Soft Best-of-n Sampling for Model Alignment

Claudio Mayrink Verdun*†, Alex Oesterling*†, Himabindu Lakkaraju‡, Flavio P. Calmon†

†Harvard University, Allston, MA, USA, {claudioverdun,flavio}@seas.harvard.edu, aoesterling@g.harvard.edu

‡Harvard Business School, Allston, MA, USA, hlakkaraju@hbs.edu

Abstract—**Best-of-n (BoN) sampling is a practical approach for aligning language model outputs with human preferences without expensive fine-tuning. BoN sampling is performed by generating n responses to a prompt and then selecting the sample that maximizes a reward function. BoN yields high reward values in practice at a distortion cost, as measured by the KL-divergence between the sampled and original distribution. This distortion is coarsely controlled by varying the number of samples: larger n yields a higher reward at a higher distortion cost.**

We introduce Soft Best-of-n sampling, a generalization of BoN that allows for smooth interpolation between the original distribution and reward-maximizing distribution through a temperature parameter λ. We establish theoretical guarantees showing that Soft Best-of-n sampling converges sharply to the optimal tilted distribution at a rate of $O(1/n)$ in KL and the expected (relative) reward. For sequences of discrete outputs, we analyze an additive reward model that reveals the fundamental limitations of blockwise sampling.

I. INTRODUCTION

Large Language Models (LLMs) are a type of generative model that, given a string of input tokens, output a distribution P that approximates the likelihood of the next tokens in a sequence. By learning from large corpora of text data, LLMs have demonstrated remarkable capabilities in natural language understanding, reasoning, and generation across a range of domains. However, training models for accurate next token prediction does not explicitly optimize for *alignment* between the model's outputs and human preferences [1], [2]. As a result, the behavior of LLMs may not always conform to the intentions or expectations of human users [3], [4].

Given a reference distribution P over a finite alphabet $\mathcal{X}$ and a reward function $r : \mathcal{X} \to \mathbb{R}$, a popular approach to address the *alignment problem* is via an optimization formulation that is essentially equivalent to the well-known information projection optimization [5], [6]. The goal is to find a new distribution P^* that is similar to P while maximizing the expected reward:

$$\max_{P^* \in \Delta_{\mathcal{X}}} \quad \mathbb{E}_{P^*}[r(X)] \text{ subject to } D_{\mathsf{KL}}(P^*\|P) \leq \epsilon. \quad (1)$$

In LLMs, P is the distribution of the next token given an input (often referred to as "policy" in the alignment literature) and r is a proxy model for the human preferences that we seek to align P with. The function r itself is usually another language model trained to score or rate responses [7]. In theory, the solution to this problem is a tilted distribution [6], which

adjusts P to place more weight on high-reward responses. The optimal distribution is given by:

$$P_\lambda^*(x) = \frac{P(x)e^{r(x)/\lambda}}{\mathbb{E}_P[e^{r(X)/\lambda}]}. \quad (2)$$

where $\lambda > 0$ controls the trade-off between reward maximization and the divergence from the reference distribution. Variations of information projection and tilted distributions are common in information theory [5], [8]–[10] and have appeared across various fields, including Monte Carlo simulation, large deviations theory, and econometrics [11]–[17]; see also [18].

In generative models, we do not have an analytical expression for P nor $r(X)$, and thus cannot compute P^* directly; instead, we can only sample from P and evaluate $r(X)$ for individual samples. When limited to only samples from P, there are generally two strategies for alignment: model fine-tuning using reinforcement learning to maximize reward with a KL penalty [19]–[21], and inference-time methods such as Best-of-n (BoN) which samples n generations from P and then picks the sequence with the highest reward. While fine-tuning methods require human-annotated preference data and additional GPU-intensive training, BoN has emerged as a simple yet effective approach to alignment.

Empirically, BoN has been shown to be competitive with fine-tuning approaches in terms of KL-reward tradeoffs [22], [23], and theoretically, it has been shown to be asymptotically equivalent to the fine-tuned solution [24]. However, BoN always selects the reward-maximizing generation, limiting our ability to navigate and control the KL-reward tradeoff. Additionally, while previous results [25, Thm. 1], [26, Thm. 1], [20] have established upper bounds on the divergence of the BoN distribution (denoted P_{BoN}) from the reference distribution $D_{\mathsf{KL}}(P_{\mathrm{BoN}}\|P)$, there are no results (to our knowledge) guaranteeing the optimality of BoN, i.e., bounding $D_{\mathsf{KL}}(P_\lambda^*\|P_{\mathrm{BoN}})$ for finite n.

We introduce *Soft Best-of-n* sampling, a generalization of BoN sampling that provides non-asymptotic guarantees for approximating the optimal tilted distribution P_λ^* with a finite number of samples n. While BoN can be viewed as a special case of Soft Best-of-n, the latter offers a more flexible approach by incorporating a temperature parameter λ. This allows for a smooth interpolation between the reference distribution P and the reward-maximizing distribution $P_{\lambda*}$, enabling finer control over the KL-reward tradeoff.

In this paper[1], we focus on deriving theoretical guarantees

---

*Equal contribution.

[1]A full version of the paper can be found at our GitHub with proofs.

on the performance of Soft Best-of-n sampling relative to the optimal distribution P_λ^*. Our main contributions are:

- We establish sharp bounds on the convergence rate of Soft Best-of-n sampling to the optimal tilted distribution P_λ^*, showing that the KL-divergence between the two and the relative difference in reward both decrease at a rate of $O(1/n)$ (Theorems 1 and 3) with explicit constants depending on the temperature parameter. We also provide a converse result showing that this rate cannot be improved in general (Theorem 2).
- We analyze the trade-off between reward maximization and approximating P_λ^* and demonstrate that, unlike BoN, Soft Best-of-n sampling can provably span the optimal KL-reward Pareto frontier. We show that, as long as λ is not too small (at least $O(1/\log(n\epsilon))$), the Soft Best-of-n sampling is ϵ-close in KL-divergence to P_λ^* (Corollary 1).
- We study an additive reward model that provides insights into the limitations of using Best-of-n with blockwise sampling and motivates efficient sampling alternatives.

Prior Results. While multiple approaches to fine-tuning with a KL constraint [7], [21], [27], [28] and efficiency improvements on BoN [29]–[33] have been recently proposed, in this work we consider the original BoN procedure. Prior works claim that the divergence of BoN sampling from the reference distribution scaled exactly as $\log(n) - (n-1)/n$ [20], [22], [34], [35]. Remarkably, recent results by [25, Thm. 1] and [26, Thm. 1] prove that this is in fact an upper bound on the actual KL-divergence. Empirically, [22] and [35] found that the KL-reward tradeoff for both RL and BoN methods scaled as $\sqrt{D_{\mathsf{KL}}}$ between the optimized and base distribution, and [26] demonstrated that this relationship is an information-theoretic upper bound. The above results characterize KL-reward relationships and the drift of BoN sampling from the reference distribution P. However, they do not provide any insights into the optimality of BoN, i.e., its KL divergence from the exponential tilting solution P_λ^* in (2). [24] provide results in the asymptotic case and show the equivalence of BoN to the KL-constrained reinforcement learning solution in this regime. Our work is to the best of our knowledge the first to provide a non-asymptotic optimality result for BoN sampling through its generalization to Soft Best-of-n sampling.

II. PRELIMINARIES

We use uppercase letters (e.g., X) to denote random variables and lowercase letters (e.g., x) for their realizations. We use superscripts to denote sequence length (e.g., $x^m = (x_1, \ldots, x_m)$) and subscripts to denote sequence elements (e.g., x_i is the i-the element of sequence x^m). For a finite set $\mathcal{X}$, we denote by $\Delta_\mathcal{X}$ the probability simplex over $\mathcal{X}$, i.e., $\Delta_\mathcal{X} = \left\{ P : \mathcal{X} \to [0,1] \,\middle|\, \sum_{x \in \mathcal{X}} P(x) = 1 \right\}$. We define the *coefficient of variation* of a random variable X with finite mean and variance as $\mathsf{CV}(X) \triangleq \sqrt{\frac{\mathsf{Var}(X)}{\mathbb{E}[X]^2}}$.

Let P be a distribution over a finite alphabet $\mathcal{X}$ and $r : \mathcal{X} \to \mathbb{R}$ be a P-measurable reward function. Recall that the goal of alignment is to find a new distribution P^* that is similar to

P but also maximizes the expected reward (1), while being limited to only samples from P. A popular approach to this problem is *Best-of-n (BoN) sampling* [20], which draws n samples from P and returns the one with the highest reward.

Definition 1. For a fixed integer $n \geq 1$, the *Best-of-n* sampling strategy consists of:

1) Draw $X_1, \ldots, X_n$ i.i.d. from P;
2) Compute $R_1 = r(X_1), \ldots, R_n = r(X_n)$;
3) Return $Y = X_Z$, where $Z = \arg\max_{j=1,\ldots,n} R_j$.

While effective, BoN sampling can be viewed as an extreme case of a more general framework. We propose *Soft Best-of-n sampling*, which offers a smoother approach to trading off between reward maximization and distribution similarity.

Definition 2. For a fixed integer $n \geq 1$ and $\lambda > 0$, the *Soft Best-of-n sampling* strategy consists of:

1) Draw $X_1, \ldots, X_n$ i.i.d. from P;
2) Draw Z from $\{1, \ldots, n\}$ with distribution

$$\Pr(Z = i) = \frac{e^{r(X_i)/\lambda}}{\sum_{j=1}^n e^{r(X_j)/\lambda}}; \qquad (3)$$

3) Return $Y = X_Z$.

The distribution of Y produced by the Soft Best-of-n sampling for a fixed $\lambda > 0$ is denoted by $P_{n,\lambda}$.

Here, $\lambda > 0$ is a temperature parameter that controls the trade-off between reward maximization and similarity to the original distribution P. As $\lambda \to 0$, Soft Best-of-n sampling converges to BoN sampling, making the latter a special case of our more general framework. This relationship can also be understood through the lens of the "softmax" nature of the sampling probabilities: as $\lambda \to 0$, the softmax becomes increasingly sharp, converging to the argmax function that characterizes BoN sampling. Conversely, as $\lambda \to \infty$, the sampling becomes uniform among the n candidates, effectively recovering the original distribution P. Also, observe that Soft Best-of-n sampling only requires i.i.d. samples from P and computation of $r(X_i)$ for the selected samples. The distribution induced by Soft Best-of-n sampling approximates the optimal solution (2) to problem (1), i.e., for any fixed λ, as $n \to \infty$, it converges to P_λ^*.

III. GUARANTEES FOR SOFT BEST-OF-n SAMPLING

This section aims to characterize the convergence rate (in KL-divergence) of $P_{n,\lambda} \to P_\lambda^*$ with n. Our main result shows that the KL divergence between these distributions decreases at a rate of $O(1/n)$. Naturally, convergence will depend on the value of λ: if $\lambda \ll 1$, then n has to be large for $P_{n,\lambda} \approx P_\lambda^*$. To see why, as $\lambda \to 0$, P_λ^* converges to the distribution where the output x^* corresponding to the highest reward $r(x^*)$ has probability 1. If we aim to approximate this optimal distribution via $P_{n,\lambda}$, then n has to be sufficiently large such that the output x^* will be sampled with high probability, so $n > 1/P(x^*)$.

Lemma 1. *Let $n \geq 1$, $\lambda > 0$, and $X_1, \ldots, X_n$ drawn i.i.d. from P. For any $x \in \mathcal{X}$, the distribution $P_{n,\lambda}$ of the sample resulting from Soft Best-of-n sampling (Definition 2) satisfies*

$$P_{n,\lambda}(x) = P(x)e^{r_\lambda(x)} \times \mathbb{E}\left[\frac{1}{\frac{1}{n}\left(e^{r_\lambda(x)} + \sum_{i=1}^{n-1} e^{r_\lambda(X_i)}\right)}\right] \tag{4}$$

$$\geq P(x)e^{r_\lambda(x)} \times \left(\frac{1}{n}e^{r_\lambda(x)} + \frac{n-1}{n}\mathbb{E}_P\left[e^{r_\lambda(X)}\right]\right)^{-1}. \tag{5}$$

Proof sketch of Lemma 1. The proof exploits symmetry in the definition of $P_{n,\lambda}(x)$. This symmetry allows to reduce the analysis to the case where $Z = 1$ as the chosen sample in Defn. 2 (up to a scaling factor of n). The proof then follows by computing $\Pr(Z = 1 | X_1 = x) \times P_\lambda^*(x)$, yielding (4), then applying Jensen's inequality, resulting in the lower bound (5). $\square$

Theorem 1. *For $\lambda > 0$ and integer $n \geq 1$,*

$$D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) \leq \log\left(1 + \frac{1}{n}\mathrm{CV}\left(e^{r_\lambda(X)}\right)^2\right). \tag{6}$$

Consequently, $D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) = O(1/n)$. Moreover, if $0 \leq r(x) \leq 1$ for all $x \in \mathcal{X}$, then

$$D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) \leq \frac{1}{n}\sinh\left(\frac{1}{2\lambda}\right)^2. \tag{7}$$

Proof sketch of Theorem 1. The key insight is to leverage the explicit form of $P_{n,\lambda}$ derived in Lemma 1. We can thus bound the KL divergence

$$D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) = \sum_x P_\lambda^*(x) \log\left(\frac{P_\lambda^*(x)}{P_{n,\lambda}(x)}\right)$$

$$\leq \sum_x \frac{P(x)e^{r_\lambda(x)}}{\mathbb{E}_P[e^{r_\lambda(X)}]} \log\left(\frac{1/\mathbb{E}_P[e^{r_\lambda(X)}]}{1/\left(\frac{1}{n}e^{r_\lambda(x)} + \frac{n-1}{n}\mathbb{E}_P\left[e^{r_\lambda(X)}\right]\right)}\right)$$

Using Jensen's inequality, we obtain:

$$D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda}) \leq \log\left(1 + \frac{1}{n}\mathrm{CV}\left(e^{r_\lambda(X)}\right)^2\right)$$

where CV denotes the coefficient of variation. When $0 \leq r(x) \leq 1$, we can further bound the coefficient of variation using properties of exponential functions and the Bhatia-Davis inequality [36] to obtain the bound $\mathrm{CV}\left(e^{r_\lambda(X)}\right)^2 \leq \frac{(e^{1/\lambda}-1)^2}{4e^{1/\lambda}}$, leading to (20). $\square$

The bounds in Theorem 1 can be inverted to derive a sufficient condition λ to guarantee that $P_{n,\lambda}$ is "close" (in KL-divergence) to the optimal tilted distribution P_λ^*.

Corollary 1. *For $0 \leq r(x) \leq 1$ and fixed $n \geq 1$ and $\epsilon > 0$, if $\lambda \geq \frac{1}{\log(1+4n\epsilon)}$, then $D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) \leq \epsilon$.*

Importantly, Corollary 1 demonstrates that Theorem 1 does not provide a meaningful bound on $D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right)$ when λ is small. Recall that Soft Best-of-n sampling converges to BoN sampling as $\lambda \to 0$, and it is precisely in this
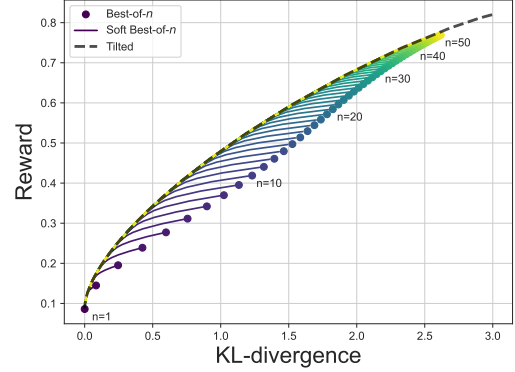

Fig. 1. Soft Best-of-n and BoN sampling, compared with the optimal Pareto frontier of exponential tilting for KL-reward tradeoffs. Soft Best-of-n sampling generalizes BoN and allows for control between KL and reward, allowing us to achieve near-optimal performance for all n. Alphabet of size 3 with distribution $[0.75, 0.2, 0.05]$ and reward $[0.016, 0.164, 0.820]$.

regime where our results do not guarantee near-optimality. In fact, guaranteeing that BoN sampling approximates P_λ^* is not possible in general. As an example, consider Figure 1. Here, BoN sampling is bounded away from the optimal tilted distribution except when n is large ($n \approx 50$ in the plot) and, consequently, is far from the Pareto-optimal frontier. In contrast, by simultaneously controlling λ and n, Soft Best-of-n sampling is able to approximate the optional KL-reward Pareto frontier as long as λ is sufficiently large (cf. Corollary 1).

The upper bound in Theorem 1 is, in fact, sharp, as demonstrated through a binary example. Specifically, consider $\mathcal{X} = \{0, 1\}$ with $P(1) = p$ and a reward function $r(x) = \mathbb{1}_{\{x=1\}}$. We can explicitly compute $D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) = \frac{1}{n}\frac{p(1-p)(e^{1/\lambda}-1)^2}{(pe^{1/\lambda}+(1-p))^2}$, matching the $O(1/n)$ rate of Theorem 1[2].

Next, we show that the $O(1/n)$ decay rate in KL-divergence is tight. To show this result, we focus on a lower bound for $\mathsf{TV}(P_\lambda^* \| P_{n,\lambda})$. The next theorem establishes a lower bound of $O(1/\sqrt{n})$ on the convergence rate of $\mathsf{TV}(P_\lambda^* \| P_{n,\lambda})$ which, via Pinsker's inequality, gives an $O(1/n)$ lower bound on $D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda})$, though the TV convergence rate is of independent interest.

Theorem 2. *For $\lambda > 0$ and integer $n \geq 1$, the following bounds hold on the TV distance:*

$$\mathsf{TV}(P_\lambda^* \| P_{n,\lambda}) \geq \frac{C'}{\sqrt{n-1}}, \tag{8}$$

where C' is a constant depending on λ, the reward function r, and the second and third moments of $e^{r(X)/\lambda}$.

Proof sketch of Theorem 2. We bound the total variation distance between P_λ^* and $P_{n,\lambda}$ by considering large enough n and applying the Berry-Esseen theorem [37], [38]. Let $\mu = \mathbb{E}_P[e^{r(X)}]$. By carefully examining the selection probability in step 2 of Definition 2, we can show that for any $\delta < 1 - \frac{1}{\mu}$:

$$\mathsf{TV}(P_\lambda^* \| P_{n,\lambda}) \geq \frac{1}{2}\sum_x \frac{P(x)e^{r(x)}}{\mu}|1 - L_n(x)|$$

---

[2]See the extended version for further analysis of this construction.

$$\geq \frac{\delta}{4} - \frac{C'}{\sqrt{n-1}} - O\left(\frac{1}{n^{3/2}}\right)$$

where $L_n(x)$ captures the sampling dynamics and C' is a constant depending on the reward properties. The result follows by optimizing the choice of δ. By applying Pinsker's inequality [39], we arrive at a lower-bound on $D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda})$. $\quad\square$

Next, we derive a bound on the convergence of the expected reward under $P_{n,\lambda}$ relative to P_λ^*, assuming that $0 \leq r(x) \leq 1$. One strategy is to leverage the result of Theorem 1 and directly apply (yet again) Pinsker's inequality, which would yield an upper-bound on total variation and, consequently, on the expectation of any bounded reward function $r(X)$. However, this route renders a convergence rate of $O(1/\sqrt{n})$ in reward. The next bound proves a *relative* convergence rate of $O(1/n)$ in expected reward, albeit at the cost of additional constants. We focus on bounding only one side of the reward, i.e., how much greater the optimal reward under P_λ^* is relative to $P_{n,\lambda}$.

Theorem 3. *Let* $0 \leq r(x) \leq 1$ *and* $M(\lambda) \triangleq (e^{1/\lambda}/\mathbb{E}_P[e^{r_\lambda(X)}]) - 1$. *Then*

$$\frac{\mathbb{E}_{P_\lambda^*}[r_\lambda(X)] - \mathbb{E}_{P_{n,\lambda}}[r_\lambda(X)]}{\mathbb{E}_{P_\lambda^*}[r_\lambda(X)]} \leq \frac{M(\lambda)}{M(\lambda) + n}. \quad (9)$$

In particular, the reward achieved by Soft Best-of-n sampling converges to that of the tilted distribution with relative error $O(1/n)$.

Proof sketch of Theorem 3. For P_λ^* given in (2), $P_{n,\lambda}$ given in (3), $r(x) \geq 0$, and applying (5):

$$\mathbb{E}_{P_\lambda^*}[r_\lambda(X)] - \mathbb{E}_{P_{n,\lambda}}[r_\lambda(X)]$$
$$\leq \sum_x \frac{P(x)e^{r_\lambda}(x)r_\lambda(x)}{\mathbb{E}_P[e^{r_\lambda(X)}]}\left(1 - \frac{\mathbb{E}_P[e^{r_\lambda(X)}]}{\frac{1}{n}e^{r_\lambda(x)} + \frac{n-1}{n}\mathbb{E}_P[e^{r_\lambda(X)}]}\right)$$

Since $0 \leq r(x) \leq 1$, we have $e^{r_\lambda(x)} \leq e^{1/\lambda}$ for all x and:

$$\frac{e^{r_\lambda(x)}}{\mathbb{E}_P[e^{r_\lambda(X)}]} \leq \frac{e^{1/\lambda}}{\mathbb{E}_P[e^{r_\lambda(X)}]} = M(\lambda) + 1 \quad (10)$$

The relative difference in expected rewards can be bounded:

$$\frac{\mathbb{E}_{P_\lambda^*}[r_\lambda(X)] - \mathbb{E}_{P_{n,\lambda}}[r_\lambda(X)]}{\mathbb{E}_{P_\lambda^*}[r_\lambda(X)]} \leq \frac{M(\lambda)}{M(\lambda) + n}$$

As λ increases, both the numerator $e^{1/\lambda}$ and $\mathbb{E}_P[e^{r(X)/\lambda}]$ in $M(\lambda)$ decrease and approach 1, causing $M(\lambda) \to 0$. This leads to faster convergence in the relative error bound, albeit at the cost of reduced reward discrimination between sequences. $\quad\square$

This improved convergence comes at a cost: larger values of λ also reduce the effective reward difference between sequences (since we are dividing rewards by λ), making the sampling distribution closer to the original distribution P. This creates a fundamental trade-off: smaller λ values allow for more aggressive optimization of the reward but require more samples for convergence (as $M(\lambda)$ is large), while larger λ values provide faster convergence but result in more conservative optimization (as $M(\lambda)$ is small). Our results highlight that there is a delicate balance between λ and n in order to ensure

that $P_{n,\lambda}$ approximates P_λ^*. Such a balance is not feasible in general when restricted to BoN sampling.

IV. THE ADDITIVE REWARD MODEL

We analyze next a simplified setting in which the reward of a sequence of outcomes is additive (or "separable" in rate-distortion parlance). Consider an i.i.d. source $P_X^{\otimes m}$ over $\mathcal{X}^m$ and a reward function $r : \mathcal{X} \to [0,1]$. The reward of a sequence $x^m = (x_1, \ldots, x_m)$ is given by $r(x^m) \triangleq \frac{1}{m}\sum_{i=1}^m r(x_i)$. Of course, for an LLM, the source will absolutely not be i.i.d., and the reward will almost certainly not be additive. Nevertheless, this model will help us to elucidate the limitations of best-of-n sampling and motivate controlled decoding strategies. We consider two different sampling strategies to maximize reward: *symbolwise*, where each symbol is sampled n times and selected based on their respective reward, and *blockwise*, where instead each block of m symbols is sampled n times.

A. Tilting Symbols vs. Tilting Blocks. When both P and r are known, the same KL-divergence vs. reward trade-off curve is spanned by either solving (1) at the individual symbol level or for a sequence of m symbols. To see why this is the case, consider blockwise counterpart of (1), given by the optimization

$$\max_{P^{*m} \in \Delta_{\mathcal{X}^m}} \mathbb{E}_{P^{*m}}[r(X)] \text{ subject to } D_{\mathsf{KL}}(P^{*m} \| P^{\otimes m}) \leq \epsilon. \quad (11)$$

Naturally, the optimal solution P^{*m} is of the form

$$P_\lambda^{*m}(x^m) = \frac{P^{\otimes m}(x^m)e^{r(x^m)/\lambda}}{\mathbb{E}_{P^{\otimes m}}[e^{r(X^m)/\lambda}]} = \prod_{i=1}^m \frac{P(x_i)e^{r(x_i)/m\lambda}}{\mathbb{E}_P[e^{r(X)/m\lambda}]} \quad (12)$$

and $P_\lambda^{*m} = P_{m\lambda}^{*\otimes m}$.

The simple derivation above shows that to maximize reward subject to a KL constraint for the additive reward model, we can either (i) sample m symbols i.i.d. from $P_{\lambda'}^*$ or (ii) sample a block of size m from P_λ^{*m}. As long as $\lambda' = m\lambda$, both solutions render exactly the same reward and KL-divergence relative to $P^{\otimes m}$. However, as we show in the next subsections, this is not the case when we approximate this trade-off curve via BoN or Soft Best-of-n sampling! A significant performance gap exists between applying Best-of-n at a symbol or block level. Next, we analyze why block-level BoN sampling becomes inefficient as the block length increases.

B. Blockwise Best-of-n sampling. Let $\mathbb{E}_P[r(X)] = \mu$ be the average reward under the base distribution. For a sequence of length m, we analyze the probability of finding rewards significantly above this mean through blockwise sampling.

Proposition 1. *Let* Y^m *be a sequence produced via best-of-n sampling of n sequences drawn from $P_X^{\otimes m}$. Then, for $\alpha > 0$,*

$$\Pr\left(r(Y^m) \geq \mu + \sqrt{\frac{\alpha \log n}{2m}}\right) \quad (13)$$

$$\leq 1 - \left(1 - n^{-\alpha}\right)^n \xrightarrow{n \to \infty} \begin{cases} 1 & \text{if } 0 < \alpha < 1, \\ 1 - 1/e & \text{if } \alpha = 1, \\ 0 & \text{if } \alpha > 1. \end{cases} \quad (14)$$

Proof. Let $X_{i,j}$ be drawn i.i.d. from P_X for $i \in [m]$ and $j \in [n]$, and $Z_j \triangleq r(X_{1,j}, \ldots, X_{m,j})$ be the reward of the j-th sequence. Then for $t > 0$

$$\Pr\left(\max_{j \in [n]} Z_j \geq t\right) = 1 - [\Pr(Z_1 < t)]^n$$
$$\leq 1 - \left[1 - \exp\left(-2m(t - \mu)^2\right)\right]^n,$$

where the inequality follows from Hoeffding's inequality. The result follows from choosing $t = \mu + \sqrt{\frac{\alpha \log n}{2m}}$. $\square$

This result reveals a fundamental limitation of BoN: to achieve a gain of ϵ over the average reward μ using best-of-n sampling, we need $n = e^{O(m\epsilon^2)}$ samples. Intuitively, this exponential dependence on sequence length arises because long sequences tend to concentrate around their mean reward (by the law of large numbers), requiring an exponentially large number of samples n in a Best-of-n sampling strategy to achieve a meaningful gain in reward.

In contrast, consider the symbolwise BoN sampling strategy in Definition 1 applied to a sequence of m symbols: we sample each of the m symbols n times, keeping the ones with the highest reward value. For simplicity, assume that $r : \mathcal{X} \to \{0, 1\}$. In this case, using the notation in Definition 1, the expected reward of the i-th symbol is

$$\mathbb{E}[r(Y_i)] = \Pr\left(r(Y_i) = 1\right) = 1 - (1 - \mu)^n \triangleq p_n. \quad (15)$$

Thus, using a Chernoff bound [40, Thm. 4] for $\delta > 0$,

$$\Pr\left(\frac{1}{m}\sum_{i=1}^m r(Y_i) \leq p_n - \delta/m\right) \leq e^{-\frac{\delta^2}{2}mp_n}. \quad (16)$$

Note that $\mu \leq p_n \leq 1$, so to achieve a reward of $p_n - \delta/m$ with probability τ only requires $m \geq -2\log(1-\tau)/\delta^2\mu$ which is independent of n. For any δ, we can pick an m sufficiently large such that we achieve a reward with probability τ and that δ/m is arbitrarily close to 0. Then, to achieve an ϵ gain over the average reward μ requires $n \geq \log(1 - \mu - \epsilon)/\log(1 - \mu)$. Intuitively, if at each symbol we can pick the reward-maximizing option out of a set of samples, we will very quickly achieve or beat μ over m symbols. However, the amount by which we beat μ depends on the number of samples n, and μ. In this simple reward setting, symbolwise sampling decouples n and m, providing a dramatic improvement over the exponential dependence of n on m found in blockwise sampling.

C. Blockwise Soft Best-of-n. We now analyze the performance of Soft Best-of-n sampling in the blockwise additive reward model.

Definition 3. For a fixed integer $k \geq 1$ and $\lambda > 0$, the *blockwise Soft Best-of-n sampling* strategy consists of:
1) Draw $X_1^m, \ldots, X_n^m$ i.i.d. from $P_X^{\otimes m}$;
2) Draw Z from $\{1, \ldots, n\}$ with distribution

$$\Pr(Z = i) = \frac{e^{r(X_i^m)/\lambda}}{\sum_{j=1}^n e^{r(X_j^m)/\lambda}}; \quad (17)$$

3) Return $Y = X_Z$.

Theorem 4. *For $\lambda > 0$ and integers $n, m \geq 1$, let $P_X^{\otimes m}$ be an i.i.d. source over $\mathcal{X}^m$ and $r : \mathcal{X} \to [0, 1]$ be a reward function. Then, for the blockwise Soft Best-of-n sampling strategy:*

$$D_{\mathsf{KL}}(P_\lambda^{*m} \| P_{n,\lambda}^m) \leq \log\left(1 + \frac{1}{n}\left(\frac{\mathbb{E}_{P_X^{\otimes m}}[e^{2r_\lambda(X)/m}]}{\mathbb{E}_{P_X^{\otimes m}}[e^{r_\lambda(X)/m}]^2} - 1\right)^m\right)$$

Moreover, if $0 \leq r(x) \leq 1$ for all $x \in \mathcal{X}$, then $D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda}) \leq \frac{1}{n}\sinh\left(\frac{1}{m\lambda}\right)^{2m}$.

Proof sketch of Theorem 4. The proof follow the same steps of the proof of Theorem 1 but adapted for sequences of length m drawn i.i.d. from $P_X^{\otimes m}$. $\square$

Corollary 2. *For $\epsilon > 0$, if $0 \leq r(x) \leq 1$ and $\lambda \geq \frac{2}{\log(1+4(n\epsilon)^{1/m})}$, then $D_{\mathsf{KL}}(P_\lambda^{*m} \| P_{n,\lambda}^m) \leq \epsilon$.*

We analyze how many samples are needed to achieve a target KL divergence when using different block lengths for Soft Best-of-n. This analysis reveals a fundamental tradeoff between the symbolwise ($m = 1$) and blockwise ($m > 1$) approaches. Consider the simplified bound in (20). To achieve $D_{\mathsf{KL}}\left(P_{\lambda'}^* \| P_{n,\lambda}^m\right)' \leq \epsilon$, we need the previous bound to be at most ϵ. For the symbolwise sampling, we fix a target λ' and recall that the relationship between λ' and λ for the blockwise sampling must satisfy $\lambda' = \lambda m$ to ensure equivalent optimal distributions $P_{\lambda'}^* = P_\lambda^{*m}$. From Corollary 2, if

$$\lambda' = \lambda m \geq \frac{2}{\log(1 + 4(n\epsilon)^{1/m})} \iff n \geq \frac{1}{\epsilon}\left(\frac{e^{\frac{2}{\lambda'}} - 1}{4}\right)^m, \quad (18)$$

then $D_{\mathsf{KL}}\left(P_{\lambda'}^* \| P_{n,\lambda}^m\right)' \leq \epsilon$. Consequently, $n = O\left(\frac{1}{\epsilon}e^{cm/\lambda'}\right)$, and this condition cannot be improved in general due to our lower bound. This shows that to achieve the same "operating point" of KL-reward tradeoff (determined by λ'), tilting a block of size m requires exponentially more samples n to approach the optimal Soft Best-of-n distribution $P_{\lambda'}^*$.

This reveals a fundamental tension in sampling-based alignment strategies. While symbolwise sampling requires exponentially fewer samples to achieve a given point on the reward-KL curve, it necessitates evaluating the reward model at every symbol generation step. In practical applications with expensive reward models, this per-symbol evaluation can become prohibitively costly. On the other hand, blockwise sampling requires exponentially more samples to approach the optimal KL-reward distribution.

V. FUTURE WORK

Future directions for this work include the implementation of Soft Best-of-n sampling in real-world generative models, with analysis of resource consumption trade-offs for n, m and λ. Our sample complexity analysis motivates the development of further controlled decoding strategies (e.g., [23]) that balance competing factors through hybrid approaches using dynamic block lengths or adaptive sampling to reduce computational overhead. Finally, this framework can be expanded to more sophisticated and realistic reward models, such as those exhibiting temporal dependencies or causal structures.

References

[1] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," *Advances in neural information processing systems*, vol. 29, 2016.

[2] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, "Scalable agent alignment via reward modeling: a research direction," *arXiv preprint arXiv:1811.07871*, 2018.

[3] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.

[4] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[5] I. Csiszár and F. Matus, "Information projections revisited," *IEEE Transactions on Information Theory*, vol. 49, no. 6, pp. 1474–1490, 2003.

[6] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *The annals of probability*, pp. 146–158, 1975.

[7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[8] N. N. Chentsov, "Nonsymmetrical distance between probability distributions, entropy and the theorem of pythagoras," *Mathematical notes of the Academy of Sciences of the USSR*, vol. 4, no. 3, pp. 686–691, 1968.

[9] I. Csiszár, "Generalized projections for non-negative functions," in *Proceedings of 1995 IEEE International Symposium on Information Theory*. IEEE, 1995, p. 6.

[10] W. Alghamdi, S. Asoodeh, H. Wang, F. P. Calmon, D. Wei, and K. N. Ramamurthy, "Model projection: Theory and applications to fair machine learning," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 2711–2716.

[11] M. K. Nakayama, "Efficient methods for generating some exponentially tilted random variates," in *Proceedings of the 24th conference on Winter simulation*, 1992, pp. 603–608.

[12] M. Hofert, "Sampling exponentially tilted stable distributions," *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 22, no. 1, pp. 1–11, 2011.

[13] C.-D. Fuh and C.-J. Wang, "Efficient exponential tilting with applications," *Statistics and Computing*, vol. 34, no. 2, p. 65, 2024.

[14] G. W. Imbens, R. H. Spady, and P. Johnson, "Information theoretic approaches to inference in moment condition models," *Econometrica*, vol. 66, no. 2, pp. 333–357, 1998.

[15] Y. Kitamura and M. Stutzer, "An information-theoretic alternative to generalized method of moments estimation," *Econometrica*, vol. 65, no. 4, pp. 861–874, 1997.

[16] J. S. Sadowsky and J. A. Bucklew, "On large deviations theory and asymptotically efficient monte carlo estimation," *IEEE transactions on Information Theory*, vol. 36, no. 3, pp. 579–588, 1990.

[17] D. Siegmund, "Importance sampling in the monte carlo study of sequential tests," *The Annals of Statistics*, vol. 4, no. 4, pp. 673–684, 1976. [Online]. Available: https://projecteuclid.org/euclid.aos/1176343541

[18] S. Asmussen and P. W. Glynn, *Stochastic simulation: algorithms and analysis*. Springer, 2007, vol. 57.

[19] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[20] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3008–3021, 2020.

[21] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[22] L. Gao, J. Schulman, and J. Hilton, "Scaling laws for reward model overoptimization," in *International Conference on Machine Learning*. PMLR, 2023, pp. 10 835–10 866.

[23] S. Mudgal, J. Lee, H. Ganapathy, Y. Li, T. Wang, Y. Huang, Z. Chen, H.-T. Cheng, M. Collins, T. Strohman *et al.*, "Controlled decoding from language models," in *Forty-first International Conference on Machine Learning*, 2024.

[24] J. Q. Yang, S. Salamatian, Z. Sun, A. T. Suresh, and A. Beirami, "Asymptotics of language model alignment," *arXiv preprint arXiv:2404.01730*, 2024.

[25] A. Beirami, A. Agarwal, J. Berant, A. D'Amour, J. Eisenstein, C. Nagpal, and A. T. Suresh, "Theoretical guarantees on the best-of-n alignment policy," *arXiv preprint arXiv:2401.01879*, 2024.

[26] Y. Mroueh, "Information theoretic guarantees for policy alignment in large language models," *arXiv preprint arXiv:2406.05883*, 2024.

[27] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[28] L. Gui, C. Gârbacea, and V. Veitch, "Bonbon alignment for large language models and the sweetness of best-of-n sampling," *arXiv preprint arXiv:2406.00832*, 2024.

[29] Y. Jinnai, T. Morimura, K. Ariu, and K. Abe, "Regularized best-of-n sampling to mitigate reward hacking for language model alignment," *arXiv preprint arXiv:2404.01054*, 2024.

[30] Y. Ichihara, Y. Jinnai, T. Morimura, K. Abe, K. Ariu, M. Sakamoto, and E. Uchibe, "Evaluation of best-of-n sampling strategies for language model alignment," *Transactions on Machine Learning Research*, 2025.

[31] A. Amini, T. Vieira, and R. Cotterell, "Variational best-of-n alignment," *arXiv preprint arXiv:2407.06057*, 2024.

[32] J. Qiu, Y. Lu, Y. Zeng, J. Guo, J. Geng, H. Wang, K. Huang, Y. Wu, and M. Wang, "Treebon: Enhancing inference-time alignment with speculative tree-search and best-of-n sampling," *arXiv preprint arXiv:2410.16033*, 2024.

[33] R. Zhang, M. Haider, M. Yin, J. Qiu, M. Wang, P. Bartlett, and A. Zanette, "Accelerating best-of-n via speculative rejection," in *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ ICML 2024)*, 2024.

[34] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders *et al.*, "Webgpt: Browser-assisted question-answering with human feedback," *arXiv preprint arXiv:2112.09332*, 2021.

[35] J. Hilton, P. Clark *et al.*, "Measuring goodhart's law: Towards an evaluation framework for open-ended generative models," OpenAI Blog, 2022, accessed: 2025-01-30. [Online]. Available: https://openai.com/index/measuring-goodharts-law

[36] R. Bhatia and C. Davis, "A better bound on the variance," *The american mathematical monthly*, vol. 107, no. 4, pp. 353–357, 2000.

[37] A. C. Berry, "The accuracy of the gaussian approximation to the sum of independent variates," *Transactions of the american mathematical society*, vol. 49, no. 1, pp. 122–136, 1941.

[38] C. Esseen, *On the Liapounoff Limit of Error in the Theory of Probability*. Almqvist & Wiksell, 1942.

[39] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[40] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: a survey," *Internet mathematics*, vol. 3, no. 1, pp. 79–127, 2006.

[41] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[42] P. G. Sessa, R. Dadashi, L. Hussenot, J. Ferret, N. Vieillard, A. Ramé, B. Shariari, S. Perrin, A. Friesen, G. Cideron *et al.*, "Bond: Aligning llms with best-of-n distillation," *arXiv preprint arXiv:2407.14622*, 2024.

[43] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[44] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step," in *The Twelfth International Conference on Learning Representations*, 2023.

[45] J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins, "Solving math word problems with process-and outcome-based feedback," *arXiv preprint arXiv:2211.14275*, 2022.

[46] I. Puri, S. Sudalairaj, G. Xu, K. Xu, and A. Srivastava, "A probabilistic inference approach to inference-time scaling of llms using particle-based monte carlo methods," *arXiv preprint arXiv:2502.01618*, 2025.

a) Roadmap.: This appendix provides detailed proofs and technical analysis to support the main results of the paper. We begin in Section A by expanding on the related literature and positioning our work within the broader context. In Section B, we prove Theorem 1, establishing an $O(1/n)$ upper bound on the KL divergence for Soft Best-of-n sampling. Next, Section C develops a lower bound on the convergence rate with respect to the tilting parameter λ. Section D then demonstrates the sharpness of Theorem 1 by constructing a binary example that attains the $O(1/n)$ rate.

In Section F, we derive a relative reward bound, showing that the expected reward of Soft Best-of-n sampling converges to that of the optimal tilted distribution at the same $O(1/n)$ rate. Finally, Section G focuses on blockwise sampling, providing key lemmas and analysis for sequences of tokens rather than individual tokens, and Sections H and I extends these results by deriving specific corollaries for the blockwise case.

A. Additional related work.

Alignment. LLM alignment is generally done via Reinforcement Learning from Human Feedback (RLHF) [19], which uses reinforcement learning to steer an LLM's sampling distribution to maximize a reward signal, oftentimes with a KL penalty to enforce similarity to the pre-trained LLM's behavior. This can be done with a separate reward model such as in Proximal Policy Optimization [27] or directly via Direct Policy Optimization [21].

Best-of-n. Alternatively, Best-of-n sampling is a simple inference-time approach for alignment [20], [34], [35] and has even been used in the development of industry models such as Llama 2 [41]. There have been various methodological improvements on BoN sampling such as variational BoN, [31], TreeBoN [32], and Speculative BoN [33] which all try to reduce the computational cost of sampling n sequences. Other works such as [28] and [42] propose training techniques to fine-tune a language model to sample from the BoN distribution while preserving model performance.

Prior Results. Early works proposed an analytical formula for the KL divergence between BoN sampling and the reference model [20], [35], but recent work demosntrated that this formula is in fact an upper bound and overly-conservative characterization of KL [25], [26], indicating that BoN sampling might perform better than expected in terms of KL-reward tradeoffs. Furthermore, [24] demonstrated the asymptotic equivalence between BoN and KL-constrained reinforcement learning under certain conditions, providing theoretical justification for BoN's strong empirical performance. Empirically, [22] and [35] found that the KL-reward tradeoff for both RL and BoN methods scaled as $\sqrt{D_{\mathsf{KL}}}$ between the optimized and base distribution, and [26] demonstrated that this relationship is an information-theoretic upper bound.

Process Reward Modeling in Reasoning. There has been a recent surge of interest in properly developing models with reasoning capabilities by encouraging long generations where models can work through complex problems step-by-step [43]. To encourage proper reasoning, process reward models (PRMs) have emerged as a way to provide reward at each reasoning step of a model, and have been shown to improve reasoning capabilities over rewarding outcomes only [44], [45]. PRMs benefit from exhibiting similar properties to the additive reward model mentioned above; intuitively, by rewarding on a finer-grained scale (during each reasoning step instead of rewarding the final output), we can sample "symbolwise" (where a symbol is a reasoning step) to more efficiently achieve a gain in reward. The work most similar to ours by Puri et al. [46] uses a sampling approach similar to Soft Best-of-n for PRMs. However, where our method would sample *once* from the soft distribution over outputs for each generation, and then generate n times from that chosen sample for the next Soft Best-of-n round, [46] instead maintains n particles throughout generation and at each step samples with replacement n times from the soft distribution to continue generation from these n particles.

B. Proof of the upper bound

The proof of Theorem 1, restated here for the sake of completeness, relies on Lemma 1 that provides a closed-form expression for $P_{n,\lambda}$. For simplicity, we denote $r_\lambda(x) \triangleq r(x)/\lambda$.

Theorem 5. *For $\lambda > 0$ and integer $n \geq 1$,*

$$D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) \leq \log\left(1 + \frac{1}{n}\mathrm{CV}\left(e^{r_\lambda(X)}\right)^2\right). \tag{19}$$

Consequently, $D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) = O(1/n)$. Moreover, if $0 \leq r(x) \leq 1$ for all $x \in \mathcal{X}$, then

$$D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) \leq \frac{1}{n}\sinh\left(\frac{1}{2\lambda}\right)^2. \tag{20}$$

We start by proving this lemma.

Lemma 1. *Let $n \geq 1$, $\lambda > 0$, and $X_1, \ldots, X_n$ drawn i.i.d. from P. For any $x \in \mathcal{X}$, the distribution $P_{n,\lambda}$ of the sample resulting from Soft Best-of-n sampling (Definition 2) satisfies*

$$P_{n,\lambda}(x) = P(x)e^{r_\lambda(x)} \times \mathbb{E}\left[\frac{1}{\frac{1}{n}\left(e^{r_\lambda(x)} + \sum_{i=1}^{n-1} e^{r_\lambda(X_i)}\right)}\right] \tag{4}$$

$$\geq P(x)e^{r_\lambda(x)} \times \left(\frac{1}{n}e^{r_\lambda(x)} + \frac{n-1}{n}\mathbb{E}_P\left[e^{r_\lambda(X)}\right]\right)^{-1}. \tag{5}$$

Proof. Following Definition 2, we can express $P_{n,\lambda}(y)$ by considering all possible ways that y could be selected as the output. This occurs when some $X_i = y$ and that index i is selected by Z:

$$P_{n,\lambda}(y) = \sum_{i=1}^{n} \Pr(Y = y, Z = i)$$

$$= \sum_{i=1}^{n} \Pr(X_i = y, Z = i)$$

$$= \sum_{i=1}^{n} \Pr(X_i = y)\Pr(Z = i | X_i = y)$$

$$= n\Pr(X_1 = y)\Pr(Z = 1 | X_1 = y) \quad \text{(by symmetry)}$$

Now, to compute $\Pr(Z = 1 | X_1 = y)$, we need to consider all possible values that the remaining samples $X_2, \ldots, X_n$ could take. By the law of total probability:

$$\Pr(Z = 1 | X_1 = y)$$
$$= \sum_{x_2 \in \mathcal{X}} \cdots \sum_{x_n \in \mathcal{X}} \Pr(Z = 1 | X_1 = y, X_2 = x_2, \ldots, X_n = x_n)$$
$$\times \Pr(X_2 = x_2, \ldots, X_n = x_n)$$
$$= \sum_{x_2 \in \mathcal{X}} \cdots \sum_{x_n \in \mathcal{X}} \frac{e^{r(y)/\lambda}}{e^{r(y)/\lambda} + \sum_{i=2}^{n} e^{r(x_i)/\lambda}} \times \prod_{i=2}^{n} P(x_i)$$
$$= \mathbb{E}\left[\frac{e^{r(y)/\lambda}}{e^{r(y)/\lambda} + \sum_{i=2}^{n} e^{r(X_i)/\lambda}}\right]$$

where the last line follows from the fact that $X_2, \ldots, X_n$ are i.i.d. samples from P. Therefore:

$$P_{n,\lambda}(y) = nP(y)\mathbb{E}\left[\frac{e^{r(y)/\lambda}}{e^{r(y)/\lambda} + \sum_{i=2}^{n} e^{r(X_i)/\lambda}}\right]$$

$$= P(y)e^{r(y)/\lambda} \times \mathbb{E}\left[\frac{1}{\frac{1}{n}(e^{r(y)/\lambda} + \sum_{i=1}^{n-1} e^{r(X_i)/\lambda})}\right],$$

expectation is with respect to the joint distribution of $(X_2, \ldots, X_n)$. The inequality follows from Jensen's inequality $\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$. $\square$

We can now finally prove Theorem 1.

Theorem 5. *For $\lambda > 0$ and integer $n \geq 1$,*

$$D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) \leq \log\left(1 + \frac{1}{n}\mathrm{CV}\left(e^{r_\lambda(X)}\right)^2\right). \tag{19}$$

Consequently, $D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) = O(1/n)$. Moreover, if $0 \leq r(x) \leq 1$ for all $x \in \mathcal{X}$, then

$$D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) \leq \frac{1}{n}\sinh\left(\frac{1}{2\lambda}\right)^2. \tag{20}$$

Proof of Theorem 1:. For P_λ^* given in (2) and $P_{n,\lambda}$ given in (4):

$$D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) = \sum_x P_\lambda^*(x) \log\left(\frac{P_\lambda^*(x)}{P_{n,\lambda}(x)}\right) \tag{21}$$

$$\leq \sum_x \frac{P(x)e^{r_\lambda(x)}}{\mathbb{E}_P[e^{r_\lambda(X)}]} \log\left(\frac{1/\mathbb{E}_P[e^{r_\lambda(X)}]}{1/\left(\frac{1}{n}e^{r_\lambda(x)} + \frac{n-1}{n}\mathbb{E}_P\left[e^{r_\lambda(X)}\right]\right)}\right) \tag{22}$$

where the inequality follows from plugging in definitions for P_λ^* and $P_{n,\lambda}$ and applying (5). Rearranging the terms, we have:

$$D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) \leq \sum_x \frac{P(x)e^{r_\lambda(x)}}{\mathbb{E}_P[e^{r_\lambda(X)}]} \log\left(\frac{1}{n}\frac{e^{r_\lambda(x)}}{\mathbb{E}_P[e^{r_\lambda(X)}]} + \frac{n-1}{n}\right)$$

$$\leq \log\left[\frac{1}{n}\left(\sum_x \frac{P(x)e^{2r_\lambda(x)}}{\mathbb{E}_P[e^{r_\lambda(X)}]^2}\right) + \frac{n-1}{n}\right] \qquad \text{(Jensen's inequality)}$$

$$= \log\left[\frac{1}{n}\left(\frac{\mathbb{E}_P[e^{2r_\lambda(X)}]}{\mathbb{E}_P[e^{r_\lambda(X)}]^2} - 1\right) + 1\right].$$

Equation (19) follows by plugging in the definition of $\mathsf{CV}\left(e^{r_\lambda(X)}\right)$. Since $\log(1+x) \leq x$ (assuming base e), then $D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) = O(1/n)$.

Now assume that the reward function $r(x)$ is bounded by 1, and denote

$$\sigma^2 \triangleq \mathbb{E}_P[e^{2r_\lambda(X)}] - \mathbb{E}_P[e^{r_\lambda(X)}]^2, \ \mu = \mathbb{E}_P[e^{r_\lambda(X)}], \ c \triangleq e^{1/\lambda}.$$

From the Bhatia-Davis inequality [36], $\sigma^2 \leq (c-\mu)(\mu-1)$. Since $\mu \in [1,c]$, a simple maximization yields

$$\mathsf{CV}\left(e^{r_\lambda(X)}\right)^2 \leq \max_{\mu \in [1,c]} \frac{(c-\mu)(\mu-1)}{\mu^2} = \frac{(c-1)^2}{4c}.$$

Consequently,

$$D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) \leq \log\left(1 + \frac{1}{n}\frac{(e^{1/\lambda}-1)^2}{4e^{1/\lambda}}\right)$$

$$\leq \frac{1}{n}\frac{(e^{1/\lambda}-1)^2}{4e^{1/\lambda}}$$

$$= \frac{1}{n}\sinh\left(\frac{1}{2\lambda}\right)^2.$$

$\square$

C. Proof of Corollary 1

We begin by restating Corollary 1.

Corollary 1. *For $0 \leq r(x) \leq 1$ and fixed $n \geq 1$ and $\epsilon > 0$, if $\lambda \geq \frac{1}{\log(1+4n\epsilon)}$, then $D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) \leq \epsilon$.*

Proof. From Theorem 1, we know that when $0 \leq r(x) \leq 1$, we have $D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda}) \leq \frac{1}{n}\sinh^2(\frac{1}{2\lambda})$. For the corollary's claim to hold, we need this upper bound to be at most ϵ. Therefore, we start by writing:

$$\frac{1}{n}\sinh^2(\frac{1}{2\lambda}) \leq \epsilon \tag{23}$$

Multiplying both sides by n, we obtain $\sinh^2(\frac{1}{2\lambda}) \leq n\epsilon$. Taking the square root of both sides and noting that since $\lambda > 0$, we have $\frac{1}{2\lambda} > 0$, we can write $\sinh(\frac{1}{2\lambda}) \leq \sqrt{n\epsilon}$. Then, applying the inverse hyperbolic sine (arcsinh) to both sides yields:

$$\frac{1}{2\lambda} \leq \operatorname{arcsinh}(\sqrt{n\epsilon})$$

Now, we can use the fact that for $x > 0$, $\operatorname{arcsinh}(x) = \ln(x + \sqrt{x^2+1})$. Therefore:

$$\lambda \geq \frac{1}{2\ln(\sqrt{n\epsilon} + \sqrt{n\epsilon+1})}$$

Note that for $x > 0$, $x + \sqrt{x^2 + 1} \geq 2x$ since

$$
\begin{aligned}
(x + \sqrt{x^2 + 1})^2 &= x^2 + 2x\sqrt{x^2 + 1} + (x^2 + 1) \\
&= 2x^2 + 2x\sqrt{x^2 + 1} + 1 \\
&\geq 2x^2 + 2x^2 + 1 \quad (\text{since } \sqrt{x^2 + 1} > x \text{ for } x > 0) \\
&> 4x^2
\end{aligned}
$$

Thus, $\ln(x + \sqrt{x^2 + 1}) \geq \ln(2x)$, which implies:

$$
\frac{1}{2\ln(\sqrt{n\epsilon} + \sqrt{n\epsilon + 1})} \leq \frac{1}{\ln(1 + 4n\epsilon)}
$$

Thus, $\lambda \geq \frac{1}{\log(1+4n\epsilon)}$ is a sufficient condition to ensure $D_{KL}(P_\lambda^* \| P_{n,\lambda}) \leq \epsilon$. $\qquad\square$

D. A sequence that attains the upper bound

In Theorem 1, we established that $D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) \leq \log\left(1 + \frac{1}{n}\mathrm{CV}\left(e^{r_\lambda(X)}\right)^2\right)$, i.e., $D_{\mathsf{KL}}\left(P_\lambda^* \| P_{n,\lambda}\right) = O(1/n)$. We will now establish that this upper bound is, indeed, sharp. Let us construct a simple binary example. Let $\mathcal{X} = \{0, 1\}$ with $P(1) = p$ and $P(0) = 1 - p$ where $p \in (0, 1)$. Define the reward function:

$$
r(x) = \begin{cases} 1 & \text{if } x = 1 \\ 0 & \text{if } x = 0 \end{cases}
$$

Then:

$$
\begin{aligned}
E_P[e^{r_\lambda(X)}] &= pe^{1/\lambda} + (1 - p) \\
E_P[e^{2r_\lambda(X)}] &= pe^{2/\lambda} + (1 - p)
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
\mathrm{Var}(e^{r_\lambda(X)}) &= E_P[e^{2r_\lambda(X)}] - (E_P[e^{r_\lambda(X)}])^2 \\
&= pe^{2/\lambda} + (1 - p) - (pe^{1/\lambda} + (1 - p))^2 \\
&= pe^{2/\lambda} + (1 - p) - (p^2 e^{2/\lambda} + 2p(1 - p)e^{1/\lambda} + (1 - p)^2) \\
&= p(1 - p)e^{2/\lambda} - 2p(1 - p)e^{1/\lambda} + (1 - p)(p) \\
&= p(1 - p)(e^{1/\lambda} - 1)^2
\end{aligned}
$$

Which, in turn, implies

$$
CV(e^{r_\lambda(X)})^2 = \frac{\mathrm{Var}(e^{r_\lambda(X)})}{(E_P[e^{r_\lambda(X)}])^2} = \frac{p(1 - p)(e^{1/\lambda} - 1)^2}{(pe^{1/\lambda} + (1 - p))^2}
$$

From Lemma 1, for $x = 1$:

$$
\begin{aligned}
P_{n,\lambda}(1) &= pe^{1/\lambda} E\left[\frac{1}{\frac{1}{n}(e^{1/\lambda} + S_{n-1})}\right] \\
&= npe^{1/\lambda} E\left[\frac{1}{e^{1/\lambda} + S_{n-1}}\right]
\end{aligned}
$$

Now, we can write the Taylor expansion for $E\left[\frac{1}{e^{1/\lambda} + S_{n-1}}\right]$. Consider the function $g(x) = \frac{1}{e^{1/\lambda} + x}$ and let $S_{n-1} = \sum_{i=1}^{n-1} e^{r_\lambda(X_i)}$. Note that $g(x)$ is analytic and uniformly bounded by $1/e^{1/\lambda}$ for all $x \geq 0$. The second-order Taylor expansion of $g(x)$ around $\mu = E[S_{n-1}]$ is:

$$
g(x) = g(\mu) + g'(\mu)(x - \mu) + \frac{g''(\mu)}{2}(x - \mu)^2 + R_3(x)
$$

The law of large numbers guarantees that for any $\varepsilon > 0$,

$$
\Pr(|S_{n-1} - \mu| \geq \varepsilon) \to 0 \quad \text{as } n \to \infty. \tag{24}
$$

Define the high-probability event $\mathcal{A} = \{\, |S_{n-1} - \mu| \leq \varepsilon \,\}$. On $\mathcal{A}$, S_{n-1} lies within the interval $[\mu - \varepsilon, \mu + \varepsilon]$, a region on which the derivatives of g are well-behaved and the Taylor expansion provides a uniformly accurate approximation (i.e., there exists a constant $C > 0$ such that $|R_3(S_{n-1})| \leq C|S_{n-1} - \mu|^3$). On the complementary event $\mathcal{A}^c$, the Taylor expansion might not be as

accurate; however, the probability of $\mathcal{A}^c$ decays exponentially in n. Therefore, even if the error in the Taylor expansion is large on $\mathcal{A}^c$, its overall contribution to expectations is negligible. Computing the derivatives of g:

$$g'(x) = -\frac{1}{(e^{1/\lambda} + x)^2}$$

$$g''(x) = \frac{2}{(e^{1/\lambda} + x)^3}$$

$$g'''(x) = -\frac{6}{(e^{1/\lambda} + x)^4}$$

Therefore:

$$g(\mu) = \frac{1}{e^{1/\lambda} + (n-1)(pe^{1/\lambda} + (1-p))}$$

$$g'(\mu) = -\frac{1}{(e^{1/\lambda} + (n-1)(pe^{1/\lambda} + (1-p)))^2}$$

$$g''(\mu) = \frac{2}{(e^{1/\lambda} + (n-1)(pe^{1/\lambda} + (1-p)))^3}$$

Taking expectations:

$$
\begin{aligned}
E[g(S_{n-1})] &= g(\mu) + g'(\mu)E[(S_{n-1} - \mu)] \\
&\quad + \frac{g''(\mu)}{2}E[(S_{n-1} - \mu)^2] + E[R_3(S_{n-1})] \\
&= g(\mu) + 0 + \frac{g''(\mu)}{2}\mathrm{Var}(S_{n-1}) + E[R_3(S_{n-1})]
\end{aligned}
$$

Substituting the values:

$$
\begin{aligned}
E\left[\frac{1}{e^{1/\lambda} + S_{n-1}}\right] &= \frac{1}{e^{1/\lambda} + (n-1)(pe^{1/\lambda} + (1-p))} \\
&\quad + \frac{(n-1)p(1-p)(e^{1/\lambda} - 1)^2}{(e^{1/\lambda} + (n-1)(pe^{1/\lambda} + (1-p)))^3} \\
&\quad + E[R_3(S_{n-1})]
\end{aligned}
$$

(This last term is definitely small. In particular, smaller than $O(1/n^2)$. We will take care of it later).
A similar calculation for $x = 0$ gives:

$$P_{n,\lambda}(0) = \frac{n(1-p)}{1 + (n-1)(pe^{1/\lambda} + (1-p))} + O\left(\frac{1}{n^2}\right)$$

This leads to our four final expressions:

$$P_\lambda^*(1) = \frac{pe^{1/\lambda}}{pe^{1/\lambda} + (1-p)}$$

$$P_\lambda^*(0) = \frac{1-p}{pe^{1/\lambda} + (1-p)}$$

$$P_{n,\lambda}(1) = \frac{npe^{1/\lambda}}{e^{1/\lambda} + (n-1)(pe^{1/\lambda} + (1-p))} + O(1/n^2)$$

$$P_{n,\lambda}(0) = \frac{n(1-p)}{1 + (n-1)(pe^{1/\lambda} + (1-p))} + O(1/n^2)$$

We can now compute the KL divergence

$$D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda}) = P_\lambda^*(1) \log \frac{P_\lambda^*(1)}{P_{n,\lambda}(1)} + P_\lambda^*(0) \log \frac{P_\lambda^*(0)}{P_{n,\lambda}(0)}$$

$$= \frac{pe^{1/\lambda}}{pe^{1/\lambda} + (1-p)}$$

$$\times \log\left(\frac{pe^{1/\lambda}}{pe^{1/\lambda} + (1-p)} \cdot \frac{e^{1/\lambda} + (n-1)(pe^{1/\lambda} + (1-p))}{npe^{1/\lambda}} + O(1/n^2)\right)$$

$$+ \frac{1-p}{pe^{1/\lambda} + (1-p)}$$

$$\times \log\left(\frac{1-p}{pe^{1/\lambda} + (1-p)} \cdot \frac{1 + (n-1)(pe^{1/\lambda} + (1-p))}{n(1-p)} + O(1/n^2)\right)$$

Let's simplify the terms inside each logarithm. For the first term:

$$\frac{e^{1/\lambda} + (n-1)(pe^{1/\lambda} + (1-p))}{n(pe^{1/\lambda} + (1-p))}$$

$$= \frac{e^{1/\lambda} - (pe^{1/\lambda} + (1-p)) + n(pe^{1/\lambda} + (1-p))}{n(pe^{1/\lambda} + (1-p))}$$

$$= 1 + \frac{e^{1/\lambda} - (pe^{1/\lambda} + (1-p))}{n(pe^{1/\lambda} + (1-p))}$$

$$= 1 + \frac{(1-p)(e^{1/\lambda} - 1)}{n(pe^{1/\lambda} + (1-p))}$$

Similarly for the second term:

$$\frac{1 + (n-1)(pe^{1/\lambda} + (1-p))}{n(pe^{1/\lambda} + (1-p))}$$

$$= 1 + \frac{1 - (pe^{1/\lambda} + (1-p))}{n(pe^{1/\lambda} + (1-p))}$$

$$= 1 - \frac{p(e^{1/\lambda} - 1)}{n(pe^{1/\lambda} + (1-p))}$$

This leads to

$$D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda}) =$$

$$\frac{pe^{1/\lambda}}{pe^{1/\lambda} + (1-p)} \log\left(1 + \frac{(1-p)(e^{1/\lambda} - 1)}{n(pe^{1/\lambda} + (1-p))} + O(1/n^2)\right)$$

$$+ \frac{1-p}{pe^{1/\lambda} + (1-p)} \log\left(1 - \frac{p(e^{1/\lambda} - 1)}{n(pe^{1/\lambda} + (1-p))} + O(1/n^2)\right)$$

For the first logarithm, let's denote $a_1 = \frac{(1-p)(e^{1/\lambda}-1)}{n(pe^{1/\lambda}+(1-p))}$. Then:

$$\log(1 + a_1 + O(1/n^2)) = (a_1 + O(1/n^2))$$

$$- \frac{(a_1 + O(1/n^2))^2}{2} + O((a_1 + O(1/n^2))^3)$$

$$= a_1 + O(1/n^2) - \frac{a_1^2}{2} - O(a_1/n^2) + O(1/n^3)$$

$$= \frac{(1-p)(e^{1/\lambda} - 1)}{n(pe^{1/\lambda} + (1-p))} - \frac{(1-p)^2(e^{1/\lambda} - 1)^2}{2n^2(pe^{1/\lambda} + (1-p))^2} + O(1/n^3)$$

Similarly for the second logarithm, let $a_2 = -\frac{p(e^{1/\lambda}-1)}{n(pe^{1/\lambda}+(1-p))}$:

$$\log(1 + a_2 + O(1/n^2)) = (a_2 + O(1/n^2))$$
$$- \frac{(a_2 + O(1/n^2))^2}{2} + O((a_2 + O(1/n^2))^3)$$
$$= a_2 + O(1/n^2) - \frac{a_2^2}{2} - O(a_2/n^2) + O(1/n^3)$$
$$= -\frac{p(e^{1/\lambda} - 1)}{n(pe^{1/\lambda} + (1-p))} - \frac{p^2(e^{1/\lambda} - 1)^2}{2n^2(pe^{1/\lambda} + (1-p))^2} + O(1/n^3)$$

Multiplying by the respective coefficients and combining terms:

$$D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda}) = \frac{pe^{1/\lambda}}{pe^{1/\lambda} + (1-p)} \cdot \frac{(1-p)(e^{1/\lambda} - 1)}{n(pe^{1/\lambda} + (1-p))}$$
$$- \frac{pe^{1/\lambda}}{pe^{1/\lambda} + (1-p)} \cdot \frac{(1-p)^2(e^{1/\lambda} - 1)^2}{2n^2(pe^{1/\lambda} + (1-p))^2}$$
$$+ \frac{1-p}{pe^{1/\lambda} + (1-p)} \cdot \left(-\frac{p(e^{1/\lambda} - 1)}{n(pe^{1/\lambda} + (1-p))} \right)$$
$$- \frac{1-p}{pe^{1/\lambda} + (1-p)} \cdot \frac{p^2(e^{1/\lambda} - 1)^2}{2n^2(pe^{1/\lambda} + (1-p))^2}$$
$$+ O(1/n^3)$$

The first-order terms (in $1/n$) combine to:

$$\frac{p(1-p)(e^{1/\lambda} - 1)^2}{n(pe^{1/\lambda} + (1-p))^2}$$

The second-order terms (in $1/n^2$) combine to:

$$-\frac{p(1-p)(e^{1/\lambda} - 1)^2}{2n^2(pe^{1/\lambda} + (1-p))^2} \cdot \frac{p(1-p)(e^{1/\lambda} - 1)^2}{(pe^{1/\lambda} + (1-p))^2}$$

Therefore:

$$D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda}) = \frac{p(1-p)(e^{1/\lambda} - 1)^2}{n(pe^{1/\lambda} + (1-p))^2} + O(1/n^2)$$

This matches our earlier coefficient of variation calculation:

$$\frac{1}{n} CV(e^{r_\lambda(X)})^2 = \frac{1}{n} \frac{p(1-p)(e^{1/\lambda} - 1)^2}{(pe^{1/\lambda} + (1-p))^2}$$

E. Proof of the lower bound

We begin by restating Theorem 2.

Theorem 2. *For $\lambda > 0$ and integer $n \geq 1$, the following bounds hold on the TV distance:*

$$\mathsf{TV}(P_\lambda^* \| P_{n,\lambda}) \geq \frac{C'}{\sqrt{n-1}}, \tag{8}$$

where C' is a constant depending on λ, the reward function r, and the second and third moments of $e^{r(X)/\lambda}$.

Let

$$P_\lambda^*(x) \triangleq \frac{P(x)e^{r(x)}}{\mathbb{E}_P\left[e^{r(X)}\right]}$$

and

$$P_{n,\lambda}(x) \triangleq P(x)e^{r(x)} \times \mathbb{E}_{P^{\otimes(n-1)}}\left[\frac{n}{e^{r(x)} + \sum_{i=1}^{n-1} e^{r(X_i)}} \right],$$

where $r : \mathcal{X} \to [0, a]$. Note that a will be $e^{1/\lambda}$ once we add the dependence on λ.

Denoting $\mu \triangleq \mathbb{E}_P\left[e^{r(X)}\right]$, we have

$$\mathsf{TV}\left(P_\lambda^* \| P_{n,\lambda}\right) = \frac{1}{2} \sum_x |P_\lambda^*(x) - P_{n,\lambda}(x)|$$

$$= \frac{1}{2} \sum_x \frac{P(x)e^{r(x)}}{\mu} \times \left| 1 - \mathbb{E}_{P^{\otimes(n-1)}} \left[\frac{1}{\frac{e^{r(x)}}{n\mu} + \left(\frac{n-1}{n}\right)\left(\frac{1}{n-1} \sum_{i=1}^{n-1} \frac{e^{r(X_i)}}{\mu}\right)} \right] \right|.$$

Our goal is to bound the term

$$L_n(x) \triangleq \mathbb{E}_{P^{\otimes(n-1)}} \left[\frac{1}{\frac{e^{r(x)}}{n\mu} + \left(\frac{n-1}{n}\right)\left(\frac{1}{n-1} \sum_{i=1}^{n-1} \frac{e^{r(X_i)}}{\mu}\right)} \right]. \tag{25}$$

Denote $Y_i \triangleq \frac{e^{r(X_i)}}{\mu}$, and

$$S_k \triangleq \frac{1}{k} \sum_{i=1}^{k} Y_i. \tag{26}$$

Note that $\mathbb{E}[Y_i] = 1$, and S_k has finite variance and third moment, denoted by σ^2 and ρ, respectively. For all x, we have $S_{n-1} \geq \frac{1}{\mu}$ since $e^{r(x)}/\mu \geq 1/\mu$ for all x. This allows us to introduce the unit step function $u(t)$, we can thus rewrite $L_n(x)$ as

$$L_n(x) = \left(\frac{n}{n-1}\right) \mathbb{E} \left[\frac{1}{\frac{e^{r(x)}}{(n-1)\mu} + S_{n-1}} \right]$$

We can now calculate this expectation and bound it with the Berry-Esseen theorem. First, note that S_{n-1} is an average of i.i.d. random variables with:

- $E[Y_i] = 1$ (by definition)
- $\mathrm{Var}(Y_i) = \sigma^2$ (finite by assumption)
- $E[|Y_i - E[Y_i]|^3] = \rho$ (finite by assumption)

Then, the Berry-Esseen theorem says

Theorem 6 (Berry-Esseen). *For any $t \in \mathbb{R}$:*

$$\left| P(S_{n-1} \leq t) - \Phi\left(\frac{(t-1)\sqrt{n-1}}{\sigma}\right) \right| \leq \frac{C\rho}{\sigma^3\sqrt{n-1}}$$

where Φ is the standard normal CDF and C is an absolute constant.

As a consequence of the theorem, we have the upper and lower bounds:

$$P(S_{n-1} \leq t) \leq \Phi\left(\frac{(t-1)\sqrt{n-1}}{\sigma}\right) + \frac{C\rho}{\sigma^3\sqrt{n-1}}$$

$$P(S_{n-1} \leq t) \geq \Phi\left(\frac{(t-1)\sqrt{n-1}}{\sigma}\right) - \frac{C\rho}{\sigma^3\sqrt{n-1}}$$

Denoting $c = \frac{e^{r(x)}}{(n-1)\mu}$ and using that $E[X] = \int_0^\infty P(X > t)dt$, we have

$$\mathbb{E}\left[\frac{1}{\frac{e^{r(x)}}{(n-1)\mu} + S_{n-1}} \right] = \int_0^\infty \Pr\left[\frac{1}{c + S_{n-1}} > t \right] dt$$

$$= \int_0^{\mu/(\mu c+1)} \Pr\left[S_{n-1} < \frac{1}{t} - c \right] dt \quad \text{(since } S_{n-1} \geq \frac{1}{\mu} \text{ always)}$$

$$= \int_{\frac{1}{1-\delta+c}}^{\mu/(\mu c+1)} \Pr\left[S_{n-1} < \frac{1}{t} - c \right] dt + \int_0^{\frac{1}{1-\delta+c}} \Pr\left[S_{n-1} < \frac{1}{t} - c \right] dt$$

$$\leq \int_{\frac{1}{1-\delta+c}}^{\mu/(\mu c+1)} \left[\Phi\left(\frac{\sqrt{n-1}}{\sigma}\left(\frac{1}{t} - c - 1\right)\right) + \frac{C\rho}{\sigma^3\sqrt{n-1}} \right] dt + \int_0^{\frac{1}{1-\delta+c}} 1 dt$$

$$\leq \int_{\frac{1}{1-\delta+c}}^{\mu/(\mu c+1)} \Phi\left(\frac{\sqrt{n-1}}{\sigma}\left(\frac{1}{t} - c - 1\right)\right) dt + \frac{C\rho}{\sigma^3\sqrt{n-1}} \cdot \frac{\mu}{\mu c+1} + \frac{1}{1-\delta+c}$$

$$\mathbb{E}\left[\frac{1}{c+S_{n-1}}\right] \leq \int_{\frac{1}{1-\delta+c}}^{\mu/(\mu c+1)} \Phi\left(\frac{\sqrt{n-1}}{\sigma}\left(\frac{1}{t}-c-1\right)\right) dt + \frac{C\rho}{\sigma^3\sqrt{n-1}} \cdot \frac{\mu}{\mu c+1} + \frac{1}{1-\delta+c}$$

Now, for the main integral term, when $t > \frac{1}{1-\delta+c}$, we have $\frac{1}{t} - c - 1 < (1-\delta+c) - c - 1 = -\delta$. Therefore $\Phi\left(\frac{\sqrt{n-1}}{\sigma}(\frac{1}{t}-c-1)\right) \leq \Phi\left(-\delta\frac{\sqrt{n-1}}{\sigma}\right)$. Using the bound for normal CDF: For $z > 0$, $\Phi(-z) = 1 - \Phi(z) \leq \frac{1}{z\sqrt{2\pi}}e^{-z^2/2}$ Applying this with $z = \delta\frac{\sqrt{n-1}}{\sigma}$:

$$\int_{\frac{1}{1-\delta+c}}^{\mu/(\mu c+1)} \Phi\left(\frac{\sqrt{n-1}}{\sigma}\left(\frac{1}{t}-c-1\right)\right) dt \leq \frac{\mu}{\mu c+1} \cdot \frac{\sigma}{\delta\sqrt{2\pi(n-1)}} e^{-\delta^2(n-1)/(2\sigma^2)}$$

Therefore:

$$\mathbb{E}\left[\frac{1}{c+S_{n-1}}\right] \leq \frac{\mu}{\mu c+1} \cdot \frac{\sigma}{\delta\sqrt{2\pi(n-1)}} e^{-\delta^2(n-1)/(2\sigma^2)} + \frac{C\rho}{\sigma^3\sqrt{n-1}} \cdot \frac{\mu}{\mu c+1} + \frac{1}{1-\delta+c}$$

The exponential term decays very fast with n, so it can be absorbed into the error term. Note that δ appears as a free parameter in our bounds, with the only constraint that $\delta < 1 - \frac{1}{\mu}$. Choosing $\delta = \frac{4C'}{\sqrt{n-1}}$ for some constant C' (which is valid for large enough n since we need $\delta < 1 - \frac{1}{\mu}$), we get:

$$\mathbb{E}\left[\frac{1}{c+S_{n-1}}\right] \leq 1 - \frac{\delta}{2} + \frac{C'}{\sqrt{n-1}} + O\left(\frac{1}{n^{3/2}}\right) = 1 - \frac{2C'}{\sqrt{n-1}} + \frac{C'}{\sqrt{n-1}} + O\left(\frac{1}{n^{3/2}}\right) = 1 - \frac{C'}{\sqrt{n-1}} + O\left(\frac{1}{n^{3/2}}\right)$$

This gives us the desired lower bound on $|1 - L_n(x)|$:

$$|1 - L_n(x)| \geq \frac{C'}{\sqrt{n-1}} - O\left(\frac{1}{n^{3/2}}\right)$$

Substituting this back into the TV distance:

$$\begin{aligned}
\mathsf{TV}(P_\lambda^*\|P_{n,\lambda}) &= \frac{1}{2}\sum_x \frac{P(x)e^{r(x)}}{\mu}|1 - L_n(x)| \\
&\geq \frac{1}{2}\sum_x \frac{P(x)e^{r(x)}}{\mu}\left(\frac{C'}{\sqrt{n-1}} - O\left(\frac{1}{n^{3/2}}\right)\right) \\
&= \frac{1}{2}\left(\frac{C'}{\sqrt{n-1}} - O\left(\frac{1}{n^{3/2}}\right)\right)\sum_x \frac{P(x)e^{r(x)}}{\mu} \\
&= \frac{1}{2}\left(\frac{C'}{\sqrt{n-1}} - O\left(\frac{1}{n^{3/2}}\right)\right) \cdot 1 \\
&= \frac{C'}{2\sqrt{n-1}} - O\left(\frac{1}{n^{3/2}}\right)
\end{aligned}$$

The last step uses the fact that $\sum_x \frac{P(x)e^{r(x)}}{\mu} = 1$ since P_λ^* is a probability distribution. This lower bound on TV distance then implies through Pinsker's inequality that:

$$D_{D_{\mathsf{KL}}}(P_\lambda^*\|P_{n,\lambda}) \geq 2[\mathsf{TV}(P_\lambda^*\|P_{n,\lambda})]^2 \geq \Omega\left(\frac{1}{\sqrt{n}}\right) \tag{27}$$

F. Relative Reward Bound.

While our previous analysis focused on establishing convergence rates in KL divergence between Soft Best-of-n sampling and the optimal tilted distribution, we now turn to analyzing how quickly the expected rewards converge. Specifically, we establish that the relative difference in expected rewards decays as $O(1/n)$, matching our KL divergence bounds. This provides a complete picture of how Soft Best-of-n sampling approximates both the distribution and reward properties of the optimal solution.

Theorem 3. *Let $0 \leq r(x) \leq 1$ and $M(\lambda) \triangleq (e^{1/\lambda}/\mathbb{E}_P[e^{r_\lambda(X)}]) - 1$. Then*

$$\frac{\mathbb{E}_{P_\lambda^*}[r_\lambda(X)] - \mathbb{E}_{P_{n,\lambda}}[r_\lambda(X)]}{\mathbb{E}_{P_\lambda^*}[r_\lambda(X)]} \leq \frac{M(\lambda)}{M(\lambda) + n}. \tag{9}$$

In particular, the reward achieved by Soft Best-of-n sampling converges to that of the tilted distribution with relative error $O(1/n)$.

G. Blockwise sampling

Having analyzed Soft Best-of-n sampling for individual tokens, we now extend our analysis to sequences of tokens. In oarticular, we study how the sampling strategy performs when applied to blocks of m tokens at once rather than token-by-token. This analysis reveals fundamental trade-offs between sequence length, sample complexity, and approximation quality that are crucial for practical applications. We begin by introducing a useful Lemma, similar to Lemma 1 but for blocks of size m.

Lemma 2. *For a sequence $y^m \in \mathcal{X}^m$, $n \geq 1$, $\lambda > 0$, and a sequence of independent samples $X_1^m, \ldots, X_n^m$ drawn i.i.d. from $P_X^{\otimes m}$, let Z be the random index drawn according to step 2 of Definition 3. Then the probability that the n-tilted sampling strategy outputs y^m is given by:*

$$P_{n,\lambda}(y^m) = n \cdot P_X^{\otimes m}(y^m) \cdot \mathbb{E}\left[\frac{e^{r(y^m)/\lambda}}{e^{r(y^m)/\lambda} + \sum_{j=2}^n e^{r(X_j^m)/\lambda}}\right]$$

$$\geq P_X^{\otimes m}(y^m) e^{r(y^m)/\lambda} \cdot \frac{1}{\frac{1}{n} e^{r(y^m)/\lambda} + \frac{n-1}{n}\left(\mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}]\right)}$$

where

$$\mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}] = \left(\mathbb{E}_P[e^{r(X)/m\lambda}]\right)^m$$

due to the additive structure of the reward and independence.

Proof. Following Definition 2 for sequences, we have:

$$P_{n,\lambda}(y^m) = \sum_{i=1}^n \Pr(Y^m = y^m, Z = i)$$

$$= \sum_{i=1}^n \Pr(X_i^m = y^m, Z = i)$$

$$= \sum_{i=1}^n \Pr(X_i^m = y^m) \cdot \Pr(Z = i | X_i^m = y^m)$$

By symmetry, we can focus on $i = 1$ and multiply by n:

$$P_{n,\lambda}(y^m) = n \cdot P_X^{\otimes m}(y^m) \cdot \Pr(Z = 1 | X_1^m = y^m)$$

$$= n \cdot P_X^{\otimes m}(y^m) \cdot \mathbb{E}\left[\frac{e^{r(y^m)/\lambda}}{e^{r(y^m)/\lambda} + \sum_{j=2}^n e^{r(X_j^m)/\lambda}}\right]$$

Since $f(x) = 1/x$ is convex for $x > 0$, we can apply Jensen's inequality:

$$\mathbb{E}\left[\frac{e^{r(y^m)/\lambda}}{e^{r(y^m)/\lambda} + \sum_{j=2}^n e^{r(X_j^m)/\lambda}}\right] \geq \frac{e^{r(y^m)/\lambda}}{e^{r(y^m)/\lambda} + (n-1)\mathbb{E}[e^{r(X^m)/\lambda}]}$$

$$= \frac{1}{\frac{1}{n} + \frac{n-1}{n}\frac{\mathbb{E}[e^{r(X^m)/\lambda}]}{e^{r(y^m)/\lambda}}}$$

For the expectation term, due to the additive structure of the reward and independence:

$$\mathbb{E}[e^{r(X^m)/\lambda}] = \mathbb{E}\left[\exp\left(\frac{1}{m\lambda}\sum_{i=1}^m r(X_i)\right)\right] = \mathbb{E}\left[\prod_{i=1}^m e^{r(X_i)/m\lambda}\right] = \left(\mathbb{E}_P[e^{r(X)/m\lambda}]\right)^m$$

Combining these results yields the stated bound. $\square$

H. Convergence Upper Bound for the Blockwise Case

We begin by restating Theorem 4.

Theorem 4. *For $\lambda > 0$ and integers $n, m \geq 1$, let $P_X^{\otimes m}$ be an i.i.d. source over $\mathcal{X}^m$ and $r : \mathcal{X} \to [0, 1]$ be a reward function. Then, for the blockwise Soft Best-of-n sampling strategy:*

$$D_{\mathsf{KL}}(P_\lambda^{*m} \| P_{n,\lambda}^m) \leq \log\left(1 + \frac{1}{n}\left(\frac{\mathbb{E}_{P_X^{\otimes m}}[e^{2r_\lambda(X)/m}]}{\mathbb{E}_{P_X^{\otimes m}}[e^{r_\lambda(X)/m}]^2} - 1\right)^m\right)$$

Moreover, if $0 \leq r(x) \leq 1$ for all $x \in \mathcal{X}$, then $D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda}) \leq \frac{1}{n} \sinh\left(\frac{1}{m\lambda}\right)^{2m}$.

Proof. We begin by examining the blockwise n-tilted sampling procedure on sequences of length m. In this setting, we sample n sequences $X_1^m, \ldots, X_n^m$ independently from $P_X^{\otimes m}$ and select one according to the blockwise n-tilted sampling from Def. 3. Let us start with the KL divergence expression:

$$D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda}) = \sum_{y^m} P_\lambda^*(y^m) \log \frac{P_\lambda^*(y^m)}{P_{n,\lambda}(y^m)}$$

$$= \sum_{y^m} \frac{P_X^{\otimes m}(y^m) e^{r(y^m)/\lambda}}{\mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}]} \log \frac{P_X^{\otimes m}(y^m) e^{r(y^m)/\lambda}}{\mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}] \cdot P_{n,\lambda}(y^m)}$$

Using the lower bound from Lemma 2:

$$P_{n,\lambda}(y^m) \geq P_X^{\otimes m}(y^m) e^{r(y^m)/\lambda} \cdot \frac{1}{\frac{1}{n} e^{r(y^m)/\lambda} + \frac{n-1}{n} \mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}]}$$

Substituting this bound:

$$D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda}) \leq \sum_{y^m} \frac{P_X^{\otimes m}(y^m) e^{r(y^m)/\lambda}}{\mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}]} \log \frac{\frac{1}{n} e^{r(y^m)/\lambda} + \frac{n-1}{n} \mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}]}{\mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}]}$$

$$= \sum_{y^m} \frac{P_X^{\otimes m}(y^m) e^{r(y^m)/\lambda}}{\mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}]} \log \left(1 + \frac{1}{n} \left(\frac{e^{r(y^m)/\lambda}}{\mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}]} - 1 \right) \right)$$

Since $\log$ is concave, we can apply Jensen's inequality:

$$\leq \log \left(1 + \frac{1}{n} \sum_{y^m} \frac{P_X^{\otimes m}(y^m) e^{r(y^m)/\lambda}}{\mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}]} \left(\frac{e^{r(y^m)/\lambda}}{\mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}]} - 1 \right) \right)$$

$$= \log \left(1 + \frac{1}{n} \left(\frac{\mathbb{E}_{P_X^{\otimes m}}[e^{2r(X^m)/\lambda}]}{(\mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}])^2} - 1 \right) \right)$$

For the additive reward model where $r(x^m) = \frac{1}{m} \sum_{i=1}^m r(x_i)$, using the independence of samples:

$$\frac{\mathbb{E}_{P_X^{\otimes m}}[e^{2r(X^m)/\lambda}]}{(\mathbb{E}_{P_X^{\otimes m}}[e^{r(X^m)/\lambda}])^2} - 1 = \frac{\mathbb{E}_{P_X^{\otimes m}}[e^{\frac{2}{m\lambda} \sum_{i=1}^m r(X_i)}]}{(\mathbb{E}_{P_X^{\otimes m}}[e^{\frac{1}{m\lambda} \sum_{i=1}^m r(X_i)}])^2} - 1$$

$$= \left(\frac{\mathbb{E}_P[e^{2r(X)/m\lambda}]}{(\mathbb{E}_P[e^{r(X)/m\lambda}])^2} \right)^m - 1$$

$$= \mathsf{CV}(e^{r(X)/m\lambda})^{2m}$$

where $\mathsf{CV}(e^{r(X)/m\lambda})^2 = \frac{\mathbb{E}_P[e^{2r(X)/m\lambda}]}{(\mathbb{E}_P[e^{r(X)/m\lambda}])^2} - 1$ is the squared coefficient of variation. Therefore:

$$D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda}) \leq \log \left(1 + \frac{1}{n} \mathsf{CV}(e^{r(X)/m\lambda})^{2m} \right)$$

$\square$

I. Proof of Corollary 2

We begin by restating Corollary 2.

Corollary 2. *For $\epsilon > 0$, if $0 \leq r(x) \leq 1$ and $\lambda \geq \frac{2}{\log(1+4(n\epsilon)^{1/m})}$, then $D_{\mathsf{KL}}(P_\lambda^{*m} \| P_{n,\lambda}^m) \leq \epsilon$.*

Proof. From Theorem 4, we have:

$$D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda}) \leq \log \left(1 + \frac{1}{n} \mathsf{CV}(e^{r(X)/m\lambda})^{2m} \right)$$

When $0 \leq r(x) \leq 1$, we can bound the coefficient of variation using properties of exponential functions and the Bhatia-Davis inequality:

$$\mathrm{CV}(e^{r(X)/m\lambda})^2 = \frac{\mathbb{E}_P[e^{2r(X)/m\lambda}]}{(\mathbb{E}_P[e^{r(X)/m\lambda}])^2} - 1$$

$$\leq \frac{e^{2/m\lambda} - 1}{4}$$

In our case, let $Y = e^{r(X)/m\lambda}$. Since $0 \leq r(x) \leq 1$, we have:

$$m = \min_x e^{r(x)/m\lambda} = e^0 = 1$$

$$M = \max_x e^{r(x)/m\lambda} = e^{1/m\lambda}$$

Let $\mu = \mathbb{E}[Y]$. By Bhatia-Davis inequality:

$$\mathrm{Var}(Y) \leq (e^{1/m\lambda} - \mu)(\mu - 1)$$

The right-hand side is quadratic in μ, and its maximum occurs at $\mu = \frac{e^{1/m\lambda}+1}{2}$. At this value:

$$\mathrm{Var}(Y) \leq \left(e^{1/m\lambda} - \frac{e^{1/m\lambda} + 1}{2}\right)\left(\frac{e^{1/m\lambda} + 1}{2} - 1\right)$$

$$= \left(\frac{e^{1/m\lambda} - 1}{2}\right)\left(\frac{e^{1/m\lambda} - 1}{2}\right)$$

$$= \frac{(e^{1/m\lambda} - 1)^2}{4}$$

Now, recall that the coefficient of variation is:

$$\mathrm{CV}(Y)^2 = \frac{\mathbb{E}[Y^2]}{(\mathbb{E}[Y])^2} - 1 = \frac{\mathrm{Var}(Y)}{(\mathbb{E}[Y])^2} = \frac{\mathbb{E}[e^{2r(X)/m\lambda}]}{(\mathbb{E}[e^{r(X)/m\lambda}])^2} - 1$$

Therefore:

$$\mathrm{CV}(e^{r(X)/m\lambda})^2 = \frac{\mathrm{Var}(Y)}{(\mathbb{E}[Y])^2} \leq \frac{(e^{1/m\lambda} - 1)^2}{4(\mathbb{E}[Y])^2} \leq \frac{e^{2/m\lambda} - 1}{4}$$

where the last inequality follows from $\mathbb{E}[Y] \geq 1$ (since $Y \geq 1$ almost surely).
Therefore:

$$D_{\mathsf{KL}}(P_\lambda^* \| P_{n,\lambda}) \leq \log\left(1 + \frac{1}{n}\left(\frac{e^{2/m\lambda} - 1}{4}\right)^m\right) \leq \epsilon$$

For this inequality to hold, it suffices to have:

$$1 + \frac{1}{n}\left(\frac{e^{2/m\lambda} - 1}{4}\right)^m \leq e^\epsilon$$

$$\frac{1}{n}\left(\frac{e^{2/m\lambda} - 1}{4}\right)^m \leq e^\epsilon - 1$$

$$\left(\frac{e^{2/m\lambda} - 1}{4}\right)^m \leq n(e^\epsilon - 1)$$

$$\frac{e^{2/m\lambda} - 1}{4} \leq (n(e^\epsilon - 1))^{1/m}$$

$$e^{2/m\lambda} \leq 1 + 4(n(e^\epsilon - 1))^{1/m}$$

$$\frac{2}{m\lambda} \leq \log(1 + 4(n(e^\epsilon - 1))^{1/m})$$

Solving for λ:

$$\lambda \geq \frac{2}{m \log(1 + 4(n(e^\epsilon - 1))^{1/m})}$$

□