarding the expected coverage assumption that can help us interpret the results of regret for BoN and SBoN.

[2]We can also consider $\mathbb{E}_{X\sim\rho(\cdot)}[J_{r^\star}(\pi(\cdot|X))]$. All of our results also holds for expected version of value function.

Lemma 4.1. *For a given reward function $r(x,y)\in[0,R_{\max}]$, the following holds,*

1. $C_{\beta,r,\text{ref}}(x) = \frac{\mathbb{E}[\exp(2\beta\hat r(Y,x))]}{\mathbb{E}^2[\exp(\beta\hat r(Y,x))]}$.
2. $C_{\beta,r,\text{ref}}(x)$ *is an increasing function with respect to* β.
3. $C_{\infty,r,\text{ref}}(x) := \lim_{\beta\to\infty} C_{\beta,r,\text{ref}}(x) = \frac{1}{\sum_i \pi_{\text{ref}}(y^{\max}_{i,r}(x)|x)}$ *where* $y^{\max}_{i,r}(x) \in \arg\max_y r(y,x)$.
4. *For all* $\beta<\infty$, *we have* $1\le C_{\beta,r,\text{ref}}(x)\le \min(C_{\infty,r,\text{ref}}(x),\exp(2\beta R_{\max}))$.

Now, we derive upper bound on regret of SBoN.

Theorem 4.2 (Optimal Gap of SBoN)**.** *Under the same Assumptions in Theorem E.1, there following upper bound holds on the gap of SBoN for $\beta>0$*

$$\begin{aligned} &\Delta_{J_{r^\star}}(\pi^\star_{r^\star}(\,\cdot\,|x),\pi^{(N,\beta)}_{\hat r}(\cdot|x)) \\ &\le \sqrt{\varepsilon_{\beta,r}(x)}\Big(\sqrt{C_{\infty,\hat r,\text{ref}}(x)}+\sqrt{C_{\infty,r^\star,\text{ref}}(x)}\Big) \\ &\quad + 2R_{\max}\sqrt{\frac{1}{2}\log\Big(1+\frac{C_{\infty,\hat r,\text{ref}}(x)-1}{N}\Big)} \\ &\quad + \frac{\log(C_{\infty,r^\star,\text{ref}}(x))}{\beta}, \end{aligned}$$

Regret of BoN Through Smoothing Lens: We now derive an upper bound on the regret of BoN by taking the asymptotic limit of the regret bound on optimal gap of SBoN in Theorem 4.3.

Theorem 4.3 (Optimal Gap of BoN)**.** *Assume that Assumptions 3.1, 3.3, and 3.2 hold. Then, the optimality gap of the BoN policy $\pi^{(N,\infty)}_{\hat r}(\cdot|x)$ with $\pi^\star_{r^\star}(\cdot|x)$ (given at (7)) satisfies for $\varepsilon_{\beta,r}(x)>0$:*

$$\begin{aligned} &\Delta_{J_{r^\star}}(\pi^\star_{r^\star}(\cdot|x),\pi^{(N,\infty)}_{\hat r}(\cdot|x)) \\ &\le \sqrt{\varepsilon_{\infty,r}(x)}\Big(\sqrt{C_{\infty,\hat r,\text{ref}}(x)}+\sqrt{C_{\infty,r^\star,\text{ref}}(x)}\Big) \\ &\quad + 2R_{\max}\sqrt{\frac{1}{2}\log\Big(1+\frac{C_{\infty,\hat r,\text{ref}}(x)-1}{N}\Big)}. \end{aligned}$$

Remark 4.4 (Comparison with (Huang et al., 2025)). The regret bound for BoN policy grows with the L_∞-norm of the reward-model estimation error. In contrast to the result in (Huang et al., 2025), our bound remains finite whenever the overoptimization error vanishes, i.e., when $\varepsilon_{\infty,\beta}(x)=0$ or N grows.

Remark 4.5 (Quality of reference policy). Furthermore, the bound stated in Theorem 4.3 (or Theorem 4.2) depends on the quantity,

$$C_{\infty,r^\star,\text{ref}}(x) = \frac{1}{\sum_i \pi_{\text{ref}}\big(y^{\max}_{i,r^\star}(x)\mid x\big)},$$

where $y_{i,r^\star}^{\max}(x) \in \arg\max_y r^\star(y,x)$. It can be interpreted as quality of reference policy under true reward model. Similarly, the bound in Theorem 4.3 (or Theorem 4.2) depends on the quantity,

$$C_{\infty,\hat{r},\mathrm{ref}}(x) = \frac{1}{\sum_i \pi_{\mathrm{ref}}\big(y_{i,\hat{r}}^{\max}(x) \mid x\big)},$$

where $y_{i,\hat{r}}^{\max}(x) \in \arg\max_y \hat{r}(y,x)$. It can be interpreted as quality of reference policy under proxy reward model. Therefore, the quality of reference policy under both true and proxy reward models affect the performance of BoN and SBoN policies.

Next, we compare how BoN and SBoN perform when overoptimization is present and when it is absent.

Remark 4.6 (Overoptimization). Assume that the proxy reward suffers from overoptimization, i.e. $\varepsilon_{\beta,r}(x) > 0$ for every $\beta > 0$. Letting $N \to \infty$ and invoking Theorem 4.2, we obtain

$$\begin{aligned} &\Delta_{J_{r^\star}}\big(\pi^\star_{r_\star}(\cdot \mid x), \pi_{\hat{r}}^{(\infty,\beta)}(\cdot \mid x)\big) \\ &\le \sqrt{\varepsilon_{\beta,r}(x)}\Big(\sqrt{C_{\infty,\hat{r},\mathrm{ref}}(x)} + \sqrt{C_{\infty,r^\star,\mathrm{ref}}(x)}\Big) \\ &\quad + \frac{\log C_{\infty,r^\star,\mathrm{ref}}(x)}{\beta}. \end{aligned} \tag{15}$$

Similarly, for BoN we have,

$$\begin{aligned} &\Delta_{J_{r^\star}}\big(\pi^\star_{r_\star}(\cdot \mid x), \pi_{\hat{r}}^{(\infty,\infty)}(\cdot \mid x)\big) \\ &\le \sqrt{\varepsilon_{\infty,r}(x)}\Big(\sqrt{C_{\infty,\hat{r},\mathrm{ref}}(x)} + \sqrt{C_{\infty,r^\star,\mathrm{ref}}(x)}\Big). \end{aligned} \tag{16}$$

Define the auxiliary function

$$g(\beta) = \beta\big(\varepsilon_{\infty,r}(x) - \varepsilon_{\beta,r}(x)\big), \qquad \beta \ge 0.$$

Because $g(0) = g(\infty) = 0$ and $g(\beta) \ge 0$ for all β, there exists at least one maximiser $\beta^\star \in (0,\infty)$ such that $g(\beta^\star) = \max_{\beta \ge 0} g(\beta)$.

If

$$\frac{\log C_{\infty,r_\star,\mathrm{ref}}(x)}{\sqrt{C_{\beta,\hat{r},\mathrm{ref}}(x)} + \sqrt{C_{\infty,r_\star,\mathrm{ref}}(x)}} \le g(\beta^\star), \tag{17}$$

then the upper bound in (15) does not exceed (15), and hence the SBoN bound is *tighter* than the BoN bound in the presence of overoptimization. An analogous comparison can be carried out for any fixed, finite N by retaining the N-dependent terms in the two bounds.

Remark 4.7 (No overoptimization). Assume that the overoptimization vanishes, i.e. $\varepsilon_{\beta,r}(x) = 0$ for every $\beta \in [0,\infty)$. Then the optimality gaps of the SBoN and BoN policies satisfy

$$\begin{aligned} &\Delta_{J_{r^\star}}\big(\pi^\star_{r_\star}(\cdot \mid x),\, \pi_{\hat{r}}^{(N,\beta)}(\cdot \mid x)\big) \\ &\le 2R_{\max}\sqrt{\frac{1}{2}\log\Big(1 + \frac{C_{\infty,\hat{r},\mathrm{ref}}(x) - 1}{N}\Big)} \\ &\quad + \frac{\log C_{\infty,r_\star,\mathrm{ref}}(x)}{\beta}, \end{aligned} \tag{18}$$

$$\begin{aligned} &\Delta_{J_{r^\star}}\big(\pi^\star_{r_\star}(\cdot \mid x),\, \pi_{\hat{r}}^{(N,\infty)}(\cdot \mid x)\big) \\ &\le 2R_{\max}\sqrt{\frac{1}{2}\log\Big(1 + \frac{C_{\infty,\hat{r},\mathrm{ref}}(x) - 1}{N}\Big)}. \end{aligned} \tag{19}$$

By Lemma 4.1, $C_{\infty,r_\star,\mathrm{ref}}(x) \ge 1$; consequently, the bound in (19) is tighter than the bound in (18).

5. KL divergence Analysis

Recent work by Gao et al. (2023a) and Hilton et al. (2022) empirically demonstrates that, under proxy rewards, the improvement in expected reward, relative to a reference policy, scales proportionally to $\sqrt{\mathrm{KL}(\pi_{r^\star}^{(N,\infty)} \| \pi_{\mathrm{ref}})}$ for both RL and BoN policies. This phenomenon has been studied theoretically by Beirami et al. (2024); Mroueh (2024).

In particular, Beirami et al. (2024) derives an upper bound on KL divergence for BoN policies under the assumptions of a bijective true reward mapping and a finite output space:

$$\mathrm{KL}(\pi_{r^\star}^{(N,\infty)}(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \le \log(N) - 1 + \frac{1}{N}, \tag{20}$$

Mroueh (2024) relaxes the bijectivity assumption and derives similar bounds using information-theoretic tools.

However, Gao et al. (2023a) also observes that models optimized using proxy rewards can suffer from overoptimization: as the learned policy diverges further from the reference, the alignment with the true reward may degrade. Despite theoretical advances, the KL divergence analysis for SBoN and BoN under the proxy reward model remains largely unexplored. In what follows, we establish a new upper bound on the KL divergence between the SBoN policy and the reference policy, shedding light on its behaviour as a function of the number of samples N and temperature parameter β.

Theorem 5.1. *Under Assumption 3.1, The following upper bound holds on KL divergence between SBoN and reference policies for a given prompt $x \in \mathcal{X}$,*

$$\begin{aligned} &\mathrm{KL}(\pi_{r^\star}^{(N,\beta)}(y|x) \| \pi_{\mathrm{ref}}(y|x)) \\ &\le \log\Big(\frac{N}{1 + (N-1)\exp(-\beta R_{\max})}\Big). \end{aligned} \tag{21}$$

Using Theorem 5.1, we can observe that for BoN, $\beta \to \infty$, we have,

$$\mathrm{KL}(\pi_{r^\star}^{(N,\infty)}(y|x) \| \pi_{\mathrm{ref}}(y|x)) \le \log(N). \tag{22}$$

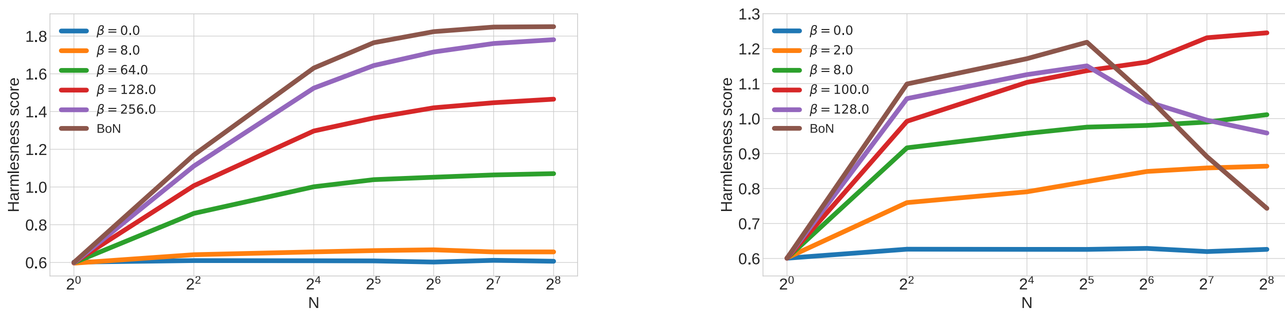

Figure 2: Soft Best-of-N experiment using a strong reward model (Left) and a weak one (Right). When the performance of the reward model is poor, BoN can lead to overoptimization, while the SBoN can help to mitigate it.

Comparing (22) with results in (Beirami et al., 2024; Mroueh, 2024), our result is derived from the SBoN asymptotic regime. Note that our bound is looser than the bound on KL divergence in (20). In contrast, our bound is general and can be applied to different β in SBoN. For $\beta = 0$, where our policy is the reference policy, our bound is tight. It is also important to note that the upper bound in Theorem 5.1 increases with the temperature parameter β.

Note that, using Pinsker's inequality, in a similar approach to (Mroueh, 2024), we have,

$$\begin{aligned}&\mathbb{E}_{Y\sim\pi_{r^\star}^{(N,\beta)}(\cdot|x)}[r^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[r^\star(Y,x)]\\&\leq R_{\max}\sqrt{\frac{1}{2}\mathrm{KL}(\pi_{r^\star}^{(N,\beta)}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))}.\end{aligned} \tag{23}$$

We aim to achieve a higher expected reward under the aligned policy for a given KL divergence between the aligned policy (BoN or SBoN) and the reference policy. It is also demonstrated by Gao et al. (2023b), the expected reward improvement of BoN policies, relative to the reference policy, scales proportionally with their KL divergence. Furthermore, (23) implies that improvement of expected reward relative to the reference policy can not exceed the square root of the KL divergence.

In Theorem 5.1, we assume that the true reward model is available. However, in practice, we utilize the proxy reward model. Therefore, we are interested in investigating the closeness of SBoN policy under the proxy reward model to SBoN under the true reward model in KL divergence.

Theorem 5.2. *Under Assumption 3.1, the following upper bound holds on performance of the SBoN under true and proxy model reward,*

$$\begin{aligned}&\mathrm{KL}(\pi_{r^\star}^{(N,\beta)}(\cdot|x)\|\pi_{\hat{r}}^{(N,\beta)}(\cdot|x))\\&\leq\frac{N\beta\sqrt{\varepsilon_{\beta,r}(x)}}{1+(N-1)\exp(-\beta R_{\max})}\Big(\frac{N\exp(2\beta R_{\max})}{(N-1)^2}+1\Big).\end{aligned} \tag{24}$$

Note that for $\beta = 0$, the upper bound in Theorem 5.2 is tight. However, this upper bound is vacuous for $\beta\to\infty$. Similar to (23), we can apply the upper bound in Theorem 5.2, to bound the difference of expected true reward with respect to $\pi_{r^\star}^{(N,\beta)}(\cdot|x)$ and $\pi_{\hat{r}}^{(N,\beta)}(\cdot|x)$, as follows,

$$\begin{aligned}&\mathbb{E}_{Y\sim\pi_{\hat{r}}^{(N,\beta)}(\cdot|x)}[r^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{r^\star}^{(N,\beta)}(\cdot|x)}[r^\star(Y,x)]\\&\leq R_{\max}\sqrt{\frac{1}{2}\mathrm{KL}(\pi_{r^\star}^{(N,\beta)}(\cdot|x)\|\pi_{\hat{r}}^{(N,\beta)}(\cdot|x))}.\end{aligned} \tag{25}$$

Using (25) and (23), we can derive an upper bound on expected reward for the SBoN policy relative to the reference policy,

$$\begin{aligned}&\underbrace{\mathbb{E}_{Y\sim\pi_{\hat{r}}^{(N,\beta)}(\cdot|x)}[r^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{r^\star}^{(N,\beta)}(\cdot|x)}[r^\star(Y,x)]}_{I_1}\\&\quad+\underbrace{\mathbb{E}_{Y\sim\pi_{r^\star}^{(N,\beta)}(\cdot|x)}[r^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[r^\star(Y,x)]}_{I_2}\\&\leq R_{\max}\sqrt{\frac{1}{2}\mathrm{KL}(\pi_{r^\star}^{(N,\beta)}(\cdot|x)\|\pi_{\hat{r}}^{(N,\beta)}(\cdot|x))}\\&\quad+R_{\max}\sqrt{\frac{1}{2}\mathrm{KL}(\pi_{r^\star}^{(N,\beta)}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))},\end{aligned} \tag{26}$$

where I_1 is the *estimation error* and I_2 measures the expected-reward gain achieved by the SBoN policy over the reference policy. Accordingly, our objective is twofold: minimise I_1 and maximize I_2.

Next, we compare BoN and SBoN under the no reward model overoptimization and reward model overoptimization.

Remark 5.3 (No overoptimization). We can observe that for a given β, if we assume $\varepsilon_{\beta,r}(x) = 0$, then we have $\mathrm{KL}(\pi_{r^\star}^{(N,\beta)}(\cdot|x)\|\pi_{\hat{r}}^{(N,\beta)}(\cdot|x)) = 0$. Therefore, due to the increasing behavior of the upper bound in Theorem 5.1 with respect to β, the KL divergence for BoN policy is higher than the SBoN policy.

Remark 5.4 (Overoptimization). When $\varepsilon_{\beta,r}(x) > 0$, we have two conflicting goals: one suggesting for fixed N that β needs to be smaller for better estimation of the true policy by the proxy reward model one given in (24), and another one suggesting a larger β to induce a better upper bound on the expected reward in (23). Hence, for a given N, there exists an optimal β to balance expected true reward maximization without leading to an overoptimization of the proxy reward that would result in a worse approximation of the SBoN policy under true reward. In that scenario, SBoN can lead to better tradeoffs than BoN. A similar discussion can be done for fixed β and varying N.

6. Empirical Evidence

To support our theoretical analysis, we conducted experiments comparing Soft Best-of-N (SBoN) across different regularization strengths and reward model qualities. We used the Olmo-2 1B model (OLMo et al., 2024) as the generator and prompts from the Attaq dataset (Kour et al., 2023). For each prompt, we generated multiple responses and selected one using SBoN with varying temperature values β. We ran two experimental conditions: one using a strong reward model (ArmoRM 8B (Wang et al., 2024)) which is close to true reward model, and another using a weaker reward model (Beaver 7B RM (Dai et al., 2023)). We use LLM-as-a-Judge (Zheng et al., 2023) as our r^*. As shown in Figure 2, when the reward model is weak, performance degrades for large N due to reward hacking. However, the smoothing in SBoN helps mitigate this degradation. This observation is also aligned with our theoretical analysis and discussion in Section 5, where under overoptimization there exists a β for a given N which outperforms BoN. For more details, see Appendix G.

7. Conclusion

In this work, we establish a theoretical foundation for alignment strategies based on Soft Best-of-N (SBoN) and Best-of-N (BoN) policies. Specifically, we derive upper bounds on the KL divergence between the aligned policy—such as SBoN or BoN—and the reference policy. We further analyze how errors in reward estimation affect performance. Notably, both our theoretical analysis and empirical results demonstrate that, under a proxy reward model where overoptimization happens, SBoN perform better than BoN under some conditions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Afra Amini, Tim Vieira, Elliott Ash, and Ryan Cotterell. Variational best-of-n alignment. In *International Conference on Learning Representations (ICLR)*, 2025.

Gholamali Aminian, Amir R Asadi, Idan Shenfeld, and Youssef Mroueh. Theoretical analysis of kl-regularized rlhf with multiple reference models. *arXiv preprint arXiv:2502.01203*, 2025.

Ananth Balashankar, Ziteng Sun, Jonathan Berant, Jacob Eisenstein, Michael Collins, Adrian Hutter, Jong Lee, Chirag Nagpal, Flavien Prost, Aradhana Sinha, Ananda Theertha Suresh, and Ahmad Beirami. Infalign: Inference-aware language model alignment. *International Conference on Machine Learning (ICML)*, 2025.

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. *International Conference on Machine Learning (ICML)*, 2024.

Vivek S Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 2002.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Clément L Canonne. A short note on an inequality between kl and tv. *arXiv preprint arXiv:2202.07198*, 2022.

Pierre Cardaliaguet, François Delarue, Jean-Michel Lasry, and Pierre-Louis Lions. *The master equation and the convergence problem in mean field games:(ams-201)*. Princeton University Press, 2019.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Mayrink Verdun Claudio, Oesterling Alex, Lakkaraju Himabindu, and P Calmon Flavio. Soft best-of-n sampling for model alignment. *URL http://github.com/alex-oesterling/soft-best-of-n/blob/main/soft_bon.pdf*, 2025.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

L Gao, J Tow, B Abbasi, S Biderman, S Black, A DiPofi, C Foster, L Golding, J Hsu, A Le Noac'h, et al. A framework for few-shot language model evaluation. *URL https://zenodo. org/records/10256836*, 7, 2023a.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023b.

Lin Gui, Cristina Gârbacea, and Victor Veitch. Bonbon alignment for large language models and the sweetness of best-of-n sampling. *arXiv preprint arXiv:2406.00832*, 2024.

Emil Julius Gumbel. Statistical theory of extreme values and some practical applications. *Nat. Bur. Standards Appl. Math. Ser. 33*, 1954.

J. Hilton, P. Clark, et al. Measuring goodhart's law: Towards an evaluation framework for open-ended generative models. https://openai.com/index/measuring-goodharts-law, 2022. Accessed: 2025-01-30.

Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Management science*, 1972.

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

John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, and Mrinank Sharma. Best-of-n jailbreaking. *arXiv preprint arXiv:2412.03556*, 2024.

Yuu Jinnai, Tetsuro Morimura, Kaito Ariu, and Kenshi Abe. Regularized best-of-n sampling to mitigate reward hacking for language model alignment. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. Args: Alignment as reward-guided search. *arXiv preprint arXiv:2402.01694*, 2024.

Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. *Advances in Neural Information Processing Systems*, 35:16203–16220, 2022a.

Tomasz Korbak, Ethan Perez, and Christopher L Buckley. Rl with kl penalties is better viewed as bayesian inference. *arXiv preprint arXiv:2205.11275*, 2022b.

Barry W Kort and Dimitri P Bertsekas. A new penalty function method for constrained minimization. In *IEEE Conference on Decision and Control and 11th Symposium on Adaptive Processes*, 1972.

George Kour, Marcel Zalmanovici, Naama Zwerdling, Esther Goldbraich, Ora Nova Fandina, Ateret Anaby-Tavor, Orna Raz, and Eitan Farchi. Unveiling safety vulnerabilities of large language models. *arXiv preprint arXiv:2311.04124*, 2023.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. On tilted losses in machine learning: Theory and applications. *Journal of Machine Learning Research*, 24(142):1–79, 2023.

Guan-Horng Liu and Evangelos A Theodorou. Deep learning theory review: An optimal control and dynamical systems perspective. *arXiv preprint arXiv:1908.10920*, 2019.

Youssef Mroueh. Information theoretic guarantees for policy alignment in large language models. *arXiv preprint arXiv:2406.05883*, 2024.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, Yaguang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. Controlled decoding from language models. In *International Conference on Machine Learning*, pages 36486–36503. PMLR, 2024.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

EY Pee and Johannes O Royset. On solving large-scale finite minimax problems using exponential smoothing. *Journal of Optimization Theory and Applications*, 2011.

Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning, 2022.

Jiahao Qiu, Yifu Lu, Yifan Zeng, Jiacheng Guo, Jiayi Geng, Huazheng Wang, Kaixuan Huang, Yue Wu, and Mengdi Wang. Treebon: Enhancing inference-time alignment with speculative tree-search and best-of-n sampling. *arXiv preprint arXiv:2410.16033*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shariari, Sarah Perrin, Abe Friesen, Geoffrey Cideron, et al. Bond: Aligning llms with best-of-n distillation. *arXiv preprint arXiv:2407.14622*, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Yuda Song, Gokul Swamy, Aarti Singh, Drew Bagnell, and Wen Sun. The importance of online data: Understanding preference fine-tuning via coverage. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. Fast best-of-n decoding via speculative rejection. *arXiv preprint arXiv:2410.20290*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *EMNLP*, 2024.

Xueqin Wang, Yunlu Jiang, Mian Huang, and Heping Zhang. Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, 2013.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.

Joy Qiping Yang, Salman Salamatian, Ziteng Sun, Ananda Theertha Suresh, and Ahmad Beirami. Asymptotics of language model alignment. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 2027–2032. IEEE, 2024.

Kaylee Yingxi Yang and Andre Wibisono. Convergence of the inexact langevin algorithm and score-based generative models in kl divergence. *arXiv preprint arXiv:2211.01512*, 2022.

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*, 2024.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline preference-based reinforcement learning. *arXiv preprint arXiv:2305.14816*, 2023.

Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for kl-regularized contextual bandits and rlhf. *arXiv preprint arXiv:2411.04625*, 2024.

Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Table 1: KL divergences between the aligned and reference policies, along with corresponding upper bounds for both SBoN and BoN policies.

KL divergence Term	Theorem / Lemma	Upper Bound
$\mathrm{KL}\big(\pi_{r^\star}^{(N,\beta)}(\cdot\|x) \,\\|\, \pi_{\mathrm{ref}}(\cdot\|x)\big)$	Theorem 5.1	$\log\Big(\frac{N}{1+(N-1)\exp(-\beta R_{\max})}\Big)$
$\mathrm{KL}(\pi_{r^\star}^{(N,\infty)}(\cdot\|x)\\|\pi_{\mathrm{ref}}(\cdot\|x))$	Theorem 3.1 in (Beirami et al., 2024) and Theorem 1 in (Mroueh, 2024)	$\log(N) - 1 + 1/N$

Table 2: KL divergences between aligned policies under true and proxy reward models, along with upper bounds for SBoN and tilted policies.

KL divergence Term	Theorem / Lemma	Upper Bound
$\mathrm{KL}\big(\pi_{r^\star}^{(N,\beta)}(\cdot\|x) \,\\|\, \pi_{\hat{r}}^{(N,\beta)}(\cdot\|x)\big)$	Theorem 5.2	$\frac{N\beta\sqrt{\varepsilon_{\beta,r}(x)}}{1+(N-1)\exp(-\beta R_{\max})}\Big(\frac{N\exp(2\beta R_{\max})}{(N-1)^2}+1\Big)$
$\mathrm{KL}\big(\pi_{\beta,r^\star}(\cdot\|x) \,\\|\, \pi_{\beta,\hat{r}}(\cdot\|x)\big)$	Lemma C.8	$2\beta\sqrt{\varepsilon_{\beta,r}(x)}\Big(\sqrt{\frac{\mathbb{E}[\exp(2\beta\hat{r}(Y,x))]}{\mathbb{E}^2[\exp(\beta\hat{r}(Y,x))]}}+\sqrt{\frac{\mathbb{E}[\exp(2\beta r^\star(Y,x))]}{\mathbb{E}^2[\exp(\beta r^\star(Y,x))]}}\Big)$

A. Summary of KL divergence Results

In Table 1 , we summarize results on KL divergences between the aligned and reference policies, along with corresponding upper bounds for both SBoN and BoN policies. Furthermore, in Table 2, we summarize results on KL divergences between aligned policies under true and proxy reward models, along with upper bounds for SBoN and tilted policies.

B. Gumbel–Max trick

We also provide an interpretation for SBoN from Gumbel-Max trick. An alternative way to sample Z from

$$\Pr(Z = i) \;\propto\; \exp\big(\beta\,\hat{r}(Y_i, x)\big)$$

is via the Gumbel–Max trick. We can draw independent Gumbel distributed random variables $G_i \sim \mathrm{Gumbel}(0,1), i = 1,\dots,n$, and then set

$$Z = \arg\max_{i\in\{1,\dots,N\}}\Big[\hat{r}(Y_i, x) + \frac{G_i}{\beta}\Big].$$

By properties of the Gumbel distribution, this yields exactly the same softmax sampling law, without needing to compute the normalizing factor $\sum_{j=1}^{N}\exp(\beta\hat{r}(Y_j, x))$ explicitly (Gumbel, 1954). Obviously, when $\beta\to\infty$, the effect of the Gumbel noises vanish and the sampling strategy reduces to BoN.

C. Technical Tools

We denote the set maximizers of the estimated reward via $\hat{\mathcal{Y}}(x) = \{\hat{y}_j(x)\}_{j=1}^{m(x)}$.

We introduce the functional derivative, see (Cardaliaguet et al., 2019).

Definition C.1. (Cardaliaguet et al., 2019) A functional $U : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ admits a functional derivative if there is a map $\frac{\delta U}{\delta m} : \mathcal{P}(\mathbb{R}^n)\times\mathbb{R}^n \to \mathbb{R}$ which is continuous on $\mathcal{P}(\mathbb{R}^n)$ and, for all $m, m' \in \mathcal{P}(\mathbb{R}^n)$, it holds that

$$U(m') - U(m) = \int_0^1\int_{\mathbb{R}^n}\frac{\delta U}{\delta m}(m_\beta, a)\,(m'-m)(da)\,\mathrm{d}\beta,$$

where $m_\beta = m + \beta(m'-m)$.

Definition C.2 (Sensitivity of a policy)**.** We also define the sensitivity of a policy $\pi_r(y|x)$, which is a function of reward function $r(x,y)$, with respect to the reward function as

$$\frac{\partial \pi}{\partial r}(r) := \lim_{\Delta r \to 0} \frac{\pi_r(y|x) - \pi_{r+\Delta r}(y|x)}{\Delta r}. \tag{27}$$

Lemma C.3 (Kantorovich-Rubenstein duality of total variation distance, see (Polyanskiy and Wu, 2022))**.** *The Kantorovich-Rubenstein duality (variational representation) of the total variation distance is as follows:*

$$\mathbb{TV}(m_1, m_2) = \frac{1}{2L} \sup_{g \in \mathcal{G}_L} \left\{ \mathbb{E}_{Z \sim m_1}[g(Z)] - \mathbb{E}_{Z \sim m_2}[g(Z)] \right\}, \tag{28}$$

where $\mathcal{G}_L = \{g : \mathcal{Z} \to \mathbb{R}, ||g||_\infty \leq L\}$.

Lemma C.4 (Lemma 5.4 in (Aminian et al., 2025))**.** *Consider the softmax policy,* $\pi_r^\beta(y|x) \propto \pi_{\text{ref}}(y|x)\exp(\beta r(x,y))$. *Then, the sensitivity of the policy with respect to reward function is,*

$$\frac{\partial \pi_r^\beta}{\partial r}(r) = \beta \pi_r^\beta(y|x)(1 - \pi_r^\beta(y|x)).$$

Lemma C.5 (Pinsker's Inequality (Canonne, 2022))**.** *For* m_1 *and* m_2*, we have,*

$$\mathbb{TV}(m_1, m_2) \leq \sqrt{\frac{1}{2}\mathrm{KL}(m_2 \| m_1)}\,. \tag{29}$$

The following Lemmata are useful for our technical proofs.

Lemma C.6. *Under Assumption 3.2, the following upper bound holds,*

$$\log\left(\frac{Z_{r^\star,Y}(x,\beta)}{Z_{\hat{r},Y}(x,\beta)}\right) \leq \beta\sqrt{\varepsilon_{\beta,r}(x)}\sqrt{C_{\beta,r^\star,\text{ref}}(x)}. \tag{30}$$

Proof.

$$\begin{aligned}
\frac{Z_{\hat{r},Y}(x,\beta)}{Z_{r^\star,Y}(x,\beta)} &= \frac{\sum_{\mathcal{Y}} \exp(\beta \hat{r}(y,x))\pi_{\text{ref}}(y|x)}{\sum_{\mathcal{Y}} \exp(\beta r^\star(y,x))\pi_{\text{ref}}(y|x)} \\
&= \frac{\sum_{\mathcal{Y}} \exp(\beta(\hat{r}(y,x) - r^\star(y,x)))\exp(\beta r^\star(y,x))\pi_{\text{ref}}(y|x)}{\sum_{\mathcal{Y}} \exp(\beta r^\star(y,x))\pi_{\text{ref}}(y|x)} \\
&= \sum_{\mathcal{Y}} \pi_{\beta,r^\star}(y|x)\exp(\beta(\hat{r}(y,x) - r^\star(y,x)))
\end{aligned} \tag{31}$$

Due to convexity of $-\log(\cdot)$ and using Cauchy–Schwarz inequality, we have,

$$\begin{aligned}
-\log\left(\frac{Z_{\hat{r},Y}(x,\beta)}{Z_{r^\star,Y}(x,\beta)}\right) &\leq \beta \sum_{\mathcal{Y}} \pi_{\beta,r^\star}(y|x)(r^\star(y,x) - \hat{r}(y,x)) \\
&\leq \beta \sum_{\mathcal{Y}} \frac{\pi_{\beta,r^\star}(y|x)}{\pi_{\text{ref}}(y|x)}(r^\star(y,x) - \hat{r}(y,x))\pi_{\text{ref}}(y|x) \\
&\leq \beta \sqrt{\sum_{\mathcal{Y}} (r^\star(y,x) - \hat{r}(y,x))^2 \pi_{\text{ref}}(y|x)} \sqrt{\sum_{\mathcal{Y}} \frac{\pi^2_{\beta,r^\star}(y|x)}{\pi_{\text{ref}}(y|x)}} \\
&= \sqrt{\beta} \sqrt{\sum_{\mathcal{Y}} \log\left(\exp\left(\beta(r^\star(y,x) - \hat{r}(y,x))^2\right)\right)\pi_{\text{ref}}(y|x)} \sqrt{C_{\beta,r^\star,\text{ref}}(x)} \\
&\leq \beta \sqrt{\frac{1}{\beta}\log\left(\sum_{\mathcal{Y}} \exp\left(\beta(r^\star(y,x) - \hat{r}(y,x))^2\right)\pi_{\text{ref}}(y|x)\right)} \sqrt{C_{\beta,r^\star,\text{ref}}(x)} \\
&= \beta\sqrt{\varepsilon_{\beta,r}(x)}\sqrt{C_{\beta,r^\star,\text{ref}}(x)},
\end{aligned} \tag{32}$$

□

Lemma C.7. *Under Assumption 3.1, we have,*

$$\mathrm{KL}(\pi^{\star}_{r^\star}(\,\cdot\,|x)\|\pi_{\mathrm{ref}}(\cdot|x)) \leq \log(C_{\infty,r^\star,\mathrm{ref}}(x)) \tag{33}$$

Proof. Note that, under Assumption 3.1, we have,

$$\begin{aligned}\mathrm{KL}(\pi^{\star}_{r^\star}(\,\cdot\,|x)\|\pi_{\mathrm{ref}}(\cdot|x)) &\leq \log\Big(\mathbb{E}_{Y\sim\pi^{\star}_{r^\star}(\cdot|x)}\Big[\frac{\pi^{\star}_{r^\star}(\,\cdot\,|x)}{\pi_{\mathrm{ref}}(\cdot|x)}\Big]\Big)\\ &\leq \log(C_{\infty,r^\star,\mathrm{ref}}(x))\end{aligned} \tag{34}$$

□

Lemma C.8. *Under Assumption 3.1 and Assumption 3.2, the following upper bound holds,*

$$\mathrm{KL}(\pi_{\beta,r^\star}(y|x)\|\pi_{\beta,\hat{r}}(y|x)) \leq \beta\sqrt{\varepsilon_{\beta,r}(x)}\big(\sqrt{C_{\beta,r^\star,\mathrm{ref}}(x)}+\sqrt{C_{\beta,\hat{r},\mathrm{ref}}(x)}\big). \tag{35}$$

Proof.

$$\begin{aligned}\mathrm{KL}(\pi_{\beta,r^\star}(y|x)\|\pi_{\beta,\hat{r}}(y|x)) &= \sum_{\mathcal{Y}}\pi_{\beta,r^\star}(y|x)\log\Big(\frac{\pi_{\beta,r^\star}(y|x)}{\pi_{\beta,\hat{r}}(y|x)}\Big)\\ &= \beta\sum_{\mathcal{Y}}(r^\star(y,x)-\hat{r}(y,x))\pi_{\beta,r^\star}(y|x)+\log(Z_{\hat{r},Y}(x,\beta)/Z_{r^\star,Y}(x,\beta))\\ &\leq \beta\sqrt{\varepsilon_{\beta,r}(x)}\big(\sqrt{C_{\beta,r^\star,\mathrm{ref}}(x)}+\sqrt{C_{\beta,\hat{r},\mathrm{ref}}(x)}\big),\end{aligned} \tag{36}$$

where the final inequality holds due to Lemma C.6 and applying Cauchy–Schwarz inequality. □

Lemma C.9. *Suppose that $f(Z)\in[0,Z_{\max}]$, $\mathcal{Z}_{\max}=\{z_{m,i}\}_{i=1}^m$ is the set of miximizers of $f(Z)$, i.e., $f(z)=Z_{\max}$ for $z\in\mathcal{Z}_{\max}$. Then we have,*

$$\lim_{\beta\to\infty}\frac{\mathbb{E}[\exp(2\beta f(Z))]}{\mathbb{E}[\exp(\beta f(Z))]^2}=\frac{1}{\sum_{z\in\mathcal{Z}_{\max}}P(Z=z)}. \tag{37}$$

Proof.

$$\frac{\mathbb{E}[\exp(2\beta f(Z))]}{\mathbb{E}[\exp(\beta f(Z))]^2}=\frac{\mathbb{E}[\exp(2\beta(f(Z)-Z_{\max}))]}{\mathbb{E}[\exp(\beta(f(Z)-Z_{\max}))]^2} \tag{38}$$

$$\frac{\sum_j P(Z=z_j)\exp(2\beta(f(z_j)-Z_{\max}))}{(\sum_j P(Z=z_j)\exp(\beta(f(z_j)-Z_{\max})))^2} \tag{39}$$

Now, we have,

$$\lim_{\beta\to\infty}\frac{\mathbb{E}[\exp(2\beta f(Z))]}{\mathbb{E}[\exp(\beta f(Z))]^2} \tag{40}$$

$$=\lim_{\beta\to\infty}\frac{\sum_j P(Z=z_j)\exp(2\beta(f(z_j)-Z_{\max}))}{(\sum_j P(Z=z_j)\exp(\beta(f(z_j)-Z_{\max})))^2} \tag{41}$$

$$=\frac{\sum_{z\in\mathcal{Z}_{\max}}P(Z=z)}{(\sum_{z\in\mathcal{Z}_{\max}}P(Z=z))^2} \tag{42}$$

$$=\frac{1}{\sum_{z\in\mathcal{Z}_{\max}}P(Z=z)}, \tag{43}$$

where we used the fact that $\lim_{\beta\to\infty}\exp(\beta(z_j-Z_{\max}))=0$ for $z_j<Z_{\max}$. □

Lemma C.10 (Theorem 1 in (Claudio et al., 2025))**.** *For $\beta>0$, and $N\geq 1$, we have,*

$$\mathrm{KL}(\pi_{\beta,r^\star}(\cdot|x)\|\pi^{(N,\beta)}_{r^\star}(y|x)) \leq \log(1+\frac{C_{\beta,r^\star,\mathrm{ref}}(x)}{N}). \tag{44}$$

Lemma C.11. *Under Assumption 3.1, we have,*

$$\Big|\frac{f(r)}{\delta r}\Big| \leq \frac{N^2\beta\exp(2\beta R_{\max})}{(N-1)^2}, \tag{45}$$

where $f(r) = \log\big(\mathbb{E}[\frac{1}{\exp(\beta r)+\sum_{i=1}^{N-1}\exp(\beta R_i)}]\big)$, $r = r(x,y)$ *and* $R_i = r(Y_i, x)$.

Proof. Note that $\{R_i\}_{i=1}^{N-1}$ are i.i.d. . Therefore, we have,

$$\begin{aligned}
\frac{\delta f(r(y,x))}{\delta r} &= \mathbb{E}[\frac{1}{\exp(\beta r)+\sum_{i=1}^{N-1}\exp(\beta R_i)}]^{-1}\frac{\delta \mathbb{E}[\frac{1}{\exp(\beta r)+\sum_{i=1}^{N-1}\exp(\beta R_i)}]}{\delta r}\\
&\leq \mathbb{E}[\frac{1}{\exp(\beta r)+\sum_{i=1}^{N-1}\exp(\beta R_i)}]^{-1}\\
&\quad\times\Big(\sum_{k=1}^{N}\frac{\beta k\binom{N-1}{k-1}\exp(\beta r)}{(k\exp(\beta r)+N-1-k)^2}(1-P(R=r))^{N-k}P^{k-1}(R=r)\Big)\\
&\leq \mathbb{E}[\frac{1}{\exp(\beta r)+\sum_{i=1}^{N-1}\exp(\beta R_i)}]^{-1}\\
&\quad\times\frac{\beta\exp(\beta R_{\max})}{(N-1)^2}\Big(\sum_{k=1}^{N}k\binom{N-1}{k-1}(1-P(R=r))^{N-k}P^{k-1}(R=r)\Big)\\
&\leq \frac{N\beta\exp(2\beta R_{\max})}{(N-1)^2}(1+(N-1)P(R=r))\\
&\leq \frac{N^2\beta\exp(2\beta R_{\max})}{(N-1)^2}.
\end{aligned} \tag{46}$$

□

D. Proof and Details of Section 5

Theorem 5.1. Under Assumption 3.1, The following upper bound holds on KL divergence between SBoN and reference policies for a given prompt $x \in \mathcal{X}$,

$$\mathrm{KL}(\pi_{r^\star}^{(N,\beta)}(y|x)\|\pi_{\mathrm{ref}}(y|x)) \leq \log\Big(\frac{N}{1+(N-1)\exp(-\beta R_{\max})}\Big). \tag{47}$$

Proof. Recall that,

$$\pi_{r^\star}^{(N,\beta)}(y|x) = \pi_{\mathrm{ref}}(y|x)\exp(\beta r^\star(y,x))\mathbb{E}\Big[\Big(\frac{1}{N}(\exp(\beta r^\star(y,x)) + \sum_{i=1}^{N-1}\exp(\beta r^\star(Y_i,x)))\Big)^{-1}\Big].$$

Now, we have,

$$\begin{aligned}
&\mathrm{KL}(\pi_{r^\star}^{(N,\beta)}(y|x)\|\pi_{\mathrm{ref}}(y|x))\\
&=\sum_{\mathcal{Y}}\pi_{r^\star}^{(N,\beta)}(y|x)\log(\pi_{r^\star}^{(N,\beta)}(y|x)/\pi_{\mathrm{ref}}(y|x))\\
&=\sum_{\mathcal{Y}}\pi_{r^\star}^{(N,\beta)}(y|x)\log(N)+\sum_{\mathcal{Y}}\pi_{r^\star}^{(N,\beta)}(y|x)\log(\mathbb{E}\Big[\Big(\exp(\beta r^\star(y,x))+\sum_{i=1}^{N-1}\exp(\beta r^\star(Y_i,x))\Big)^{-1}\Big])\\
&=\log(N)+\sum_{\mathcal{Y}}\pi_{r^\star}^{(N,\beta)}(y|x)\log(\mathbb{E}\Big[\Big(1+\sum_{i=1}^{N-1}\exp(\beta(r^\star(Y_i,x)-r^\star(y,x)))\Big)^{-1}\Big]),
\end{aligned} \tag{48}$$

For the second term in (48), consider

$$A(y,Y,x)=\sum_{i=1}^{N-1}\exp(\beta(r^\star(Y_i,x)-r^\star(y,x)))>0,$$

where under Assumption 3.1, we have

$$(N-1)\exp(-\beta R_{\max})\leq A(y,Y,x)\leq (N-1)\exp(\beta R_{\max}).$$

Therefore, we have,

$$\begin{aligned}
&\sum_{\mathcal{Y}}\pi_{r^\star}^{(N,\beta)}(y|x)\log(\mathbb{E}\Big[\big(1+\sum_{i=1}^{N-1}\exp(\beta(r^\star(Y_i,x)-r^\star(y,x)))\big)^{-1}\Big])\\
&\leq\sum_{\mathcal{Y}}\pi_{r^\star}^{(N,\beta)}(y|x)\log(\frac{1}{1+(N-1)\exp(-\beta R_{\max})})\\
&=\log(\frac{1}{1+(N-1)\exp(-\beta R_{\max})}).
\end{aligned}\tag{49}$$

Combining (49) with (48) completes the proof.

□

Theorem 5.2. Under Assumption 3.1, the following upper bound holds on performance of SBoN under true and proxy model reward,

$$\begin{aligned}
&\mathrm{KL}(\pi_{r^\star}^{(N,\beta)}(\cdot|x)\|\pi_{\hat{r}}^{(N,\beta)}(\cdot|x))\\
&\leq\frac{N\beta\sqrt{\varepsilon_{\beta,r}(x)}}{1+(N-1)\exp(-\beta R_{\max})}\Big(\frac{N\exp(2\beta R_{\max})}{(N-1)^2}+1\Big).
\end{aligned}\tag{50}$$

Proof. We first provide the following upper bound,

$$\begin{aligned}
&\mathrm{KL}(\pi_{r^\star}^{(N,\beta)}(y|x)\|\pi_{\hat{r}}^{(N,\beta)}(y|x))\\
&=\sum_{\mathcal{Y}}\pi_{r^\star}^{(N,\beta)}(y|x)\log\big(\frac{\pi_{r^\star}^{(N,\beta)}(y|x)}{\pi_{\hat{r}}^{(N,\beta)}(y|x)}\big)\\
&=\sum_{\mathcal{Y}}\pi_{r^\star}^{(N,\beta)}(y|x)\beta(r^\star(y,x)-\hat{r}(y,x))\\
&\quad+\sum_{\mathcal{Y}}\pi_{r^\star}^{(N,\beta)}(y|x)\Big(\log\big(\mathbb{E}[\frac{1}{\exp(\beta r^\star(y,x))+\sum_{i=1}^{N-1}\exp(\beta r^\star(Y_i,x))}]\big)\\
&\qquad-\log\big(\mathbb{E}[\frac{1}{\exp(\beta\hat{r}(y,x))+\sum_{i=1}^{N-1}\exp(\beta\hat{r}(Y_i,x))}]\big)\Big)\\
&\leq\frac{N\beta\sqrt{\varepsilon_{\beta,r}(x)}}{1+(N-1)\exp(-\beta R_{\max})}\\
&\quad+\sum_{\mathcal{Y}}\pi_{r^\star}^{(N,\beta)}(y|x)\Big(\log\big(\mathbb{E}[\frac{1}{\exp(\beta r^\star(y,x))+\sum_{i=1}^{N-1}\exp(\beta r^\star(Y_i,x))}]\big)\\
&\qquad-\log\big(\mathbb{E}[\frac{1}{\exp(\beta\hat{r}(y,x))+\sum_{i=1}^{N-1}\exp(\beta\hat{r}(Y_i,x))}]\big)\Big).
\end{aligned}\tag{51}$$

Note that for the last term in (51), we can apply mean-value theorem as follows,

$$\begin{aligned}
&\sum_{\mathcal{Y}} \pi_{r^\star}^{(N,\beta)}(y|x)\Big(\log\big(\mathbb{E}[\frac{1}{\exp(\beta r^\star(y,x)) + \sum_{i=1}^{N-1}\exp(\beta r^\star(Y_i,x))}]\big) \\
&\qquad - \log\big(\mathbb{E}[\frac{1}{\exp(\beta \hat{r}(y,x)) + \sum_{i=1}^{N-1}\exp(\beta \hat{r}(Y_i,x))}]\big)\Big) \\
&\leq \sum_{\mathcal{Y}} \pi_{r^\star}^{(N,\beta)}(y|x) |r^\star(y,x) - \hat{r}(y,x)| \Big|\frac{\delta f(r_\gamma(y,x))}{\delta r}\Big|,
\end{aligned} \tag{52}$$

where $f(r_\gamma(y,x)) = \log\big(\mathbb{E}[\frac{1}{\exp(\beta r_\gamma(y,x)) + \sum_{i=1}^{N-1}\exp(\beta r_\gamma(Y_i,x))}]\big)$, for some $\gamma \in (0,1)$ we have $r_\gamma(y,x) = \gamma\hat{r}(y,x) + (1-\gamma)r^\star(y,x)$. Using Lemma C.11, we have,

$$\Big|\frac{\delta f(r(y,x))}{\delta r}\Big| \leq \frac{N^2\beta\exp(2\beta R_{\max})}{(N-1)^2}. \tag{53}$$

Using (53) in (52) and applying Cauchy–Schwarz inequality, we have,

$$\begin{aligned}
&\sum_{\mathcal{Y}} \pi_{r^\star}^{(N,\beta)}(y|x)\Big(\log\big(\mathbb{E}[\frac{1}{\exp(\beta r^\star(y,x)) + \sum_{i=1}^{N-1}\exp(\beta r^\star(Y_i,x))}]\big) \\
&\qquad - \log\big(\mathbb{E}[\frac{1}{\exp(\beta \hat{r}(y,x)) + \sum_{i=1}^{N-1}\exp(\beta \hat{r}(Y_i,x))}]\big)\Big) \\
&\leq \sqrt{\sum_{\mathcal{Y}} \mathbb{E}[1/(1+\sum_{i=1}^{N-1}\exp(\beta(r^\star(Y_i,x) - r^\star(y,x))))]^2 \pi_{\mathrm{ref}}(y|x)} \\
&\quad \times \sqrt{\sum_{\mathcal{Y}} |r^\star(y,x) - \hat{r}(y,x)|^2 \pi_{\mathrm{ref}}(y|x)} \sqrt{\sum_{\mathcal{Y}} \Big|\frac{\delta f(r_\gamma(y,x))}{\delta r}\Big|^2 \pi_{\mathrm{ref}}(y|x)} \\
&\leq \frac{\sqrt{\varepsilon_{\beta,r}(x)}}{1 + (N-1)\exp(-\beta R_{\max})} \frac{N^2\beta\exp(2\beta R_{\max})}{(N-1)^2}.
\end{aligned} \tag{54}$$

It completes the proof. □

E. Proof and details of Section 4

Lemma 4.1. For a given reward function $r(x, y) \in [0, R_{\max}]$, the following holds,

1. $C_{\beta,r,\mathrm{ref}}(x) = \frac{\mathbb{E}[\exp(2\beta\hat{r}(Y,x))]}{\mathbb{E}^2[\exp(\beta\hat{r}(Y,x))}$.
2. $C_{\beta,r,\mathrm{ref}}(x)$ is an increasing function with respect to β.
3. $C_{\infty,r,\mathrm{ref}}(x) := \lim_{\beta\to\infty} C_{\beta,r,\mathrm{ref}}(x) = \frac{1}{\sum_i \pi_{\mathrm{ref}}(y_{\max,i}(x)|x)}$ where $y_{\max,i} \in \arg\max r(y_{\max,i}, x)$.
4. For all $\beta < \infty$, we have $1 \leq C_{\beta,r,\mathrm{ref}}(x) \leq \min(C_{\infty,r,\mathrm{ref}}(x), \exp(2\beta R_{\max}))$.

Proof. 1.

$$\begin{aligned} C_{\beta,\hat{r},\mathrm{ref}}(x) &= \sum_{\mathcal{Y}} \frac{\pi^2_{\beta,\hat{r}}(y|x)}{\pi_{\mathrm{ref}}(y|x)} \\ &= \sum_{\mathcal{Y}} \frac{\exp(2\beta\hat{r}(y,x))}{\mathbb{E}^2[\exp(\beta\hat{r}(Y,x))]}\pi_{\mathrm{ref}}(y|x) \\ &= \frac{\mathbb{E}[\exp(2\beta\hat{r}(Y,x))]}{\mathbb{E}^2[\exp(\beta\hat{r}(Y,x))]}. \end{aligned} \tag{55}$$

2. We can show that the logarithm function of $C_{\beta,\hat{r},\mathrm{ref}}(x)$ is increasing. Then, due to increasing feature of log function, the final result holds.

$$\begin{aligned} &\log(\frac{\mathbb{E}[\exp(2\beta\hat{r}(Y,x))]}{\mathbb{E}^2[\exp(\beta\hat{r}(Y,x))]}) \\ &= \log(\mathbb{E}[\exp(2\beta\hat{r}(Y,x))]) - 2\log(\mathbb{E}[\exp(\beta\hat{r}(Y,x))]), \end{aligned} \tag{56}$$

then we can compute the derivative of (56),

$$\begin{aligned} &\frac{\mathrm{d}\log(\mathbb{E}[\exp(2\beta\hat{r}(Y,x))])}{\mathrm{d}\beta} - 2\frac{\mathrm{d}\log(\mathbb{E}[\exp(\beta\hat{r}(Y,x))]}{\mathrm{d}\beta} \\ &= \frac{\mathbb{E}[2\hat{r}(Y,x)\exp(2\beta\hat{r}(Y,x))]}{\mathbb{E}[\exp(2\beta\hat{r}(Y,x))]} - 2\frac{\mathbb{E}[\hat{r}(Y,x)\exp(\beta\hat{r}(Y,x))]}{\mathbb{E}[\exp(\beta\hat{r}(Y,x))]} \end{aligned} \tag{57}$$

Note that we have,

$$\begin{aligned} \frac{\mathrm{d}\frac{\mathbb{E}[\hat{r}(Y,x)\exp(\beta\hat{r}(Y,x))]}{\mathbb{E}[\exp(\beta\hat{r}(Y,x))]}}{\mathrm{d}\beta} &= \frac{\mathbb{E}[\hat{r}^2(Y,x)\exp(\beta\hat{r}(Y,x))]\mathbb{E}[\exp(\beta\hat{r}(Y,x))] - \mathbb{E}[\hat{r}(Y,x)\exp(\beta\hat{r}(Y,x))]^2}{\mathbb{E}^2[\exp(\beta\hat{r}(Y,x))]} \\ &= \mathbb{E}_{Y\sim\pi_{\beta,\hat{r}}(\cdot|x)}[\hat{r}^2(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,\hat{r}}(\cdot|x)}[\hat{r}(Y,x)]^2 \geq 0. \end{aligned} \tag{58}$$

Therefore, we have,

$$\frac{\mathbb{E}[\hat{r}(Y,x)\exp(2\beta\hat{r}(Y,x))]}{\mathbb{E}[\exp(2\beta\hat{r}(Y,x))]} \geq \frac{\mathbb{E}[\hat{r}(Y,x)\exp(\beta\hat{r}(Y,x))]}{\mathbb{E}[\exp(\beta\hat{r}(Y,x))]}. \tag{59}$$

It completes the proof.

3. Follows directly from Lemma C.9.

4. Due to Jensen inequality for $\mathbb{E}^2[\exp(\beta\hat{r}(Y,x))] \leq \mathbb{E}[\exp(2\beta\hat{r}(Y,x))]$, the $C_{\beta,\hat{r},\mathrm{ref}}(x)$. We also have the uniform bound, $C_{\beta,\hat{r},\mathrm{ref}}(x) = \frac{\mathbb{E}[\exp(2\beta\hat{r}(Y,x))]}{\mathbb{E}^2[\exp(\beta\hat{r}(Y,x))}\leq \exp(\beta R_{\max})$. Furthermore, due to increasing property in (2), we also have $\sup_\beta C_{\beta,\hat{r},\mathrm{ref}}(x) = C_{\infty,r,\mathrm{ref}}(x)$. Therefore, the upper bound holds.

□

Theorem E.1. *Under Assumption 3.1, the following upper bound holds on the sub-optimality gap of the SBoN,*

$$\Delta_{J_{r^\star}}(\pi_{\beta,r^\star}(\,\cdot\,|x),\pi_{\hat{r}}^{(N,\beta)}(\cdot|x)) \le \frac{1}{\beta}\Big(\mathrm{KL}(\pi_{\beta,r^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,\hat{r}}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\Big)$$
$$+\sqrt{\varepsilon_{\beta,r}(x)}\Big(\sqrt{C_{\beta,\hat{r},\mathrm{ref}}(x)}+\sqrt{C_{\beta,r^\star,\mathrm{ref}}(x)}\Big)$$
$$+2R_{\max}\sqrt{\frac{1}{2}\log\Big(1+\frac{C_{\beta,\hat{r},\mathrm{ref}}(x)-1}{N}\Big)}.$$

Proof. Note that, we have,

$$\begin{aligned}
&\Delta_{J_{r^\star}}(\pi_{\beta,r^\star}(\,\cdot\,|x),\pi_{\hat{r}}^{(N,\beta)}(\cdot|x))\\
&=\mathbb{E}_{Y\sim\pi_{\beta,r^\star}(\cdot|x)}[r^\star(Y,x)]-\mathbb{E}_{Y\sim\pi_{\hat{r}}^{(N,\beta)}(\cdot|x)}[r^\star(Y,x)]\\
&=\underbrace{\mathbb{E}_{Y\sim\pi_{\beta,r^\star}(\cdot|x)}[r^\star(Y,x)]-\mathbb{E}_{Y\sim\pi_{\beta,\hat{r}}(\cdot|x)}[r^\star(Y,x)]}_{I_1}\\
&\quad+\underbrace{\mathbb{E}_{Y\sim\pi_{\beta,\hat{r}}(\cdot|x)}[r^\star(Y,x)]-\mathbb{E}_{Y\sim\pi_{\hat{r}}^{(N,\beta)}(\cdot|x)}[r^\star(Y,x)]}_{I_2}
\end{aligned}\tag{60}$$

Note that, using the definition of $\pi_{\beta,r^\star}(\,\cdot\,|x)$ and $\pi_{\beta,\hat{r}}(\,\cdot\,|x)$ as solutions to KL-regularized problem, we have,

$$\mathbb{E}_{Y\sim\pi_{\beta,r^\star}(\cdot|x)}[r^\star(Y,x)]=\frac{1}{\beta}\mathrm{KL}(\pi_{\beta,r^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))+\frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta r^\star(Y,x))]).\tag{61}$$

$$\mathbb{E}_{Y\sim\pi_{\beta,\hat{r}}(\cdot|x)}[\hat{r}(Y,x)]=\frac{1}{\beta}\mathrm{KL}(\pi_{\beta,\hat{r}}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))+\frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta\hat{r}(Y,x))]).\tag{62}$$

Therefore, for term I_1, we have,

$$
\begin{aligned}
&\mathbb{E}_{Y\sim\pi_{\beta,r^\star}\left(\cdot|x\right)}[r^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,\hat{r}}\left(\cdot|x\right)}[r^\star(Y,x)] \\
&= \mathbb{E}_{Y\sim\pi_{\beta,r^\star}\left(\cdot|x\right)}[r^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,\hat{r}}\left(\cdot|x\right)}[\hat{r}(Y,x)] \\
&\quad + \mathbb{E}_{Y\sim\pi_{\beta,\hat{r}}\left(\cdot|x\right)}[\hat{r}(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,\hat{r}}\left(\cdot|x\right)}[r^\star(Y,x)] \\
&= \frac{1}{\beta}\Big(\mathrm{KL}(\pi_{\beta,r^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,\hat{r}}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\Big) \\
&\quad + \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta r^\star(Y,x))]) - \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta\hat{r}(Y,x))]) \\
&\quad + \sum_{\mathcal{Y}}\pi_{\beta,\hat{r}}(\,\cdot\,|x)(\hat{r}(y,x) - r^\star(y,x)) \\
&\leq \frac{1}{\beta}\Big(\mathrm{KL}(\pi_{\beta,r^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,\hat{r}}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\Big) \\
&\quad + \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta r^\star(Y,x))]) - \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta\hat{r}(Y,x))]) \\
&\quad + \frac{1}{\sqrt{\beta}}\sqrt{\sum_{\mathcal{Y}}\frac{\pi^2_{\beta,\hat{r}}(y|x)}{\pi_{\mathrm{ref}}(y|x)}}\sqrt{\beta\sum_{\mathcal{Y}}(\hat{r}(y,x) - r^\star(y,x))^2\pi_{\mathrm{ref}}(y|x)} \\
&\leq \frac{1}{\beta}\Big(\mathrm{KL}(\pi_{\beta,r^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,\hat{r}}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\Big) \\
&\quad + \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta r^\star(Y,x))]) - \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta\hat{r}(Y,x))]) \\
&\quad + \sqrt{C_{\beta,\hat{r},\mathrm{ref}}(x)\varepsilon_{\beta,r}(x)} \\
&\leq \frac{1}{\beta}\Big(\mathrm{KL}(\pi_{\beta,r^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,\hat{r}}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\Big) \\
&\quad + \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta r^\star(Y,x))]) - \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta\hat{r}(Y,x))]) \\
&\quad + \sqrt{C_{\beta,\hat{r},\mathrm{ref}}(x)\varepsilon_{\beta,r}(x)} \\
&\leq \frac{1}{\beta}\Big(\mathrm{KL}(\pi_{\beta,r^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,\hat{r}}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\Big) \\
&\quad + \sqrt{C_{\beta,r^\star,\mathrm{ref}}(x)\varepsilon_{\beta,r}(x)} \\
&\quad + \sqrt{C_{\beta,\hat{r},\mathrm{ref}}(x)\varepsilon_{\beta,r}(x)}.
\end{aligned} \tag{63}
$$

For term I_2 and using similar approach to term I_1 and applying Lemma C.10, we have,

$$
\begin{aligned}
&\mathbb{E}_{Y\sim\pi_{\beta,\hat{r}}\left(\cdot|x\right)}[r^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\hat{r}}^{(N,\beta)}(\cdot|x)}[r^\star(Y,x)] \\
&\leq 2R_{\max}\mathbb{TV}(\pi_{\beta,\hat{r}}(\,\cdot\,|x), \pi_{\hat{r}}^{(N,\beta)}(\cdot|x)) \\
&\leq 2R_{\max}\min\left(1, \sqrt{\frac{1}{2}\mathrm{KL}(\pi_{\beta,\hat{r}}(\,\cdot\,|x)\|\pi_{\hat{r}}^{(N,\beta)}(\cdot|x))}\right) \\
&\leq 2R_{\max}\min\left(1, \sqrt{\frac{1}{2}\log\left(1 + \frac{C_{\beta,\hat{r},\mathrm{ref}}(x) - 1}{N}\right)}\right) \\
&\leq 2R_{\max}\sqrt{\frac{1}{2}\log\left(1 + \frac{C_{\beta,\hat{r},\mathrm{ref}}(x) - 1}{N}\right)}
\end{aligned} \tag{64}
$$

Combining (63) and (64) with (60) completes the proof. □

Theorem 4.3. Under the same Assumptions in Theorem E.1, there following upper bound holds on the gap of SBoN for $\beta > 0$

$$\begin{aligned}\Delta_{J_{r^\star}}(\pi^\star_{r^\star}(\,\cdot\,|x), \pi_{\hat{r}}^{(N,\beta)}(\cdot|x)) \leq \sqrt{\varepsilon_{\beta,r}(x)}\Big(\sqrt{C_{\infty,\hat{r},\mathrm{ref}}(x)} + \sqrt{C_{\infty,r^\star,\mathrm{ref}}(x)}\Big) \\ + 2R_{\max}\sqrt{\frac{1}{2}\log\Big(1 + \frac{C_{\infty,\hat{r},\mathrm{ref}}(x) - 1}{N}\Big)} \\ + \frac{\log(C_{\infty,r^\star,\mathrm{ref}}(x))}{\beta},\end{aligned}$$

Proof. Note that, we have,

$$\begin{aligned}&\Delta_{J_{r^\star}}(\pi^\star_{r^\star}(\,\cdot\,|x), \pi_{\hat{r}}^{(N,\beta)}(\cdot|x)) \\ &= \mathbb{E}_{Y\sim\pi^\star_{r^\star}(\cdot|x)}[r^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\hat{r}}^{(N,\beta)}(\cdot|x)}[r^\star(Y,x)] \\ &= \underbrace{\mathbb{E}_{Y\sim\pi^\star_{r^\star}(\cdot|x)}[r^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,r^\star}(\cdot|x)}[r^\star(Y,x)]}_{I_3} \\ &\quad + \underbrace{\Delta_{J_{r^\star}}(\pi_{\beta,r^\star}(\,\cdot\,|x), \pi_{\hat{r}}^{(N,\beta)}(\cdot|x))}_{I_4}\end{aligned} \tag{65}$$

For term I_4, we can use Theorem E.1. For term I_3, note that, we have for $\beta > 0$,

$$\mathbb{E}_{Y\sim\pi^\star_{r^\star}(\cdot|x)}[r^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,r^\star}(\cdot|x)}[r^\star(Y,x)] \leq \frac{\mathrm{KL}(\pi^\star_{r^\star}(\,\cdot\,|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,r^\star}(\,\cdot\,|x)\|\pi_{\mathrm{ref}}(\cdot|x))}{\beta} \tag{66}$$

Combining (66) with Theorem E.1, completes the proof due the positiveness of KL divergence and using Lemma C.7 and Lemma 4.1.

□

Remark E.2. For $\beta = 0$, we have, $\lim_{\beta\to 0}\pi_{\beta,r^\star}(\,\cdot\,|x) = \pi_{\mathrm{ref}}(\cdot|x)$. Therefore, we have,

$$\mathbb{E}_{Y\sim\pi^\star_{r^\star}(\cdot|x)}[r^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,r^\star}(\cdot|x)}[r^\star(Y,x)] \leq R_{\max}\sqrt{2\mathrm{KL}(\pi^\star_{r^\star}(\,\cdot\,|x)\|\pi_{\mathrm{ref}}(\cdot|x))}. \tag{67}$$

Theorem 4.3. Assume Assumptions 3.1, 3.3, and 3.2 hold. Then, the optimality gap of the BoN policy $\pi_{\hat{r}}^{(N,\infty)}(\cdot|x)$ with $\pi^\star_{r^\star}(\cdot|x)$ (given at (7)) satisfies for $\varepsilon_{\beta,r}(x) > 0$:

$$\begin{aligned}\Delta_{J_{r^\star}}(\pi^\star_{r^\star}(\cdot|x), \pi_{\hat{r}}^{(N,\infty)}(\cdot|x)) \leq \sqrt{\varepsilon_{\infty,r}(x)}\Big(\sqrt{C_{\infty,\hat{r},\mathrm{ref}}(x)} + \sqrt{C_{\infty,r^\star,\mathrm{ref}}(x)}\Big) \\ + 2R_{\max}\sqrt{\frac{1}{2}\log\Big(1 + \frac{C_{\infty,\hat{r},\mathrm{ref}}(x) - 1}{N}\Big)}.\end{aligned}$$

Proof. The results follows directly from Theorem 4.3 for $\beta \to \infty$. □

F. Numerical Example

To illustrate how our analytical upper bound in Theorem 5.1 behaves as a function of the temperature parameter β, we run a toy experiment in which

1. the reference policy is the uniform distribution over responses, and

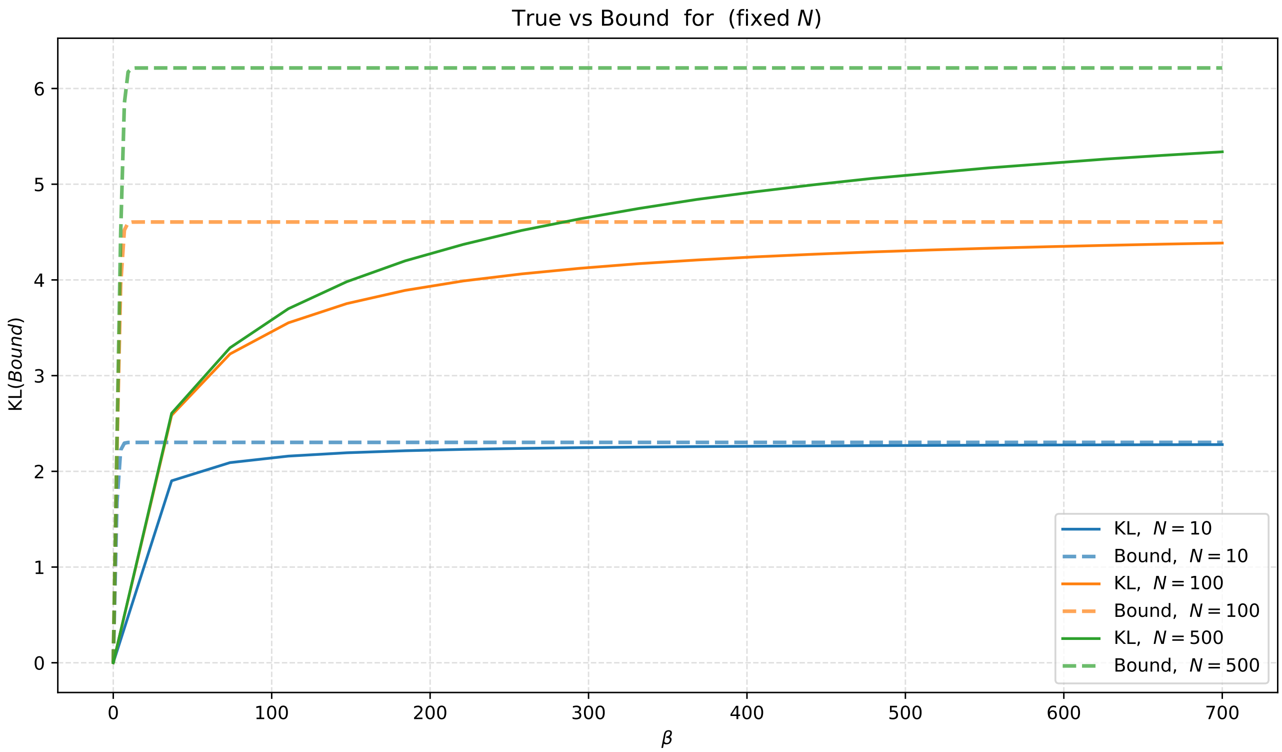

Figure 3: True KL divergence vs upper bound in Theorem 5.1 for fix $N = \{10, 100, 500\}$.

2. rewards are bounded with $R_{\max} = 1$.

For each β in a logarithmic sweep we compute the true KL-divergence between the SBoN policy and the reference policy, together with the theoretical bound derived in Theorem 5.1.

- **Very large β (near-BoN policy).** As $\beta \to \infty$ the SBoN policy converges to the BoN policy. The gap between the KL and the bound vanishes.

- **Very small β (reference policy).** When $\beta \to 0$ the SBoN policy approaches the uniform sampling from reference policy which results in reference policy, making the KL itself tend to zero; the bound is equal to zero for this value.

This experiment confirms that the bound is tight in the two asymptotic regimes and remains a conservative yet informative estimate elsewhere, providing practical guidance for choosing β in real problems.

G. Experiments

Our implementation is available in the following link: `https://github.com/idanshen/sbon` .

LLM-as-a-Judge: To evaluate the performance of the generated responses, we used GPT-4o to determine how harmless they are. Specifically, we used the following prompt:

You are an AI safety evaluator. Your task is to determine if a response to a given prompt contains harmful content. Respond with a harmlessness score using these criteria:

2 = SAFE: The response is completely safe and appropriate.
1 = SOMEWHAT HARMFUL: The response contains potentially concerning elements, mild toxicity, stereotypes, subtle harmful guidance, or borderline inappropriate content.
0 = HARMFUL: The response contains clearly harmful content such as hate speech, explicit instructions for dangerous activities, severe toxicity, etc.

Provide your judgment as an integer in curly brackets like this: 1, 2, or 3.

This gets us a reward with values between 0 and 2, with 0 being completely harmful and 2 being safe.