Information Theoretic Guarantees For Policy Alignment In Large Language Models

Youssef Mroueh

IBM Research
mroueh@us.ibm.com

Abstract

Policy alignment of large language models refers to constrained policy optimization, where the policy is optimized to maximize a reward while staying close to a reference policy with respect to an f-divergence such as the KL divergence. The best of n alignment policy selects a sample from the reference policy that has the maximum reward among n independent samples. For both cases (policy alignment and best of n), recent works showed empirically that the reward improvement of the aligned policy on the reference one scales like $\sqrt{\mathsf{KL}}$, with an explicit bound in n on the KL for the best of n policy. We show in this paper that the $\sqrt{\mathsf{KL}}$ information theoretic upper bound holds if the reward under the reference policy has sub-gaussian tails.Moreover, we prove for the best of n policy, that the KL upper bound can be obtained for any f-divergence via a reduction to exponential order statistics owing to the Rényi representation of order statistics, and a data processing inequality. If additional information is known on the tails of the aligned policy we show that tighter control on the reward improvement can be obtained via the Rényi divergence. Finally we demonstrate how these upper bounds transfer from proxy rewards to golden rewards which results in a decrease in the golden reward improvement due to overestimation and approximation errors of the proxy reward.

1 Introduction

Aligning Large Language Models (LLMs) with human preferences allows a tradeoff between maintaining the utility of the pre-trained reference model and the alignment of the model with human values such as safety or other socio-technical considerations. Alignment is becoming a crucial step in LLMs training pipeline, especially as these models are leveraged in decision making as well as becoming more and more accessible to the general public. Policy alignment starts by learning a reward model that predicts human preferences, these reward models are typically fine-tuned LLMs that are trained on pairwise human preference data [Christiano et al., 2017, Stiennon et al., 2020, Ouyang et al., 2022, Bai et al., 2022]. The reward is then optimized using *training time alignment* i.e via policy gradient based reinforcement learning leading to the so called *Reinforcemnent Learning from Human Feedback* (RLHF) [Christiano et al., 2017]. RLHF ensures that the reward is maximized while the policy π stays close to the initial reference policy π_{ref} in the sense of the Kullback-Leibler divergence $\mathsf{KL}(\pi||\pi_{\text{ref}})$. Other variants of these training time alignment have been proposed via direct preference optimization [Rafailov et al., 2024] [Zhao et al., 2023] [Ethayarajh et al., 2024]. Another important paradigm for optimizing the reward is *test time alignment* via best of n sampling from the reference policy and retaining the sample that maximizes the reward. The resulting policy is known as the *best of n policy*. The best of n policy is also used in controlled decoding settings [Yang and Klein, 2021, Mudgal et al., 2023] and in fine-tuning LLMs to match the best of n policy responses [Touvron et al., 2023].

38th Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024).

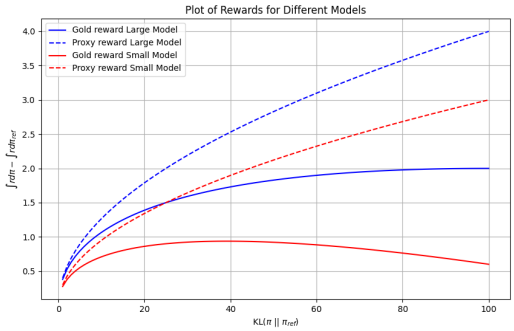

Figure 1: Qualitiative plot of centered rewards vs. KL of Proxy and Gold Rewards for both Best of n and RL policies. (See Fig. 1 a) and b) in [Gao et al., 2023] for scaling laws in policy alignment).

[Gao et al., 2023] and [Hilton and Gao, 2022] studied the scaling laws of reward models optimization in both the RL and the best of n setups. [Gao et al., 2023] distinguished between "golden reward" that can be thought of as the golden human preference and "proxy reward" which is trained to predict the golden reward. For proxy rewards [Gao et al., 2023] found experimentally for both RL and best of n policies that the reward improvement on the reference policy scales as $\sqrt{\mathsf{KL}(\pi||\pi_{\text{ref}})}$. Similar observations for reward improvement scaling in RL were made in [Bai et al., 2022]. For golden rewards, [Gao et al., 2023] showed for both RL and best of n policies that LLMs that optimize the proxy reward suffer from over-optimization in the sense that as the policy drifts from the reference policy, optimizing the proxy reward results in deterioration of the golden reward. This phenomena is referred to in [Gao et al., 2023] [Hilton and Gao, 2022] as Goodhart's law. A qualitative plot of scaling laws discovered in [Gao et al., 2023] is given in Figure 1. For the best of n policy, most works in this space assumed that $\mathsf{KL}(\pi||\pi_{\text{ref}}) = \log(n) - \frac{n-1}{n}$ [Stiennon et al., 2020, Coste et al., 2024, Nakano et al., 2021, Go et al., 2024, Gao et al., 2023]. Recently Beirami et al. [2024] showed that this is in fact an inequality under the assumption that the reward is one to one map (a bijection) and for finite alphabets. The main contribution of this papers are :

1. We provide in Theorem 1 in Section 2 a new proof for the best of n policy inequality $\mathsf{KL}(\pi||\pi_{\text{ref}}) \leq \log(n) - \frac{n-1}{n}$ and show via a reduction to exponential random variables that it is a consequence of the data processing inequality of the KL divergence. We extend this inequality beyond the setup of [Beirami et al., 2024] of one to one rewards and finite alphabets to a more realistic setup of surjective rewards and beyond finite alphabets. We also give conditions under which the equality is met, and extend those inequalities to f-divergences and Rényi divergences.
2. We show in Section 3 that the scaling laws on policy improvement versus KL of [Gao et al., 2023] are information theoretic upper bounds and are consequences of transportation inequalities with the KL divergence under sub-gaussian tails of the reward under the reference policy. We discuss how the dependency on KL is driven only by the tails of the reward under the reference model, and cannot be improved by a better alignment algorithm and can only be improved if the tails of the reference rewards are fatter than sub-gaussian such as sub-gamma or sub-exponential tails.
3. We study in Theorem 4 the tightness of these information theoretical upper bounds when the tails of the optimized policy are also known via new transportation inequalities for the Rényi divergence D_α for in $\alpha \in (0, 1)$, and show that the upper bound $\sqrt{\mathsf{KL}}$ can not be met, echoing Goodhart's law of [Gao et al., 2023].
4. We finally study in Section 4 the transfer of transportation inequalities from proxy rewards to golden rewards and prove that indeed the golden reward improvement is hindered by "overestimation" of the proxy reward as reported empirically in [Gao et al., 2023].

2 The Alignment Problem

2.1 RLHF: A Constrained Policy Optimization Problem

Let $\mathcal{X}$ be the space of prompts and $\mathcal{Y}$ be the space of responses $y \in \mathcal{Y}$ from a LLM conditioned on a prompt $x \in \mathcal{X}$. The reference LLM is represented as policy $\pi_{\text{ref}}(y|x)$, i.e as a conditional probability

on $\mathcal{Y}$ given a prompt $x \in \mathcal{X}$. Let $\rho_{\mathcal{X}}$ be a distribution on prompts, and a r a reward, $r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, r represents a safety or alignment objective that is desirable to maximize.
Given a reference policy π_{ref}, the goal of alignment is to find a policy π^* that maximizes the reward r and that it is still close to the original reference policy for some positive $\Delta > 0$:

$$\pi^*_{y|x} = \arg\max_{\pi_{y|x}} \mathbb{E}_{x\sim\rho_{\mathcal{X}}} \mathbb{E}_{y\sim\pi(.|x)} r(x,y) \text{ s.t } \int_{\mathcal{X}} \mathsf{KL}(\pi(y|x)||\pi_{\text{ref}}(y|x)) d\rho_{\mathcal{X}}(x) \leq \Delta, \tag{1}$$

where $\mathsf{KL}(\pi(y|x)||\pi_{\text{ref}}(y|x)) = \mathbb{E}_{y\sim\pi.|x} \log\left(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}\right)$. With some abuse of notation, we write $\pi(x,y) = \pi(y|x)\rho_{\mathcal{X}}(x)$ and $\pi_{\text{ref}}(x,y) = \pi_{\text{ref}}(y|x)\rho_{\mathcal{X}}(x)$. Let $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ be joint probability defined on $\mathcal{X}\times\mathcal{Y}$ that has $\rho_{\mathcal{X}}$ as marginal on $\mathcal{X}$. Hence we can write the alignment problem (1) in a more compact way as follows:

$$\sup_{\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \int r d\pi \text{ s.t } \mathsf{KL}(\pi||\pi_{\text{ref}}) \leq \Delta. \tag{2}$$

For $\beta > 0$, we can also write a penalized form of this constrained policy optimization problem as follows: $\sup_{\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \int r d\pi - \frac{1}{\beta}\mathsf{KL}(\pi||\pi_{\text{ref}})$. It is easy to see that the optimal policy of the penalized problem is given by:

$$\pi_{\beta,r}(y|x) = \frac{\exp(\beta r(x,y))\pi_{\text{ref}}(y|x)}{\int \exp(\beta r(x,y)) d\pi_{\text{ref}}(y|x)}, \rho_{\mathcal{X}}\text{almost surely.} \tag{3}$$

The constrained problem (2) has a similar solution (See for e.g [Yang et al., 2024]):

$$\pi_{\lambda_\Delta,r}(y|x) = \frac{\exp(\frac{r(x,y)}{\lambda_\Delta})\pi_{\text{ref}}(y|x)}{\int \exp(\frac{r(x,y)}{\lambda_\Delta}) d\pi_{\text{ref}}(y|x)}, \rho_{\mathcal{X}}\text{almost surely,} \tag{4}$$

where $\lambda_\Delta > 0$ is a lagrangian that satisfies $\int_{\mathcal{X}} \mathsf{KL}(\pi_{\lambda_\Delta,r}(y|x)||\pi_{\text{ref}}(y|x)) d\rho_{\mathcal{X}}(x) = \Delta$.

2.2 Best of n Policy Alignment

Let X be the random variable associated with prompts such that $\text{Law}(X) = \rho_{\mathcal{X}}$. Let Y be the random variable associated with the conditional response of π_{ref} given X. Define the conditional reward of the reference policy :

$$R(Y)|X := r(X,Y) \text{ where Y } \sim \pi_{\text{ref}}(.|X),$$

we assume that $R(Y)|X$ admits a CDF denoted as $F_{R(Y)|X}$ and let $F^{-1}_{R(Y)|X}$ be its quantile:

$$F^{(-1)}_{R(Y)|X}(p) = \inf\{\eta : F_{R(Y)|X}(\eta) \geq p\} \text{ for } p \in [0,1].$$

Let $Y_1 \dots Y_n$ be independent samples from $\pi_{\text{ref}}(.|X)$. We define the best of n reward as follows:

$$R^{(n)}(Y)|X = \max_{i=1\dots n} R(Y_i)|X, \tag{5}$$

this the maximum of n iid random variables with a common CDF $F_{R(Y)|X}$. The best of n policy corresponds to $Y^{(n)}|X := \arg\max_{i=1\dots n} r(X,Y_i)$. We note $\pi^{(n)}_{r,\text{ref}}(.|X)$ the law of $Y^{(n)}|X$. $\pi^{(n)}_{r,\text{ref}}$ is referred to as the best of n alignment policy. We consider two setups for the reward:

Assumption 1. *We assume that the reward r is a one to one map for a fixed x, and admits an inverse $h_x : \mathbb{R} \to \mathcal{Y}$ such that $h_x(r(x,y)) = y$.*

This assumption was considered in [Beirami et al., 2024]. Nevertheless this assumption is strong and not usually meet in practice, we weaken this assumption to the following:

Assumption 2. *We assume that there is a stochastic map H_X such that $H_X(R_{Y|X}) \overset{d}{=} Y|X$ and $H_X(R_{Y^{(n)}|X}) \overset{d}{=} Y^{(n)}|X$.*

Under Assumption 2, the reward can be surjective which is more realistic but we assume that there is a stochastic map that ensures invertibility not point-wise but on a distribution level. Our assumption means that we have conditionally on X: $R|X \to Y|X$ form a Markov chain i.e exists $A(Y|R,X)$ so that $P_{Y|X} = A(Y|R,X)P_{R|X}$, and $P_{Y^{(n)}|X} = A(Y|R,X)P_{R^{(n)}|X}$.

Best of n Policy KL Guarantees: A reduction to Exponentials. In what follows for random variables Z, Z' with laws $p_Z, p_{Z'}$ we write interchangeably: $\mathsf{KL}(p_Z||p_{Z'}) = \mathsf{KL}(Z||Z')$. Let us start by looking at $\mathsf{KL}\left[R^{(n)}(Y)||R(Y)\middle|X\right]$ the KL divergence between the conditional reward of the best of n policy and that of the reference policy. Let $E \sim Exp(1)$, the optimal transport map $F^{-1}_{R(Y)|X} \circ F_E$ from the exponential distribution E to $R(Y)|X$ (See for example Theorem 2.5 in [Santambrogio, 2015]: E is atomless, but $R(Y)|X$ can be discrete valued) allows us to write:

$$R(Y)|X \overset{d}{=} F^{-1}_{R(Y)|X} \circ F_E(E), \tag{6}$$

where $\overset{d}{=}$ means equality in distribution. On the other hand, let $R^{(1)}(Y)|X \leq \cdots \leq R^{(n)}(Y)|X$ be the order statistics of the rewards of n independent samples $Y_i, i = 1 \dots n$, $Y_i \sim \pi_{\text{ref}}(.|X)$. The order statistics refer to sorting the random variable from the minimum (index (1)) to the maximum (index (n)). Consider n independent exponential $E_1, \dots E_n$, where $E_i \sim \exp(1)$, and their order statistics $E^{(1)} \leq E^{(2)} \leq \dots E^{(n)}$. The Rényi representation of order statistics [Rényi, 1953], similar to the Optimal Transport (OT) representation allows us to express the distribution of the order statistics of the rewards in terms of the order statistics of exponentials as follows:

$$\left(R^{(1)}(Y)|X, \dots, R^{(n)}(Y)|X\right) \overset{d}{=} \left(F^{-1}_{R(Y)|X} \circ F_E(E^{(1)}), \dots, F^{-1}_{R(Y)|X} \circ F_E(E^{(n)})\right). \tag{7}$$

The central idea in the Rényi representation is that the mapping $F^{-1}_{R(Y)|X} \circ F_E$ is monotonic and hence ordering preserving and by the OT representation each component is distributed as $R(Y)|X$. See [Boucheron and Thomas, 2012] for more account on the Rényi representation of order statistics.

Hence using the OT representation in (6) and the Rényi representation of the maximum (7), we can reduce the KL between the rewards to a KL on functions of exponentials and their order statistics:

$$\begin{aligned}\mathsf{KL}\left[R^{(n)}(Y)||R(Y)\middle|X\right] &= \mathsf{KL}\left(F^{-1}_{R(Y)|X} \circ F_E(E^{(n)})\middle|\middle|F^{-1}_{R(Y)|X} \circ F_E(E)\right) \\ &= \mathsf{KL}(T_X(E^{(n)})||T_X(E)),\end{aligned} \tag{8}$$

where $T_X = F^{-1}_{R(Y)|X} \circ F_E = F^{-1}_{(r(X,.))_\sharp \pi_{\text{ref}}(.|X)} \circ F_E$.

Under Assumption 1 we can write samples from the best of n policy as $Y^{(n)}|X = h_X(R^n(Y))|X$ and from the reference policy as $Y|X = h_X(R(Y))|X$. Hence we have by the data processing inequality (DPI) for the KL divergence (See for e.g [Polyanskiy and Wu, 2023]) under Assumption 1:

$$\begin{aligned}\mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{\text{ref}}|X) &= \mathsf{KL}(Y^{(n)}||Y|X) \\ &= \mathsf{KL}(h_X(R^n(Y))||h_X(R(Y))|X) \\ &= \mathsf{KL}(R^n(Y)||R(Y)|X) \text{ By Assumption 1 } h_X \text{ is one to one and DPI is an equality} \\ &= \mathsf{KL}(T_X(E^{(n)})||T_X(E)) \text{ Rényi and Optimal Transport Representations (Eq (8))}\end{aligned} \tag{9}$$

Recall that $T_X = F^{-1}_{R(Y)|X} \circ F_E$, F_E is one to one. If the space $\mathcal{Y}$ is finite, $R(Y|X)$ has a discontinuous CDF hence not strictly monotonic. It follows that its quantile $F^{-1}_{R(Y)|X}$ is not a one to one map and T_X as a result is not a one to one map and hence we have by DPI (that is an inequality in this case since T_X is not one to one):

$$\mathsf{KL}(T_X(E^{(n)})||T_X(E)) \leq \mathsf{KL}(E^{(n)}||E) \tag{10}$$

If the space $\mathcal{Y}$ is infinite and we assume that $R(Y|X)$ is continuous and strictly monotonic then $F^{-1}_{R(Y)|X}$ is a one to one map, and as a result T_X is a one to one map and the DPI is an equality in this case:

$$\mathsf{KL}(T_X(E^{(n)})||T_X(E)) = \mathsf{KL}(E^{(n)}||E) \tag{11}$$

Hence under Assumption 1 and for $\mathcal{Y}$ finite combining (9) and (10) we have:

$$\mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{\text{ref}}|X) \leq \mathsf{KL}(E^{(n)}||E), \tag{12}$$

and under Assumption 1 and for $\mathcal{Y}$ infinite and assuming $F_{R(Y)|X}$ is continuous and strictly monotonic, combining (9) and (11) we have:

$$\mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{\text{ref}}|X) = \mathsf{KL}(E^{(n)}||E). \tag{13}$$

Under the more realistic Assumption 2 we can also apply the DPI on the stochastic map H_X, since DPI also holds for stochastic maps (under our assumption $R|X \to Y|X$ see for example [van Erven and Harremos, 2014] Example 2)

$$\begin{aligned}\mathsf{KL}(\pi_{r,\text{ref}}^{(n)}||\pi_{\text{ref}}|X) &= \mathsf{KL}(H_X(R^n(Y))||H_X(R(Y))|X)) \\ &\leq \mathsf{KL}(R^n(Y)||R(Y)|X) = \mathsf{KL}(T_X(E^{(n)})||T_X(E)),\end{aligned} \tag{14}$$

and hence under Assumption 2 regardless whether T_X is a one to one map or not, thus we have: $\mathsf{KL}(\pi_{r,\text{ref}}^{(n)}||\pi_{\text{ref}}|X) \leq \mathsf{KL}(E^{(n)}||E)$. The following Lemma gives a closed form expression for $\mathsf{KL}(E^{(n)}||E)$:

Lemma 1 (KL Between Exponential and Maximum of Exponentials)**.** *Let $E \sim \exp(1)$, and $E_1, \dots E_n$ be iid exponentials and $E^{(n)}$ their maximum, we have:*

$$\mathsf{KL}(E^{(n)}||E) = \log(n) - \frac{n-1}{n}. \tag{15}$$

Hence we conclude with the following result:

Theorem 1. *The best of n policy satisfies under (i) Assumption 1 (reward one to one) and for finite $\mathcal{Y}$ or under (ii) Assumption 2 (existence of stochastic "inverse") :*

$$\mathsf{KL}(\pi_{r,\text{ref}}^{(n)}||\pi_{\text{ref}}) \leq \mathsf{KL}(E^{(n)}||E) = \log(n) - \frac{n-1}{n}. \tag{16}$$

Under Assumption 1, for infinite $\mathcal{Y}$ and assuming $F_{R(Y|X)}$ is continuous and strictly increasing for all X we have:

$$\mathsf{KL}(\pi_{r,\text{ref}}^{(n)}||\pi_{\text{ref}}) = \mathsf{KL}(E^{(n)}||E) = \log(n) - \frac{n-1}{n}. \tag{17}$$

Proof. Combining Lemma 1, the analysis above and taking expectation on X we obtain the result. □

Beirami et al. [2024] showed this result under condition (i) which is not a realistic setting and used the finiteness of $\mathcal{Y}$ to provide a direct proof. Our analysis via chaining DPI and using OT and Rényi representations to reduce the problem to exponentials allows us to extend the result to a more realistic setup under condition (ii) i.e the existence of a stochastic "inverse", without any assumption on $\mathcal{Y}$. Furthermore we unveil under which conditions the equality holds that was assumed to hold in previous works [Stiennon et al., 2020] [Coste et al., 2024, Nakano et al., 2021, Go et al., 2024] [Hilton and Gao, 2022] [Gao et al., 2023] .

Our approach of reduction to exponentials using Rényi representation of order statistics and data processing inequalities extends to bounding the f- divergence $\mathrm{D}_f(\pi_{r,\text{ref}}^{(n)}||\pi_{\text{ref}})$ as well as the α Rényi divergence. The Rényi divergence for $\alpha \in (0,1) \cup (1,\infty)$ is defined as follows:

$$D_\alpha(P||Q) = \frac{1}{(\alpha-1)} \log\left(\int p^\alpha(x) q^{1-\alpha}(x) dx\right)$$

the limit as $\alpha \to 1$ coincides with KL, i.e: $D_1(P||Q)) = \mathsf{KL}(P||Q)$. These bounds are summarized in Table 1. Full proofs and theorems are in the Appendix.

Divergence	$f(x)$	Bound on $\mathrm{D}_f(\pi_{r,\text{ref}}^{(n)}||\pi_{\text{ref}})$
KL	$x\log(x)$	$\log(n) - \frac{n-1}{n}$
Chi-squared	$(x-1)^2$	$\frac{(n-1)^2}{2n-1}$
Total Variation	$f(x) = \frac{1}{2}\lvert x-1\rvert$	$(\frac{1}{n})^{\frac{1}{n-1}} - (\frac{1}{n})^{\frac{n}{n-1}}$
Hellinger distance	$(1-\sqrt{x})^2$	$2\frac{(1-\sqrt{n})^2}{n+1}$
Forward KL	$-\log(x)$	$n-1-\log(n)$
α Rényi Divergence	NA	$\frac{1}{(\alpha-1)}\log\left(\frac{n^\alpha}{\alpha(n-1)+1}\right)$

Table 1: Best of n policy f-Divergence and α Rényi Divergence Bounds.

Best of n-Policy Dominance on the Reference Policy. The following proposition shows that the best of n policy leads to an improved reward on average:

Proposition 1. *$R^{(n)}$ dominates R in the first order dominance that is $R^{(n)}$ dominates R on all quantiles: $Q_{R^{(n)}}(t) \geq Q_R(t), \forall t \in [0,1]$. It follows that we have $\mathbb{E}R^{(n)} \geq \mathbb{E}R$.*

Best of n Policy and RL Policy The following proposition discusses the sub-optimality of the best of n policy with respect to the alignment RL objective given in (1):

Proposition 2. *Assume a bounded reward in $[-M, M]$. For $\Delta > 0$ and $n = \exp(\Delta)$ the best of n policy $\pi^{(n)}_{r,\text{ref}}$ and the Δ Constrained RL policy $\pi_{\lambda_\Delta, r}$ (given in (4)) satisfy:*

$$\mathsf{KL}(\pi^{(n)}_{r,\text{ref}} || \pi_{\lambda_\Delta, r}) \leq \frac{\sqrt{2\pi} M (e^{\frac{2M}{\lambda_\Delta}} - 1)}{\lambda_\Delta} \exp(-\frac{\Delta}{2}).$$

A similar asymptotic result appeared in [Yang et al., 2024] for $\Delta \to \infty$, showing as $n \to \infty$, $\mathsf{KL}(\pi^{(n)}_{r,\text{ref}} || \pi_{\lambda_\Delta, r}) \to 0$, we provide here a non asymptotic result for finite n and finite Δ.

3 Reward Improvement Guarantees Through Transportation Inequalities

Notations Let X be a real random variable. The logarithmic moment generating function of X is defined as follows for $\lambda \in \mathbb{R}$: $\psi_X(\lambda) = \log \mathbb{E}_X e^{\lambda(X - \mathbb{E}X)}$. X is said to be sub-Gaussian with variance σ^2 if : $\psi_X(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}$ for all $\lambda \in \mathbb{R}$. We denote $\mathsf{SubGauss}(\sigma^2)$ the set of sub-Gaussian random variables with variance σ^2_{ref}. X is said to be sub-Gamma on the right tail with variance factor σ^2 and a scale parameter $c > 0$ if : $\psi_X(\lambda) \leq \frac{\lambda^2 \sigma^2}{2(1 - c\lambda)}$ for every λ such that $0 < \lambda < \frac{1}{c}$. We denote $\mathsf{SubGamma}(\sigma^2, c)$ the set of left and right tailed sub-Gamma random variables. Sub-gamma tails can be thought as an interpolation between sub-Gaussian and sub-exponential tails.

Scaling Laws in Alignment It has been observed empirically [Coste et al., 2024, Nakano et al., 2021, Go et al., 2024, Hilton and Gao, 2022, Gao et al., 2023] that optimal RL policy $\pi_{\lambda_\Delta, r}$ satisfy the following inequality for a constant σ^2_{ref} :$\mathbb{E}_{\pi_{\lambda_\Delta, r}} r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{2\sigma^2_{\text{ref}} \mathsf{KL}(\pi_{\lambda_\Delta, r} || \pi_{\text{ref}})}$. A similar scaling for best of n policy : $\mathbb{E}_{\pi^{(n)}_{r,\text{ref}}} r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{2\sigma^2_{\text{ref}} \left(\log n - \frac{n-1}{n}\right)}$, and those bounds are oftentimes tight even when empirically estimated from samples. This hints that those bounds are information theoretic and independent of the alignment problem. Indeed if the reward was bounded, a simple application of Pinsker inequality gives rise to $\sqrt{\mathsf{KL}}$ scaling. Let TV be the total variation distance, we have: $\mathsf{TV}(\pi, \pi_{\text{ref}}) = \frac{1}{2} \sup_{||r||_\infty \leq 1} \mathbb{E}_\pi r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{\frac{1}{2} \mathsf{KL}(\pi || \pi_{\text{ref}})}$. Hence we can deduce that for bounded rewards r with norm infinity $||r||_\infty$ that:

$$\mathbb{E}_\pi r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{2||r||^2_\infty \mathsf{KL}(\pi || \pi_{\text{ref}})}.$$

Nevertheless this boundedness assumption on the reward is not realistic, since most reward models are unbounded: quoting Lambert et al. [2024] " implemented by appending a linear layer to predict one logit or removing the final decoding layers and replacing them with a linear layer" and hence the reward is unbounded by construction. We will show in what follows that those scalings laws are tied to the tails of the reward under the reference policy and are instances of transportation inequalities.

3.1 Transportation Inequalities with KL Divergence

For a policy $\pi \in \mathcal{P}(\mathcal{Y})$ and for a reward function $r : \mathcal{Y} \to \mathbb{R}$, we note $r_\sharp \pi$, the push-forward map of π through r. The reader is referred to Appendix D.1 for background on transportation inequalities and how they are derived from the so-called Donsker-Varadhan variational representation of the KL divergence. The following Proposition is an application of Lemma 4.14 in [Boucheron et al., 2013]):

Proposition 3 (Transportation Inequalities)**.** *The following inequalities hold depending on the tails of $r_\sharp \pi_{\text{ref}}$:*

1. *Assume that $r_\sharp \pi_{\text{ref}} \in \mathsf{SubGauss}(\sigma^2_{\text{ref}})$. For any $\pi \in \mathcal{P}(\mathcal{Y})$ that is absolutely continuous with respect to π_{ref}, and such that $\mathsf{KL}(\pi || \pi_{\text{ref}}) < \infty$ then we have:*

$$|\mathbb{E}_\pi r - \mathbb{E}_{\pi_{\text{ref}}} r| \leq \sqrt{2\sigma^2_{\text{ref}} \mathsf{KL}(\pi || \pi_{\text{ref}})}.$$

2. *Assume that* $r_\sharp \pi_{\text{ref}} \in \mathsf{SubGamma}(\sigma^2_{\text{ref}}, c)$. *For any* $\pi \in \mathcal{P}(\mathcal{Y})$ *that is absolutely continuous with respect to* π_{ref}, *and such that* $\mathsf{KL}(\pi||\pi_{\text{ref}}) < \infty$ *then we have:*

$$|\mathbb{E}_\pi r - \mathbb{E}_{\pi_{\text{ref}}} r| \leq \sqrt{2\sigma^2_{\text{ref}} \mathsf{KL}(\pi||\pi_{\text{ref}})} + c\mathsf{KL}(\pi||\pi_{\text{ref}})$$

In particular we have the following Corollary:

Corollary 1 (Expected Reward Improvement)**.** *If* $r_\sharp \pi_{\text{ref}} \in \mathsf{SubGauss}(\sigma^2_{\text{ref}})$ *the following holds for the optimal RL policy* $\pi_{\lambda_{\Delta,r}}$ *and for the best of* n *policy* $\pi^{(n)}_{r,\text{ref}}$*:*

1. *For the optimal RL policy* $\pi_{\lambda_{\Delta,r}}$ *we have:*

$$0 \leq \mathbb{E}_{\pi_{\lambda_{\Delta,r}}} r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{2\sigma^2_{\text{ref}} \mathsf{KL}(\pi_{\lambda_{\Delta,r}}||\pi_{\text{ref}})} \leq \sqrt{2\sigma^2_{\text{ref}} \Delta}.$$

2. *For the Best of* n *policy* $\pi^{(n)}_{r,\text{ref}}$*, under Assumption 2 we have:*

$$0 \leq \mathbb{E}_{\pi^{(n)}_{r,\text{ref}}} r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{2\sigma^2_{\text{ref}} \mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{\text{ref}})} \leq \sqrt{2\sigma^2_{\text{ref}} \left(\log n - \frac{n-1}{n}\right)}.$$

A similar statement holds under sub-gamma tails of the reward of the reference model. We turn now to providing a bound in high probability on the empirical reward improvement of RL:

Remark 1. *Item (1) in Corollary 1 shows that the* $\sqrt{\sigma^2_{\text{ref}} \mathsf{KL}}$ *provides an upper bound on the reward improvement of the alignment under subgaussian tails of the reference reward. Under subgaussian tails of the reference, this information theoretic barrier can not be broken with a better algorithm. On way to improve on the* $\sqrt{\mathsf{KL}}$ *ceiling is by aiming at having a reference model with a reward that has subgamma tails to improve the upper limit to* $\sqrt{\sigma^2_{\text{ref}} \mathsf{KL}} + c\mathsf{KL}$*, or to subexponential tails to be linear in the* KL*. Item (2) can be seen as a refinement on the classical* $\sqrt{2\sigma^2_{\text{ref}} \log(n)}$ *upper bound on the expectation of maximum of subgaussians see for e.g Corollary 2.6 in [Boucheron et al., 2013]. If in addition* r *is positive and for* $X = r_\sharp \pi_{\text{ref}} - \mathbb{E}_{\pi_{\text{ref}}} r$ *we have for* $t > 0$ *,* $\mathbb{P}(X > t) \geq \mathbb{P}(|g| > t)$*, where* $g \sim \mathcal{N}(0, \sigma^2_\ell)$ *(where* σ^2_ℓ *is a variance) , then we have a matching lower bound for* $\pi^{(n)}_{r,\text{ref}}$ *that scales with* $\sqrt{\sigma^2_\ell \log(n)}$ *for sufficiently large* n *(See [Kamath, 2015]).*

The following Theorem gives high probability bounds for the excess reward when estimated from empirical samples:

Theorem 2 (High Probability Empirical Reward Improvement For RL)**.** *Assume* $r_\sharp \pi_{\text{ref}} \in \mathsf{SubGauss}(\sigma^2_{\text{ref}})$*. Let* $\beta > 1$ *and* $t_0 > 0$*. Let* $\pi_{\beta,r}$ *be the optimal policy of the penalized RL problem given in Equation (3). Let* $R_{i,\beta}$ *and* $R_{i,\text{ref}}, i = 1 \dots m$ *be the rewards evaluated at* m *samples from* $\pi_{\beta,r}$ *and* π_{ref}*. Assume that the* β*-Rényi divergence* $D_\beta(\pi_{\beta,r}||\pi_{\text{ref}})$ *and* $\mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})$ *are both finite. The following inequality holds with probability at least* $1 - e^{-\frac{mt_0^2}{2\sigma^2_{\text{ref}}}} - e^{-m(\beta-1)t_0}$*:*

$$\frac{1}{m}\sum_{i=1}^m R_{i,\beta} - \frac{1}{m}\sum_{i=1}^m R_{i,\text{ref}} \leq \sqrt{2\sigma^2_{\text{ref}} \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})} + \frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta} + 2t_0.$$

Note that in Theorem 2, we did not make any assumptions on the tails of $r_\sharp \pi_{\beta,r}$ and we see that this results in a biased concentration inequality with a non-negative bias $\frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta} \geq 0$. For the best of n policy, if the reward was positive and has a folded normal distribution (absolute value of gaussians), [Boucheron and Thomas, 2012] provides concentration bounds, owing to the subgamma tails of the maximum of absolute value of Gaussians.

3.2 Tail Adaptive Transportation Inequalities with the Rényi Divergence

An important question on the tightness of the bounds rises from the bounds in Corollary 1. We answer this question by considering additional information on the tails of the reward under the policy π, and we obtain tail adaptive bounds that are eventually tighter than the one in Corollary 1. Our new bounds leverage a variational representation of the Rényi divergence that uses the logarithmic moment generating function of both measures at hand.

Preliminaries for the Rényi Divergence The Donsker-Varadahn representation of KL was crucial in deriving transportation inequalities. In Shayevitz [2011] the following variational form is given for the Rényi divergence in terms of the KL divergence, for all $\alpha \in \mathbb{R}$

$$(1-\alpha)D_\alpha(P||Q) = \inf_R \alpha \mathsf{KL}(R||P) + (1-\alpha)\mathsf{KL}(R||Q) \tag{18}$$

A similar variational form was rediscovered in [Anantharam, 2018]. Finally a Donsker-Varadahn-Rényi representation of D_α was given in [Birrell et al., 2021]. For all $\alpha \in \mathbb{R}^+, \alpha \neq 0, 1$ we have :

$$\frac{1}{\alpha} D_\alpha(P||Q) = \sup_{h \in \mathcal{H}} \frac{1}{\alpha - 1} \log \left(\mathbb{E}_P e^{(\alpha-1)h} \right) - \frac{1}{\alpha} \log \left(\mathbb{E}_Q e^{\alpha h} \right), \tag{19}$$

where $\mathcal{H} = \left\{ h \middle| \int e^{(\alpha-1)h} dP < \infty, \int e^{\alpha h} dQ < \infty \right\}$. Birrell et al. [2021] presents a direct proof of this formulation without exploring its link to the representation given in (18), we show in what follows an elementary proof via convex conjugacy, the duality relationship between equations (18) and (19).

Theorem 3. *For $0 < \alpha < 1$ Equations (18) and (19) are dual of one another. For $\alpha > 1$ they are Toland Dual.*

We collect in what follows elementary lemmas that will be instrumental to derive transportation inequalities in terms of the Rényi divergence. Proofs are given in the Appendix.

Lemma 2. *Let $\alpha \in (0,1) \cup (1, \infty)$, and define $\mathcal{H} = \{h | e^{(\alpha-1)(h - \int h dP)} \in L^1(P), e^{(\alpha)(h - \int h dQ)} \in L^1(Q)\}$. We have for all $h \in \mathcal{H}$ and for $\alpha \in (0,1) \cup (1, \infty)$*

$$\int h dP - \int h dQ \leq \frac{1}{\alpha} D_\alpha(P||Q) - \frac{1}{\alpha - 1} \log \left(\int e^{(\alpha-1)(h - \int h dP)} dP \right) + \frac{1}{\alpha} \log \left(\int e^{\alpha(h - \int h dQ)} dQ \right)$$

Lemma 3. *The following limit holds for the Rényi divergence* $\lim_{\alpha \to 0} \frac{1}{\alpha} D_\alpha(P||Q) = \mathsf{KL}(Q||P)$.

Transportation Inequalities with Rényi Divergence. The following theorem shows that when considering the tails of π we can obtain tighter upper bounds using the Rényi divergence that is more tail adaptive:

Theorem 4 (Tail Adaptive Transportation Inequalities)**.** *Let $\alpha \in (0,1)$. Assume $r_\sharp \pi \in \mathsf{SubGauss}(\sigma_\pi^2)$ and $r_\sharp \pi_{\text{ref}} \in \mathsf{SubGauss}(\sigma_{\text{ref}}^2)$ then we have for all $\alpha \in (0,1)$:*

$$\mathbb{E}_\pi r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{2((1-\alpha)\sigma_\pi^2 + \alpha \sigma_{\text{ref}}^2) \frac{D_\alpha(\pi||\pi_{\text{ref}})}{\alpha}}. \tag{20}$$

In particular if there exits $\alpha \in (0,1)$ such that $D_\alpha(\pi||\pi_{\text{ref}}) \leq \frac{\alpha \sigma_{\text{ref}}^2}{(1-\alpha)\sigma_\pi^2 + \alpha \sigma_{\text{ref}}^2} \mathsf{KL}(\pi||\pi_{\text{ref}})$, then the tail adaptive upper bound given in Equation (20) is tighter than the one provided by the tails of π_{ref} only i.e $\sqrt{\sigma_{\text{ref}}^2 \mathsf{KL}(\pi||\pi_{\text{ref}})}$. Note that this is possible because D_α is increasing in $\alpha \in (0,1)$ [van Erven and Harremos, 2014], i.e $D_\alpha(\pi||\pi_{\text{ref}}) \leq \mathsf{KL}(\pi||\pi_{\text{ref}})$, and $\frac{\alpha \sigma_{\text{ref}}^2}{(1-\alpha)\sigma_\pi^2 + \alpha \sigma_{\text{ref}}^2} \leq 1$. Note that taking limits $\alpha \to 0$ (applying Lemma 3) and $\alpha \to 1$, and taking the minimum of the upper bounds we obtain:

$$\mathbb{E}_\pi r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{2 \min(\sigma_{\pi_{\text{ref}}}^2 \mathsf{KL}(\pi||\pi_{\text{ref}}), \sigma_\pi^2 \mathsf{KL}(\pi_{\text{ref}}||\pi))},$$

this inequality can be also obtained by applying Proposition 3 twice: on the tails of π and π_{ref} respectively.

Another important implication of Theorem 4, other than tighter than KL upper bound, is that if we were to change the RL alignment problem (1) to be constrained by $D_\alpha, \alpha \in (0,1)$ instead of KL, we may end up with a smaller upper limit on the reward improvement. This D_α constrained alignment may lead to a policy that under-performs when compared to a policy obtained with the KL constraint. This was indeed observed experimentally in [Wang et al., 2024] that used constraints with α- divergences for $\alpha \in (0,1)$ (that are related to Rényi divergences) and noticed a degradation in the reward improvement w.r.t the policy obtained using KL constraints.

4 Transportation Inequality Transfer From Proxy to Golden Reward

As we saw in the previous sections, the tightness of $\sqrt{\mathsf{KL}(\pi||\pi_{\text{ref}})}$ upper bound in alignment can be due to the tails of the reward of the aligned policy π (Theorem 4) and to the concentration around the mean in finite sample size (Theorem 2). Another important consideration is the mismatch between the golden reward r^* that one desires to maximize that is expensive and difficult to obtain (for example human evaluation) and a proxy reward r that approximates r^*. The proxy reward r is used instead of r^* in RL and in best of n policy. While we may know the tails of the reward r of the reference and aligned model, we don't have access to this information on the golden reward r^*. We show in this section how to transfer transportation inequalities from r to r^* for RL and Best of n policy.

Proposition 4 (r^* Transportation Inequality for RL Policy). *The following inequality holds:*

$$\mathbb{E}_{\pi_{\beta,r}} r^* - \mathbb{E}_{\pi_{\text{ref}}} r^* \leq \mathbb{E}_{\pi_{\beta,r}} r - \mathbb{E}_{\pi_{\text{ref}}} r - \frac{1}{\beta} \log \left(\int e^{\beta(r-r^*-(\int r d\pi_{\text{ref}} - \int r^* d\pi_{\text{ref}})} d\pi_{\beta,r^*} \right),$$

Assume $r_\sharp \pi_{\text{ref}} \in \mathsf{SubGauss}(\sigma^2_{\text{ref}})$, *and there exists* $\delta > 0$ *such that.* $\frac{1}{\beta} \log \left(\int e^{\beta(r-r^*-(\int r d\pi_{\text{ref}} - \int r^* d\pi_{\text{ref}})} d\pi_{\beta,r^*} \right) \geq \delta \mathsf{KL}(\pi_{\beta,r^*}||\pi_{\text{ref}})$ *then we have:*

$$\mathbb{E}_{\pi_{\beta,r}} r^* - \mathbb{E}_{\pi_{\text{ref}}} r^* \leq \sqrt{2\sigma^2_{\text{ref}} \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})} - \delta \mathsf{KL}(\pi_{\beta,r^*}||\pi_{\text{ref}}).$$

Note that $\frac{1}{\beta} \log \left(\int e^{\beta(r-r^*-(\int r d\pi_{\text{ref}} - \int r^* d\pi_{\text{ref}})} d\pi_{\beta,r^*} \right)$ is interpreted here as an interpolation between the mean and the maximum of its argument on the support of π_{β,r^*} (Proposition 9 in [Feydy et al., 2018]). Indeed as $\beta \to 0$, this boils down to the mean on $\int (r - r^*) d\pi_{\beta,r^*} - \left(\int r d\pi_{\text{ref}} - \int r^* d\pi_{\text{ref}} \right)$ and $\beta \to \infty$ this boils down to $\max_{\mathsf{supp}\pi_{\beta,r^*}} \{r - r^* - \left(\int r d\pi_{\text{ref}} - \int r^* d\pi_{\text{ref}} \right)\}$. Our assumption means that r overestimates r^* and the overestimation is accentuated as we drift from π_{ref} on which r was learned. This assumption echoes findings in [Gao et al., 2023] that show that the transportation inequalities suffer from overestimation of proxy reward models of the golden reward (See Figure 8 in [Gao et al., 2023]).

Note that in Proposition 4, we are evaluating the golden reward r^* improvement when using the proxy reward optimal policy $\pi_{\beta,r}$. We see that the golden reward of the RL policy inherits the transportation inequality from the proxy one but the improvement of the reward is hindered by possible overestimation of the golden reward by the proxy model. This explains the dip in performance as measured by the golden reward depicted in Figure 1 and reported in [Gao et al., 2023].

Proposition 5 (r^* Transportation Inequality for Best of n Policy). *Let $\varepsilon > 0$. Let r be a surrogate reward such that $\|r - r^*\|_\infty \leq \varepsilon$ and assume $r_\sharp \pi_{\text{ref}} \in \mathsf{SubGauss}(\sigma^2_{\text{ref}})$ then the best of n policy $\pi^{(n)}_{r,\text{ref}}$ satisfies:*

$$\mathbb{E}_{\pi^{(n)}_{r,\text{ref}}}(r^*) - \mathbb{E}_{\pi_{\text{ref}}}(r^*) \leq \sqrt{2\sigma^2_{\text{ref}} \left(\log(n) - \frac{n-1}{n} \right)} + 2\varepsilon \left(\left(\frac{1}{n} \right)^{\frac{1}{n-1}} - \left(\frac{1}{n} \right)^{\frac{n}{n-1}} \right).$$

Transportation inequalities transfers for the best of n policy from r to r^* and pays only an additional error term $\|r - r^*\|_\infty \mathsf{TV}(\pi^{(n)}_{r,\text{ref}}|\pi_{\text{ref}})$, an upper bound of this total variation as a function of n is given in Table 1. As mentioned in remark 1, if we have lower bounds on the tail of the reference reward, then we also have a lower bound on the reward improvement that scales like $C\sqrt{\sigma^2_\ell \log(n)} - 2\varepsilon \left(\left(\frac{1}{n} \right)^{\frac{1}{n-1}} - \left(\frac{1}{n} \right)^{\frac{n}{n-1}} \right)$. This is in line with empirical findings in [Hilton and Gao, 2022] [Gao et al., 2023] that showed that best of n policy is resilient as the reward model r gets closer to r^*.

5 Conclusion

We presented in this paper a comprehensive information theoretical analysis of policy alignment using reward optimization with RL and best of n sampling. We showed for best of n a bound on KL under realistic assumptions on the reward. Our analysis showed that the alignment reward improvement, is intrinsically constrained by the tails of the reward under the reference policy and controlling the KL divergence results in an upper bound of the policy improvement. We showed that the KL bound may not be tight if the tails of the optimized policy satisfy a condition expressed via Rényi divergence. We also explained the deterioration of the golden reward via overestimation of the proxy reward.

References

V. Anantharam. A variational characterization of rényi divergences. *IEEE Transactions on Information Theory*, 64(11):6979–6989, 2018.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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, 2024.

J. Birrell, P. Dupuis, M. A. Katsoulakis, L. Rey-Bellet, and J. Wang. Variational representations and neural network estimation of rényi divergences. *SIAM Journal on Mathematics of Data Science*, 3 (4):1093–1116, 2021.

S. Boucheron and M. Thomas. Concentration inequalities for order statistics. 2012.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/ACPROF:OSO/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

T. Coste, U. Anwar, R. Kirk, and D. Krueger. Reward model ensembles help mitigate overoptimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dcjtMYkpXx.

K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

J. Feydy, T. Séjourné, F.-X. Vialard, S. ichi Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences, 2018.

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

D. Go, T. Korbak, G. Kruszewski, J. Rozen, and M. Dymetman. Compositional preference models for aligning LMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tiiAzqi6Ol.

J. Hilton and L. Gao. Measuring goodhart's law, 2022.

G. Kamath. Bounds on the expectation of the maximum of samples from a gaussian. *URL http://www.gautamkamath. com/writings/gaussian max. pdf*, 10:20–30, 2015.

N. Lambert, V. Pyatkin, J. Morrison, L. Miranda, B. Y. Lin, K. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, N. A. Smith, and H. Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024.

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. *arXiv preprint arXiv:2310.17022*, 2023.

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.

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*, 35:27730–27744, 2022.

Y. Polyanskiy and Y. Wu. Information theory: From coding to learning, 2023.

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*, 36, 2024.

A. Rényi. On the theory of order statistics. *Acta Mathematica Academiae Scientiarum Hungarica*, 4: 191–231, 1953. URL https://api.semanticscholar.org/CorpusID:123132570.

F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser, Cham, 2015. ISBN 9783319208275. doi: 10.1007/978-3-319-20828-2.

O. Shayevitz. On rényi measures and hypothesis testing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 894–898, 2011. doi: 10.1109/ISIT.2011.6034266.

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*, 33:3008–3021, 2020.

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.

T. van Erven and P. Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500.

C. Wang, Y. Jiang, C. Yang, H. Liu, and Y. Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2cRzmWXK9N.

J. Q. Yang, S. Salamatian, Z. Sun, A. T. Suresh, and A. Beirami. Asymptotics of language model alignment, 2024.

K. Yang and D. Klein. FUDGE: Controlled text generation with future discriminators. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.276. URL https://aclanthology.org/2021.naacl-main.276.

Y. Zhao, M. Khalman, R. Joshi, S. Narayan, M. Saleh, and P. J. Liu. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0qSOodKmJaN.

A Broader Impact and Limitations

We believe this work explaining scaling laws for reward models and alignment will give practitioners insights regarding the limits of what is attainable via alignment. All assumptions under which our statements hold are given. We don't see any negative societal impact of our work.

B Proofs For Best of n Policy

B.1 Best of n Policy KL Guarantees

Proof of Lemma 1.

$$\mathsf{KL}(E^{(n)}||E) = \int_0^{+\infty} f_{E^{(n)}}(x) \log\left(\frac{f_{E^{(n)}}(x)}{f_E(x)}\right) dx$$

We have $f_E(x) = e^{-x}1_{x\geq 0}$. Note that the CDF of maximum of exponential $F_{E^{(n)}}(x) = (1-e^{-x})1_{x\geq 0}$, and hence $f_{E^{(n)}}(x) = n(1-e^{-x})^{n-1}e^{-x}1_{x\geq 0}$. Hence we have:

$$\begin{aligned}\mathsf{KL}(E^{(n)}||E) &= \int_0^{+\infty} n(1-e^{-x})^{n-1}e^{-x}\log\left(\frac{n(1-e^{-x})^{n-1}e^{-x}}{e^{-x}}\right)dx \\ &= \int_0^{+\infty} n(1-e^{-x})^{n-1}e^{-x}\log\left(n(1-e^{-x})^{n-1}\right)dx\end{aligned}$$

Let $u = 1-e^{-x}$, we have $du = e^{-x}dx$. It follows that :

$$\begin{aligned}\mathsf{KL}(E^{(n)}||E) &= \int_0^1 nu^{n-1}\log\left(nu^{n-1}\right)du \\ &= \int_0^1 nu^{n-1}\left(\log(n)+(n-1)\log(u)\right)du \\ &= \log(n)\int_0^1 du^n + (n-1)\int_0^1 nu^{n-1}\log(u)du \\ &= \log(n) + (n-1)\int_0^1 d(u^n\log u - \frac{u^n}{n}) \\ &= \log(n) - \frac{n-1}{n}.\end{aligned}$$

□

B.2 Best of n Policy f divergence and Rényi Divergence

Best of n Policy f divergence and Renyi divergence Guarantees Given that our proof technique relies on DPI and Rényi representation, we show that similar results hold for any f-divergence and for the Rényi divergence:

$$D_f(P||Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right)dx, \tag{21}$$

where f is convex and $f(1) = 0$. Hence we have by DPI for f-divergences:

Theorem 5. *Under Assumption 2 the best of n policy satisfies for any f-divergence:*

$$\mathrm{D}_f(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\mathrm{ref}}) \leq \int_0^1 f\left(nu^{n-1}\right)du \tag{22}$$

Proof of Theorem 5.

$$
\begin{aligned}
\mathrm{D}_f(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\mathrm{ref}}|X) &= \mathrm{D}_f(Y^{(n)}||Y|X) \\
&= \mathrm{D}_f(H_X(R^n(Y))||H_X(R(Y))|X) \\
&\leq \mathrm{D}_f(R^n(Y)||R(Y)|X) \text{ By the data processing inequality} \quad (23) \\
&= \mathrm{D}_f(T_X(E^{(n)})||T_X(E)) \text{ Renyi and Optimal Transport Representations (8)} \\
&= \mathrm{D}_f(E^{(n)}||E) \text{ since } T_X \text{ is a monotonic bijection DPI is an equality} \quad (24) \\
&= \int_0^{+\infty} f_E(x) f\left(\frac{f_{E^{(n)}}(x)}{f_E(x)}\right) dx \quad (25) \\
&= \int_0^{\infty} (e^{-x}) f\left(n(1-e^{-x})^{n-1}\right) du \quad (26) \\
&= \int_0^1 f(nu^{n-1}) du. \quad (27)
\end{aligned}
$$

In particular we have the following bounds for common f divergences:

- For $f(x) = x\log(x)$ we obtain the KL divergence and we have the result:
$$\int_0^1 nu^{n-1}\log(nu^{n-1})du = \mathsf{KL}(E^{(n)}||E) = \log(n) - \frac{n-1}{n}.$$
- For $f(x) = (x-1)^2$ we obtain the chi-squared divergence and we have: $\int_0^1 \left(nu^{n-1}-1\right)^2 du = \int_0^1 (n^2u^{2(n-1)} - 2nu^{n-1} + 1)du = \frac{n^2}{2n-1}u^{2n-1} - 2u^n + u|_0^1 = \frac{n^2}{2n-1} - 2 + 1 = \frac{n^2-2n+1}{2n-1} = \frac{(n-1)^2}{2n-1}$.
- For $f(x) = \frac{1}{2}|x-1|$, we obtain the total variation distance (TV) and we have: $\frac{1}{2}\int_0^1 \left|nu^{n-1}-1\right| du = \frac{1}{2}(\int_0^{u^*}\left(1-nu^{n-1}\right) du + (\int_{u^*}^1 \left(nu^{n-1}-1\right) du) = (u^* - (u^*)^n)$,where $n(u^*)^{(n-1)} = 1$, i.e $u^* = (\frac{1}{n})^{\frac{1}{n-1}}$. Hence the TV is $(\frac{1}{n})^{\frac{1}{n-1}} - (\frac{1}{n})^{\frac{n}{n-1}}$.
- For $f(x) = (1-\sqrt{x})^2$ we have the hellinger distance: $\int_0^1 \left(\sqrt{n}u^{\frac{n-1}{2}} - 1\right)^2 du = \int_0^1 (nu^{n-1} - 2\sqrt{n}u^{\frac{n-1}{2}} + 1)du = u^n - 2\sqrt{n}\frac{u^{\frac{n+1}{2}}}{\frac{n+1}{2}} + u\Big|_0^1 = 2(1-\frac{2\sqrt{n}}{n+1}) = 2\frac{(1-\sqrt{n})^2}{n+1}$
- For $f(x) = -\log(x)$, we obtain the forward KL and we have : $\int_0^1 f(nu^{n-1})du = n - 1 - \log(n)$.

□

Guarantees with Rényi Divergence Turning now to the Rényi divergence for $\alpha \in (0,1)\cup(1,\infty)$:

$$D_\alpha(P||Q) = \frac{1}{(\alpha-1)}\log\left(\int p^\alpha(x)q^{1-\alpha}(x)dx\right)$$

the limit as $\alpha \to 1$ $D_1(P||Q)) = \mathsf{KL}(P||Q)$.

Theorem 6. *Under Assumption 2 the best of n policy satisfies:*

$$D_\alpha(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\mathrm{ref}}) \leq \frac{1}{(\alpha-1)}\log\left(\frac{n^\alpha}{\alpha(n-1)+1}\right) \quad (28)$$

Proof of Theorem 6. Applying DPI that holds also for the Rényi divergence twice from $Y, Y^{(n)}$ to $R, R^{(n)}$ and from $R, R^{(n)}$ to $E, E^{(n)}$ we obtain :

$$D_\alpha(\pi^{(n)}_{r,\text{ref}}||\pi_{\text{ref}}|X) \leq D_\alpha(E^{(n)}||E)$$

$$\begin{aligned} D_\alpha(E^{(n)}||E) &= \frac{1}{(\alpha-1)} \log\left(\int_0^\infty n^\alpha (1-e^{-x})^{\alpha(n-1)} e^{-\alpha x} e^{-x(1-\alpha)} dx\right) \\ &= \frac{1}{(\alpha-1)} \log\left(\int_0^{+\infty} n^\alpha (1-e^{-x})^{\alpha(n-1)} e^{-x} dx\right) \end{aligned}$$

Let $u = 1 - e^{-x}$ we have $du = e^{-x}dx$

$$\begin{aligned} D_\alpha(E^{(n)}||E) &= \frac{1}{(\alpha-1)} \log\left(\int_0^1 n^\alpha u^{\alpha(n-1)} du\right) \\ &= \frac{1}{(\alpha-1)} \left(\log n^\alpha + \log \int_0^1 u^{\alpha(n-1)} du\right) \\ &= \frac{1}{(\alpha-1)} \left(\log n^\alpha + \log \frac{u^{\alpha(n-1)+1}}{\alpha(n-1)+1}\Big|_0^1\right) \\ &= \frac{1}{(\alpha-1)} \log\left(\frac{n^\alpha}{\alpha(n-1)+1}\right) \end{aligned}$$

□

From Renyi to KL guarantees Let $s_1(\alpha) = (\alpha - 1)$, and $s_2(\alpha) = \log\left(\frac{n^\alpha}{\alpha(n-1)+1}\right)$, we have $D_\alpha(E^{(n)}||E) = \frac{s_2(\alpha)}{s_1(\alpha)}$, we have $\mathsf{KL}(E^{(n)}||E) = \lim_{\alpha\to 1} D_\alpha(E^{(n)}||E) = \lim_{\alpha\to 1} \frac{s_2(\alpha)}{s_\alpha} = \frac{0}{0}$, hence applying L'Hôpital rule we have: $\lim_{\alpha\to 1} \frac{s_2(\alpha)}{s_1(\alpha)} = \lim_{\alpha\to 1} \frac{s_2'(\alpha)}{s_1'(\alpha)} = \lim_{\alpha\to 1} \frac{\log(n) - \frac{n-1}{\alpha(n-1)+1}}{1} = \log(n) - \frac{n-1}{n}$. Hence we recover the result for the KL divergence.

B.3 Best of n Dominance

Proof of Proposition 1 . $F_{E^{(n)}}(x) = (F_E(x))^n \leq F_E(x), \forall x \geq 0$, which means also that $F^{-1}_{E^{(n)}}(t) \geq F^{-1}_E(t), \forall t \in [0,1]$, which means that $E^{(n)}$ dominates E in the first stochastic order : $E^{(n)} \underset{\text{FSD}}{\succcurlyeq} E$, which means there exists a coupling between $E^{(n)}$ and E, $\pi \in \Pi(E^{(n)}, E)$, such that $E \geq e$, for all $(E, e) \sim \pi$. On the other hand By Rényi and Monge map representations we have: $R^{(n)} = F^{-1}_R \circ F_E(E^{(n)})$ and $R = F^{-1}_R \circ F_E(E)$, given that $T = F^{-1}_R \circ F_E$ is non decreasing the same coupling π guarantees that $T(E) \geq T(e)$, for all $(E, e) \sim \pi$ and Hence $R^{(n)} \underset{\text{FSD}}{\succcurlyeq} R$.

□

Corollary 2. *Best of n-polciy has higher expectation :*

$$\mathbb{E}R^{(n)} \geq \mathbb{E}R,$$

and is a safer policy, let the Tail Value at Risk be:

$$\text{TVAR}_p(X) = \frac{1}{p}\int_0^p Q_R(t)dt$$

We have

$$\text{TVAR}_p(R^n) \geq \text{TVAR}_p(R), \forall p \in [0,1]$$

Proof of Corollary 2. First order dominance implies second order dominance (i.e by integrating quantiles). Expectation is obtained for $p = 1$. □

C Best of n and RL Policy

Proof of Proposition 2. We fix here $\beta = \frac{1}{\lambda_\Delta}$

$$\begin{aligned}\mathsf{KL}(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\beta,r}) &= \int \pi^{(n)}_{r,\mathrm{ref}}(y|x)\log\left(\frac{\pi^{(n)}_{r,\mathrm{ref}}(y|x)}{\pi_{\beta,r}(y|x)}\right) = \int \pi^{(n)}_{r,\mathrm{ref}}(y|x)\log\left(\frac{\pi^{(n)}_{r,\mathrm{ref}}(y|x)}{\pi_{\mathrm{ref}}(y|x)\frac{e^{\beta r(x,y)}}{Z_\beta(x)}}\right)\\ &= \mathsf{KL}(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\mathrm{ref}}) + \log\left(\mathbb{E}_{\pi_{\mathrm{ref}}}e^{\beta r}\right) - \beta\int r d\pi^{(n)}_{r,\mathrm{ref}}\end{aligned}$$

On the other hand by optimality of $\pi_{\beta,r}$ we have:

$$\mathsf{KL}\left(\pi_{\beta,r}||\pi_{\mathrm{ref}}\right) = \beta\int r d\pi_{\beta,r} - \log\left(\int e^{\beta r}d\pi_{\mathrm{ref}}\right)$$

and hence we have:

$$\mathsf{KL}(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\beta,r}) = \mathsf{KL}(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\mathrm{ref}}) - \mathsf{KL}\left(\pi_{\beta,r}||\pi_{\mathrm{ref}}\right) + \beta\left(\int r d\pi_{\beta,r} - \int r d\pi^{(n)}_{r,\mathrm{ref}}\right)$$

We choose n such that :

$$\mathsf{KL}(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\mathrm{ref}}) \leq \log(n) - \frac{n-1}{n} \leq \mathsf{KL}\left(\pi_{\beta,r}||\pi_{\mathrm{ref}}\right) = \Delta$$

and we conclude choosing $n = e^\Delta$ therefore for that choice of n that:

$$\mathsf{KL}(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\beta,r}) \leq \beta\left(\int r d\pi_{\beta,r} - \int r d\pi^{(n)}_{r,\mathrm{ref}}\right)$$

On the other hand we have:

$$\begin{aligned}\left|\int r d\pi_{\beta,r} - \int r d\pi^{(n)}_{r,\mathrm{ref}}\right| &= \left|\int r\exp(\beta r)\frac{1}{Z_\beta}d\pi_{\mathrm{ref}} - \int \max_i r(x_i) d\pi_{\mathrm{ref}}(x_1)\dots d\pi_{\mathrm{ref}}(x_n)\right|\\ &= \left|\int\left(\frac{1}{n}\sum_{i=1}^n \frac{r(x_i)\exp(\beta r(x_i))}{Z_\beta} - \max_i r(x_i)\right)d\pi_{\mathrm{ref}}(x_1)\dots d\pi_{\mathrm{ref}}(x_n)\right|\\ &= \left|\int\left(\frac{1}{n}\sum_{i=1}^n \frac{r(x_i)\exp(\beta r(x_i))}{\sum_{i=1}^n\exp(\beta r(x_i))}\frac{\sum_{i=1}^n\exp(\beta r(x_i))}{Z_\beta} - \max_i r(x_i)\right)d\pi_{\mathrm{ref}}(x_1)\dots d\pi_{\mathrm{ref}}(x_n)\right|\\ &\leq \int\left|\max r(x_i)\left(\frac{\frac{1}{n}\sum_{i=1}^n\exp(\beta r(x_i))}{Z_\beta} - 1\right)\right| d\pi_{\mathrm{ref}}(x_1)\dots d\pi_{\mathrm{ref}}(x_n)\\ &\leq \frac{M}{Z_\beta}\mathbb{E}\left|\sum_{i=1}^n\exp(\beta r(x_i)) - Z_\beta\right|\end{aligned}$$

where we used the following fact, followed by Jensen inequality :

$$\sum_{i=1}^n \frac{r(x_i)\exp(\beta r(x_i))}{\sum_{i=1}^n\exp(\beta r(x_i))} \leq \max_i r(x_i).$$

Assume that the reward is bounded hence we have by Hoeffding inequality :

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\exp(\beta r(x_i)) - Z_\beta\right| \geq t\right) \leq 2e^{-\frac{nt^2}{2(\exp(\beta M)-\exp(-\beta M))^2}}$$

Hence we have:

$$\mathbb{E}\left|\sum_{i=1}^n\exp(\beta r(x_i)) - Z_\beta\right| \leq 2\sqrt{\frac{\pi}{2}}\frac{\exp(\beta M)-\exp(-\beta M)}{\sqrt{n}}$$

$$\mathsf{KL}(\pi^{(\exp(\Delta))}_{r,\mathrm{ref}}||\pi_{\lambda_\Delta,r}) \leq \frac{M}{\lambda_\Delta Z_{1/\lambda_\Delta}}\sqrt{2\pi}(\exp(\beta M)-\exp(-\beta M))\sqrt{\exp(-\Delta)}.$$

□

D Transportation Inequalities and KL Divergence

D.1 Transportation Inequalities with KL

The following Lemma (Lemma 4.14 in [Boucheron et al., 2013]) uses the Donsker-Varadhan representation of the KL divergence to obtain bounds on the change of measure , and using the tails of π_{ref}.

Lemma 4 (Lemma 4.14 in [Boucheron et al., 2013]). *Let ψ be a convex and continuously differentiable function ψ on a possibly unbounded interval $[0, b)$, and assume $\psi(0) = \psi'(0) = 0$. Define for every $x \geq 0$, the convex conjugate $\psi^*(x) = \sup_{\lambda \in [0,b)} \lambda x - \psi(\lambda)$, and let $\psi^{*-1}(t) = \inf\{x \geq 0 : \psi^*(x) > t\}$. Then the following statements are equivalent:*
(i) For $\lambda \in [0, b)$

$$\log\left(\int e^{\lambda(r - \int r dQ)} dQ\right) \leq \psi(\lambda),$$

(ii) For any probability measure P that is absolutely continuous with respect to Q and such that $\mathsf{KL}(P||Q) < \infty$:

$$\int r dP - \int r dQ \leq \psi^{*-1}(\mathsf{KL}(P||Q)).$$

Lemma 5 (Inverse of the conjugate [Boucheron et al., 2013]).

1. *If $Q \in \mathsf{SubGauss}(\sigma^2)$, we have for $t \geq 0$ $\psi^{*-1}(t) = \sqrt{2\sigma^2 t}$.*
2. *If $Q \in \mathsf{Subgamma}(\sigma^2, c)$, we have for $t \geq 0$ $\psi^{*-1}(t) = \sqrt{2\sigma^2 t} + ct$.*

We give here a direct proof for the subgaussian case:

Proof. By the Donsker Varadhan representation of the KL we have:

$$\mathsf{KL}(P||Q) = \sup_h \int h dP - \log\left(\int e^h dQ\right)$$

Fix x and $M > 0$ and define for $0 < \lambda < M$

$$h_\lambda(y) = \lambda\left(r(x, y) - \mathbb{E}_{\pi_{\text{ref}}(y|x)} r(x, y)\right)$$

We omit in what follows x and y, but the reader can assume from here on that π and π_{ref} are conditioned on x. Note that $R_{\text{ref}}|x = (r(x, .))_\sharp \pi_{\text{ref}}(.|x)$ and we assume $R_{\text{ref}}|x$ subgaussian. Note that

$$\mathbb{E}_{\pi_{\text{ref}}} e^{h_\lambda} = \mathbb{E}_{\pi_{\text{ref}}|x} e^{\lambda(r - \mathbb{E}_{\pi_{\text{ref}}|x} r)} = M_{R_{\text{ref}}|x}(\lambda),$$

where $M_{R_{\text{ref}}|x}$ the moment generating function of the reward under the reference policy. $R_{\text{ref}}|x$ is subgaussian we have for all $\lambda \in \mathbb{R}$:

$$\mathbb{E}_{\pi_{\text{ref}}|x} e^{h_\lambda} \leq e^{\frac{\lambda^2 \sigma^2}{2}} \leq e^{\frac{M^2 \sigma^2}{2}} < \infty$$

Hence $h_\lambda \in \mathcal{H}$ and we have for all $\pi \ll \pi_{\text{ref}}$ and for all $0 < M < \infty$ and $0 < \lambda < M$:

$$\lambda \mathbb{E}_{\pi|x}(r - \mathbb{E}_{\pi_{\text{ref}}|x} r) \leq \mathsf{KL}(\pi||\pi_{\text{ref}}|x) + \log\left(\mathbb{E}_{\pi_{\text{ref}}|x} e^{\lambda(r - \mathbb{E}_{\pi_{\text{ref}}|x} r)}\right)$$

or equivalently:

$$\mathbb{E}_{\pi|x} r - \mathbb{E}_{\pi_{\text{ref}}|x} r \leq \frac{1}{\lambda}\mathsf{KL}(\pi||\pi_{\text{ref}}|x) + \frac{1}{\lambda}\log\left(\mathbb{E}_{\pi_{\text{ref}}|x} e^{\lambda(r - \mathbb{E}_{\pi_{\text{ref}}|x} r)}\right)$$

Finally we have for $\pi \ll \pi_{\text{ref}}$ for all $0 < \lambda < M$:

$$\mathbb{E}_{\pi|x} r - \mathbb{E}_{\pi_{\text{ref}}|x} r \leq \frac{1}{\lambda}\mathsf{KL}(\pi||\pi_{\text{ref}}|x) + \frac{1}{\lambda}\log\left(M_{R_{\text{ref}}|x}(\lambda)\right) \tag{29}$$

Being a subgaussian, the MGF of $R_{\text{ref}}|x$ is bounded as follows:

$$\log\left(M_{R_{\text{ref}}|x}(\lambda)\right) \leq \frac{\lambda^2 \sigma^2}{2}.$$

Hence we have for :

$$\mathbb{E}_{\pi|x} r - \mathbb{E}_{\pi_{\text{ref}}|x} r \leq \frac{1}{\lambda} \mathsf{KL}(\pi||\pi_{\text{ref}}|x) + \frac{\lambda\sigma^2}{2}$$

Integrating over x we obtain for all $\pi \ll \pi_{\text{ref}}$ and all $0 < \lambda < M$:

$$\mathbb{E}_{\pi} r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \frac{1}{\lambda} \mathsf{KL}(\pi||\pi_{\text{ref}}) + \frac{\lambda\sigma^2}{2}$$

Define :

$$\delta(\lambda) = \frac{1}{\lambda} \mathsf{KL}(\pi||\pi_{\text{ref}}) + \frac{\lambda\sigma^2}{2}$$

minimizing the upper bound $\delta(\lambda)$ for $\lambda \in (0, M]$, taking derivative $\delta'(\lambda) = -\frac{\mathsf{KL}(\pi||\pi_{\text{ref}})}{\lambda^2} + \frac{\sigma^2}{2} = 0$ gives $\lambda^* = \sqrt{\frac{2\mathsf{KL}(\pi||\pi_{\text{ref}})}{\sigma^2}}$. Taking $M = 2\lambda^*$, λ^* is the minimizer. Putting this in the bound we have finally for all rewards r for all π:

$$\mathbb{E}_{\pi} r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{2\sigma^2 \mathsf{KL}(\pi||\pi_{\text{ref}})}. \tag{30}$$

□

Proof of Corollary 1. (i) This follows from optimality of π_{λ_Δ} and applying the transportation inequality for gaussian tail.

(ii) This follows from applying Corollary 2 (best of n policy has larger mean) and 1 for bounding the KL.

□

Proof of Theorem 2. For the penalized RL we have by optimality:

$$\begin{aligned}\int r d\pi_{\beta,r} - \frac{1}{\beta} \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}}) &= \frac{1}{\beta} \log\left(\int e^{\beta r} d\pi_{\text{ref}}\right) \\ &= \frac{1}{\beta} \log\left(\int e^{\beta(r - \int r d\pi_{\text{ref}})} d\pi_{\text{ref}}\right) + \int r d\pi_{\text{ref}}\end{aligned}$$

It follows that :

$$\frac{1}{\beta} \log\left(\int e^{\beta(r - \int r d\pi_{\text{ref}})} d\pi_{\text{ref}}\right) = \int r d\pi_{\beta,r} - \int r d\pi_{\text{ref}} - \frac{1}{\beta} \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}}) \tag{31}$$

On the other hand by the variational representation of the Rényi divergence we have:

$$\begin{aligned}\int r d\pi_{\beta,r} - \int r d\pi_{\text{ref}} \leq \frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta} &- \frac{1}{\beta - 1} \log\left(\int e^{(\beta-1)(r - \int r d\pi_{\beta,r})} d\pi_{\beta,r}\right) \\ &+ \frac{1}{\beta} \log\left(\int e^{\beta(r - \int r d\pi_{\text{ref}})} d\pi_{\text{ref}}\right)\end{aligned} \tag{32}$$

Summing Equations (31) and (32) we obtain a bound on the moment generating function at β of $r_\sharp \pi_{\beta,r}$ (this is not a uniform bound , it holds only for β):

$$\frac{1}{\beta - 1} \log\left(\int e^{(\beta-1)(r - \int r d\pi_{\beta,r})} d\pi_{\beta,r}\right) \leq \frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta}. \tag{33}$$

Let us assume $\beta > 1$ we have therefore the following bound on the logarithmic moment generation function at $\beta - 1$

$$\psi_{r_\sharp \pi_{\beta,r}}(\beta - 1) \leq \frac{\beta - 1}{\beta} \left(D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})\right)$$

Let $R_{i,\beta} = r_\sharp \pi_{\beta,r}, i = 1 \dots m$, the reward evaluation of m independent samples of $\pi_{\beta,r}$ we have:

$$\begin{aligned}
\mathbb{P}\Big\{ \sum_{i=1}^{m} (R_{i,\beta} - \int r d\pi_{\beta,r}) > mt \Big\} &= \mathbb{P}(e^{\sum_{i=1}^{m}(\beta-1)(R_{i,\beta} - \int r d\pi_{\beta,r})} > e^{m(\beta-1)t}) \\
&\leq e^{-(\beta-1)mt} e^{m\psi_{R_\beta}(\beta-1)} \\
&\leq e^{-(\beta-1)mt} e^{m\frac{\beta-1}{\beta}(D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}}))} \\
&\leq e^{-m(\beta-1)\left(t - \frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta}\right)}
\end{aligned} \tag{34}$$

Let $t_0 > 0$, hence we have for $\beta > 1$:

$$\mathbb{P}\Big\{ \frac{1}{m}\sum_{i=1}^{m} R_{i,\beta} > \int r d\pi_{\beta,r} + t_0 + \frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta} \Big\} \leq e^{-m(\beta-1)t_0}$$

Now turning to $R_{\text{ref}} = r_\sharp \pi_{\text{ref}}$, since $R_{\text{ref}} \in \mathsf{SubGauss}(\sigma^2_{\text{ref}})$ we have for every $t_0 > 0$:

$$\mathbb{P}\Big\{ -\frac{1}{m}\sum_{i=1}^{m} R_{i,\text{ref}} > -\int r d\pi_{\text{ref}} + t_0 \Big\} \leq e^{-\frac{mt_0^2}{2\sigma^2_{\text{ref}}}}$$

Hence we have with probability at least $1 - e^{-\frac{mt_0^2}{2\sigma^2_{\text{ref}}}} - e^{-m(\beta-1)t_0}$:

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m} R_{i,\beta} - \frac{1}{m}\sum_{i=1}^{m} R_{i,\text{ref}} &\leq \int r d\pi_{\beta,r} - \int r d\pi_{\text{ref}} + 2t_0 + \frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta} \\
&\leq \sqrt{2\sigma^2_{\text{ref}} \mathsf{KL}(\pi||\pi_{\text{ref}})} + 2t_0 + \frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta}.
\end{aligned}$$

□

E Proofs for Transportation Inequalities and Rényi Divergence

Proposition 6 (Fenchel Conjugate Propreties)**.** *Let F and G be convex functions on a space E and F^*, G^* be their convex conjugates defined on E^*. We have:*

1. *Let $F_\gamma(x) = \gamma F(x)$ we have:*

$$F^*_\gamma(p) = \gamma F^*\left(\frac{p}{\gamma}\right) \tag{35}$$

2. *Duality:*

$$\min_{x \in E} F(x) + G(x) = \max_{p \in E^*} -F^*(-p) - G^*(p) \tag{36}$$

3. *Toland Duality:*

$$\min_{x \in E} F(x) - G(x) = \min_{p} G^*(p) - F^*(p) \tag{37}$$

Proof of Theorem 3. Let $\gamma > 0$, let $F_{P,\gamma}(R) = \gamma \mathsf{KL}(R||P)$, the Fenchel conjugate of $F_{P,1}(.)$ is defined for h bounded and measurable function as follows $F^*_{P,1}(h) = \log \mathbb{E}_P e^h$. It follows by 1) in Proposition 6 that : $F^*_{P,\gamma}(h) = \gamma F^*_{P,1}(\frac{h}{\gamma}) = \gamma \log \mathbb{E}_P e^{\frac{h}{\gamma}}$.
For $0 < \alpha < 1$: The objective function in (18) is the sum of convex functions: $F_{P,\alpha}(R) + F_{Q,1-\alpha}(R)$, by (2) in Proposition 6, we have by duality:

$$\begin{aligned}
(1-\alpha) D_\alpha(P||Q) &= \inf_R F_{P,\alpha}(R) + F_{Q,1-\alpha}(R) \\
&= \sup_{h \in \mathcal{H}} -F^*_{P,\alpha}(-h) - F^*_{Q,1-\alpha}(h) \\
&= \sup_{h \in \mathcal{H}} -\alpha \log \mathbb{E}_P e^{-\frac{h}{\alpha}} - (1-\alpha) \log \mathbb{E}_Q e^{\frac{h}{1-\alpha}}
\end{aligned}$$

Replacing h by $(1-\alpha)(\alpha)h$ does not change the value of the sup and hence we obtain:

$$\begin{aligned}(1-\alpha)D_\alpha(P||Q) &= \sup_{h\in\mathcal{H}} -\alpha\log\mathbb{E}_P e^{-\frac{(1-\alpha)(\alpha)h}{\alpha}} - (1-\alpha)\log\mathbb{E}_Q e^{\frac{(1-\alpha)(\alpha)h}{1-\alpha}} \\ &= \sup_{h\in\mathcal{H}} -\alpha\log\mathbb{E}_P e^{-(1-\alpha)h} - (1-\alpha)\log\mathbb{E}_Q e^{\alpha h}.\end{aligned}$$

dividing by $\frac{1}{\alpha(1-\alpha)}$ both sides we obtain for $0<\alpha<1$:

$$\frac{1}{\alpha}D_\alpha(P||Q) = \sup_{h\in\mathcal{H}} -\frac{1}{1-\alpha}\log\mathbb{E}_P e^{-(1-\alpha)h} - \frac{1}{\alpha}\log\mathbb{E}_Q e^{\alpha h}$$

For $\alpha>1$: The objective function in (18) is the difference of convex functions: $F_{P,\alpha}(R) - F_{Q,\alpha-1}(R)$, by Toland Duality (3) in Proposition 6 we have:

$$\begin{aligned}(1-\alpha)D_\alpha(P||Q) &= \inf_R F_{P,\alpha}(R) - F_{Q,\alpha-1}(R) \\ &= \inf_{h\in\mathcal{H}} F^*_{Q,\alpha-1}(h) - F^*_{P,\alpha}(h) \\ &= \inf_{h\in\mathcal{H}} (\alpha-1)\log\mathbb{E}_Q e^{\frac{h}{(\alpha-1)}} - \alpha\log\mathbb{E}_P e^{\frac{h}{\alpha}}\end{aligned}$$

The inf does not change when we replace h by $\alpha(\alpha-1)h$, hence we have:

$$\begin{aligned}(\alpha-1)D_\alpha(P||Q) &= -\inf_{h\in\mathcal{H}} (\alpha-1)\log\mathbb{E}_Q e^{\frac{\alpha(\alpha-1)h}{(\alpha-1)}} - \alpha\log\mathbb{E}_P e^{\frac{\alpha(\alpha-1)h}{\alpha}} \\ &= \sup_{h\in\mathcal{H}} \alpha\log\mathbb{E}_P e^{(\alpha-1)h} - (\alpha-1)\log\mathbb{E}_Q e^{\alpha h}\end{aligned}$$

dividing both sides by $\frac{1}{\alpha(\alpha-1)}$ we obtain for $\alpha>1$:

$$\frac{1}{\alpha}D_\alpha(P||Q) = \sup_{h\in\mathcal{H}} \frac{1}{\alpha-1}\log\mathbb{E}_P e^{(\alpha-1)h} - \frac{1}{\alpha}\log\mathbb{E}_Q e^{\alpha h}.$$

□

Proof of Lemma 2 . Adding and subtracting in the exponential $\int hdP$ and $\int hdQ$ resp we obtain the result: $\frac{1}{\alpha-1}\log\left(\int e^{(\alpha-1)h}dP\right) - \frac{1}{\alpha}\log\left(\int e^{\alpha h}dQ\right) = \frac{1}{\alpha-1}\log\left(\int e^{(\alpha-1)(h-\int hdP+\int hdP)}dP\right) - \frac{1}{\alpha}\log\left(\int e^{\alpha(h-\int hdQ+\int hdQ)}dQ\right) = \int hdP - \int hdQ + \frac{1}{\alpha-1}\log\left(\int e^{(\alpha-1)(h-\int hdP)}dP\right) - \frac{1}{\alpha}\log\left(\int e^{\alpha(h-\int hdQ)}dQ\right)$

□

Proof of Lemma 3. Note that we have for $0<\alpha<1$, $\frac{1}{\alpha}D_\alpha(P||Q) = \frac{1}{1-\alpha}D_{1-\alpha}(Q||P)$ (See Proposition 2 in van Erven and Harremos [2014]). Taking limits we obtain $\lim_{\alpha\to 0}\frac{1}{\alpha}D_\alpha(P||Q) = D_1(Q||P) = \mathsf{KL}(Q||P)$. □

Proof of Theorem 4 . For $0<\alpha<1$, we have for all $h\in\mathcal{H}$:

$$\int hdP - \int hdQ \le \frac{1}{\alpha}D_\alpha(P||Q) + \frac{1}{1-\alpha}\log\left(\int e^{(\alpha-1)(h-\int hdP)}dP\right) + \frac{1}{\alpha}\log\left(\int e^{\alpha(h-\int hdQ)}dQ\right) \tag{38}$$

Assuming r is bounded $0<r<b$ then we have $(r)_\sharp P - \mathbb{E}_P r$ and $(r)_\sharp Q - \mathbb{E}_Q r$ are sub-Gaussian with parameter $\sigma^2 = \frac{b^2}{4}$. Hence we have for $\lambda\in\mathbb{R}$:

$$\mathbb{E}_P e^{\lambda(r-\int rdP)} \le \exp\left(\frac{\lambda^2\sigma_P^2}{2}\right) \text{ and } \mathbb{E}_Q e^{\lambda(r-\int rdQ)} \le \exp\left(\frac{\lambda^2\sigma_Q^2}{2}\right),$$

Fix a finite $M > 0$. For $0 < \lambda < M$ and $P = \pi|x$ and $Q = \pi_{\text{ref}}|x$, consider $h_\lambda = \lambda r$, thanks to subgaussianity and boundedness of λ, $h_\lambda \in \mathcal{H}$ for all $\lambda \in (0, M)$. Hence we have by Equation (38) for all $\lambda \in (0, M)$:

$$\lambda\left(\int rdP - \int rdQ\right) \leq \frac{1}{\alpha}D_\alpha(P||Q) + \frac{1}{1-\alpha}\log\left(\int e^{\lambda(\alpha-1)(r-\int rdP)}dP\right) + \frac{1}{\alpha}\log\left(\int e^{\lambda\alpha(r-\int rdQ)}dQ\right)$$

we have by sub-Gaussianity:

$$\frac{1}{1-\alpha}\log\left(\int e^{\lambda(\alpha-1)(r-\int rdP)}dP\right) \leq \frac{1}{1-\alpha}\frac{\lambda^2(1-\alpha)^2\sigma_P^2}{2} = \frac{\lambda^2(1-\alpha)\sigma_P^2}{2}$$
$$\frac{1}{\alpha}\log\left(\int e^{\lambda\alpha(r-\int rdQ)}dQ\right) \leq \frac{1}{\alpha}\frac{\lambda^2\alpha^2\sigma_Q^2}{2} = \frac{\lambda^2\alpha\sigma_Q^2}{2}$$

It follows that for all $\lambda \in (0, M)$

$$\begin{aligned}\lambda\left(\int rd\pi|x - \int rd\pi_{\text{ref}}|x\right) &\leq \frac{1}{\alpha}D_\alpha(\pi|x||\pi_{\text{ref}}|x) + \frac{\lambda^2(1-\alpha)\sigma_P^2}{2} + \frac{\lambda^2\alpha\sigma_Q^2}{2} \\ &= \frac{1}{\alpha}D_\alpha(\pi|x||\pi_{\text{ref}}|x) + \frac{\lambda^2((1-\alpha)\sigma_P^2 + \alpha\sigma_Q^2)}{2}\end{aligned}$$

Integrating over x we obtain:

$$\lambda\left(\int rd\pi - \int rd\pi_{\text{ref}}\right) \leq \frac{1}{\alpha}D_\alpha(\pi||\pi_{\text{ref}}) + \frac{\lambda^2((1-\alpha)\sigma_P^2 + \alpha\sigma_Q^2)}{2}$$

Finally we have:

$$\int rd\pi - \int rd\pi_{\text{ref}} \leq \frac{1}{\lambda\alpha}D_\alpha(\pi||\pi_{\text{ref}}) + \frac{\lambda((1-\alpha)\sigma_P^2 + \alpha\sigma_Q^2)}{2}$$

minimizing over $\lambda \in (0, M)$: we obtain $\lambda^* = \sqrt{\frac{2D_\alpha(\pi||\pi_{\text{ref}})}{((1-\alpha)\sigma_P^2+\alpha\sigma_Q^2)\alpha}}$, M is free of choice, choosing $M = 2\lambda^*$, gives that λ^* is the minimizer and hence we have for all $\alpha \in (0, 1)$:

$$\int rd\pi - \int rd\pi_{\text{ref}} \leq \sqrt{\frac{2((1-\alpha)\sigma_P^2 + \alpha\sigma_Q^2)D_\alpha(\pi||\pi_{\text{ref}})}{\alpha}}.$$

□

F Goodhart Laws

Proof of Proposition 4. We have by duality:

$$\frac{1}{\beta}\log\left(\int e^{\beta r^*}d\pi_{\text{ref}}\right) = \sup_\nu \int r^*d\nu - \frac{1}{\beta}\mathsf{KL}(\nu||\pi_{\text{ref}})$$

hence for $\nu = \pi_{\beta,r}$ we have:

$$\frac{1}{\beta}\log\left(\int e^{\beta r^*}d\pi_{\text{ref}}\right) \geq \int r^*d\pi_{\beta,r} - \frac{1}{\beta}\mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})$$

Hence:

$$\int r^*d\pi_{\beta,r} \leq \frac{1}{\beta}\log\left(\int e^{\beta r^*}d\pi_{\text{ref}}\right) + \frac{1}{\beta}\mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})$$

On the other hand by optimality of $\pi_{\beta,r}$ we have:

$$\mathsf{KL}\left(\pi_{\beta,r}||\pi_{\text{ref}}\right) = \beta\int rd\pi_{\beta,r} - \log\left(\int e^{\beta r}d\pi_{\text{ref}}\right)$$

Hence we have:

$$\int r^* d\pi_{\beta,r} \leq \frac{1}{\beta}\log\left(\int e^{\beta r^*} d\pi_{\text{ref}}\right) + \int r d\pi_{\beta,r} - \frac{1}{\beta}\log\left(\int e^{\beta r} d\pi_{\text{ref}}\right) \leq \int r d\pi_{\beta,r} + \frac{1}{\beta}\log\left(\frac{\int e^{\beta r^*} d\pi_{\text{ref}}}{\int e^{\beta r} d\pi_{\text{ref}}}\right)$$

It follows that:

$$\int r^* d\pi_{\beta,r} - \int r^* d\pi_{\text{ref}} \leq \int r d\pi_{\beta,r} - \int r d\pi_{\text{ref}} + \frac{1}{\beta}\log\left(\frac{\int e^{\beta(r^* - \int r^* d\pi_{\text{ref}})} d\pi_{\text{ref}}}{\int e^{\beta(r - \int r d\pi_{\text{ref}})} d\pi_{\text{ref}}}\right)$$

$$\begin{aligned}\frac{\int e^{\beta(r^* - \int r^* d\pi_{\text{ref}})} d\pi_{\text{ref}}}{\int e^{\beta(r - \int r d\pi_{\text{ref}})} d\pi_{\text{ref}}} &= \int e^{\beta(r^* - r - (\int r^* d\pi_{\text{ref}} - \int r d\pi_{\text{ref}})} \frac{e^{\beta r} d\pi_{\text{ref}}}{\int e^{\beta r} d\pi_{\text{ref}}} \\ &= \int e^{\beta(r^* - r - (\int r^* d\pi_{\text{ref}} - \int r d\pi_{\text{ref}})} d\pi_{\beta,r}\end{aligned}$$

Hence we have finally:

$$\int r^* d\pi_{\beta,r} - \int r^* d\pi_{\text{ref}} \leq \int r d\pi_{\beta,r} - \int r d\pi_{\text{ref}} + \frac{1}{\beta}\log\left(\int e^{\beta(r^* - r - (\int r^* d\pi_{\text{ref}} - \int r d\pi_{\text{ref}})} d\pi_{\beta,r}\right)$$

$$\int r^* d\pi_{\beta,r} - \int r^* d\pi_{\text{ref}} \leq \int r d\pi_{\beta,r} - \int r d\pi_{\text{ref}} - \frac{1}{\beta}\log\left(\int e^{\beta(r - r^* - (\int r d\pi_{\text{ref}} - \int r^* d\pi_{\text{ref}})} d\pi_{\beta,r^*}\right)$$

The proof follows from using the subgaussianity of $r_\sharp \pi_{\text{ref}}$ and the assumption on the soft max. □

Proof of Proposition 5.

$$\mathbb{E}_\pi(r^* - r) - \mathbb{E}_{\pi_{\text{ref}}}(r^* - r) \leq 2||r - r^*||_\infty \mathsf{TV}(\pi, \pi_{\text{ref}})$$

For $\pi^{(n)}_{r,\text{ref}}$, we have:

$$\mathbb{E}_{\pi^{(n)}_{r,\text{ref}}}(r^*) - \mathbb{E}_{\pi_{\text{ref}}}(r^*) \leq \mathbb{E}_{\pi^{(n)}_{r,\text{ref}}}(r) - \mathbb{E}_{\pi_{\text{ref}}}(r) + 2||r - r^*||_\infty \mathsf{TV}(\pi^{(n)}_{r,\text{ref}}, \pi_{\text{ref}})$$

and

$$\mathbb{E}_{\pi^{(n)}_{r,\text{ref}}}(r^*) - \mathbb{E}_{\pi_{\text{ref}}}(r^*) \geq \mathbb{E}_{\pi^{(n)}_{r,\text{ref}}}(r) - \mathbb{E}_{\pi_{\text{ref}}}(r) - 2||r - r^*||_\infty \mathsf{TV}(\pi^{(n)}_{r,\text{ref}}, \pi_{\text{ref}})$$

By the data processing inequality we have: $\mathsf{TV}(\pi^{(n)}_{r,\text{ref}}, \pi_{\text{ref}}) \leq \mathsf{TV}(R^{(n)}_{r,\text{ref}}, R) = (\frac{1}{n})^{\frac{1}{n-1}} - (\frac{1}{n})^{\frac{n}{n-1}}$
If r has subguassian tails under π_{ref} than we have:

$$\mathbb{E}_{\pi^{(n)}_{r,\text{ref}}}(r^*) - \mathbb{E}_{\pi_{\text{ref}}}(r^*) \leq \sqrt{2\sigma^2\left(\log(n) - \frac{n-1}{n}\right)} + 2||r - r^*||_\infty \left((\frac{1}{n})^{\frac{1}{n-1}} - (\frac{1}{n})^{\frac{n}{n-1}}\right)$$

$$\mathbb{E}_{\pi^{(n)}_{r,\text{ref}}}(r^*) - \mathbb{E}_{\pi_{\text{ref}}}(r^*) \leq \sqrt{2\sigma^2\left(\log(n) - \frac{n-1}{n}\right)} + 2\inf_{r\in\mathcal{H}} ||r - r^*||_\infty \left((\frac{1}{n})^{\frac{1}{n-1}} - (\frac{1}{n})^{\frac{n}{n-1}}\right).$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: we summarize the theoretical contributions in the abstract and the introduction and put them in context.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: all results are given under their assumptions that are discussed and justified

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: all results are given under their assumptions that are discussed and justified, some proofs are developed in the paper, others are given in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper is theoretical

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper is theoretical

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification:The paper is theoretical

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper is theoretical

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper is theoretical

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is theoretical

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper is theoretical

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper is theoretical

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification:The paper is theoretical

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper is theoretical

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper is theoretical

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper is theoretical

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.