Aligning Language Models with Preferences through f-divergence Minimization

Dongyoung Go [1 2] **Tomasz Korbak** [3] **Germán Kruszewski** [4] **Jos Rozen** [4] **Nahyeon Ryu** [1] **Marc Dymetman** [5]

Abstract

Aligning language models with preferences can be posed as approximating a target distribution representing some desired behavior. Existing approaches differ both in the functional form of the target distribution and the algorithm used to approximate it. For instance, Reinforcement Learning from Human Feedback (RLHF) corresponds to minimizing a reverse KL from an *implicit* target distribution arising from a KL penalty in the objective. On the other hand, Generative Distributional Control (GDC) has an *explicit* target distribution and minimizes a forward KL from it using the Distributional Policy Gradient (DPG) algorithm. In this paper, we propose a new approach, f-DPG, which allows the use of *any* f-divergence to approximate *any* target distribution that can be evaluated. f-DPG unifies both frameworks (RLHF, GDC) and the approximation methods (DPG, RL with KL penalties). We show the practical benefits of various choices of divergence objectives and demonstrate that there is no universally optimal objective but that different divergences present different alignment and diversity trade-offs. We show that Jensen-Shannon divergence strikes a good balance between these objectives, and frequently outperforms forward KL divergence by a wide margin, leading to significant improvements over prior work. These distinguishing characteristics between divergences persist as the model size increases, highlighting the importance of selecting appropriate divergence objectives.

1. Introduction

Language models (LMs) have recently revolutionized the field of Natural Language Processing thanks to their generative capabilities, which are useful in a vast number of tasks (Brown et al., 2020; Srivastava et al., 2022). However, generated texts can also violate widely-held human preferences, e.g. helpfulness (Askell et al., 2021), non-offensiveness (Gehman et al., 2020), truthfulness (Lin et al., 2022) or equal treatment (Cao et al., 2022). Aligning LMs with human preferences is the problem of adapting the LM in such a way that generated content is perceived to match the human's intent (Ouyang et al., 2022) or that it is helpful, honest, and harmless (Askell et al., 2021; Bai et al., 2022b). Fundamentally, an aligned LM can be seen as a desired target distribution that we would like to generate from (Korbak et al., 2022c). Some approaches leave this distribution implicit, to be defined as a side-effect of the proposed intervention. These include prompting with natural language instructions or demonstrations (Askell et al., 2021), using scorers or safety filters while decoding (Roller et al., 2021; Xu et al., 2021), supervised fine-tuning on curated data (Solaiman & Dennison, 2021; Ngo et al., 2021; Welbl et al., 2021; Chung et al., 2022) or selected samples from the model (Zelikman et al., 2022; Scheurer et al., 2022; Dohan et al., 2022), and fine-tuning the language model using reinforcement learning with a learned reward function that approximates human feedback (Reinforcement Learning from Human Feedback or RLHF; Ziegler et al., 2019; Bai et al., 2022a; Ouyang et al., 2022). Instead, Khalifa et al. (2021) propose a framework that they name Generation with Distributional Control (GDC), where they define the target

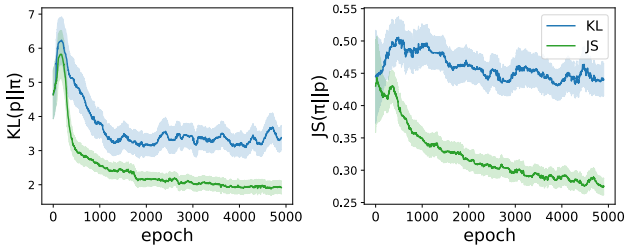


Figure 1. On many target distributions, the Jensen-Shannon (JS) divergence (green) outperforms the Kullback-Leibler (KL) divergence (blue) as an *objective,* even when performance is measured in terms of KL from the target p (left panel, ↓ better). See Sec. 4.2.

[1]Naver Corp [2]Yonsei University [3]University of Sussex [4]Naver Labs Europe [5]Independent Researcher. Correspondence to: Dongyoung Go <dongyoung.go@navercorp.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

distribution p that represents the aligned LM as an EBM (Energy Based Model), namely an unnormalized version of p that can be evaluated over any input x. They then train the generative model π_θ to approximate p via methods such as Distributional Policy Gradients (DPG; Parshakova et al., 2019), which minimize the forward Kullback-Leibler (KL) divergence $\mathrm{KL}(p||\pi_\theta)$ of p to π_θ. The advantage of such an approach is that it decouples the problem of describing the aligned LM from the problem of approximating it. Furthermore, even if RL with KL penalties (Todorov, 2006a; Kappen et al., 2012; Jaques et al., 2017; 2019), the method used to fine-tune a LM in RLHF, is defined only in terms of reward maximization, it has also been shown to be equivalent to minimizing the *reverse* KL divergence $\mathrm{KL}(\pi_\theta||p)$ of π_θ to a target distribution p that can also be written explicitly in closed-form (Korbak et al., 2022b).

The possibility of approximating various distributions according to different divergence measures begs the question: Does the choice of a divergence measure matter? In principle, all divergences lead to the same optimum, namely the target distribution p. However, when we restrict π_θ to a certain parametric family that does not include p (i.e., the search space is *mis-specified*), then the minimum can be found at different points, leading to optimal models with different properties. Moreover, different divergences present different loss landscapes: some might make it easier for stochastic gradient descent to find good minima. Finally, the space of possible divergence measures and forms of target distributions is a vast and largely uncharted terrain. Prior work has largely failed to decouple the form of a target distribution and the algorithm used for approximating it.

Here, we introduce f-DPG, a new framework for fine-tuning an LM to approximate any given target EBM, by exploiting any given divergence in the f-divergences family, which includes not only the forward KL and the reverse KL cited above, but also Total Variation (TV) distance, Jensen-Shannon (JS) divergence, among others. f-DPG generalizes existing approximation techniques both DPG and RL with KL penalties algorithms, thus allowing us to investigate new ways to approximate the target distributions defined by the GDC and RLHF frameworks. In particular, we explore the approximation of various target distributions representing different alignment goals, which include imposing lexical constraints, reducing social bias with respect to gender and religion, enforcing factual consistency in summarization, and enforcing compilability of generated code. We focus our experiments on four instantiations of f-DPG, namely KL-DPG, RKL-DPG, TV-DPG and JS-DPG, whose objective is to minimize the forward KL, reverse KL, TV and JS divergences, respectively, and evaluate each experiment in terms of approximation quality as measured by all of these f-divergences. We show that we can obtain significantly improved results over the original KL-DPG algorithm (Parshakova et al., 2019) by minimizing other f-divergences, even when the approximation quality is evaluated under the lens of the forward KL. Furthermore, we observe that while there is no single best optimization objective for all cases, JS-DPG often strikes a good balance and significantly improves upon prior work (Khalifa et al., 2021; Korbak et al., 2022a), as illustrated in Fig. 1. Lastly, we find that f-DPG with an optimal objective continues to outperform suboptimal objectives as we scale model size from 127M parameters to 1.5B parameters (Sec. 4.5). The smooth and gradual scaling trend observed with increasing model size suggests that our findings will generalize to even larger LMs.

Overall, the contributions of the paper include:

1. Introducing f-DPG, a unifying framework for approximating any EBM target distribution by minimizing any f-divergence (Sec. 3.2), and deriving a universal formula for gradient descent with f-divergences (Theorem 1).

2. Extending f-DPG to include baselines for variance reduction (Fact 1); and handling conditional target distributions (Fact 2).

3. Investigating the performance of f-DPG on a diverse array of thirteen LM alignment tasks, three forms of target distributions, four f-divergence objectives and eight metrics.

2. Background

We can organize approaches to LM alignment along two axes: how the target distribution is constructed and how it is approximated. The first problem roughly corresponds to representing human preferences through the specification of a probability distribution and the second to allowing the production of samples from that distribution.

2.1. Defining a Target Distribution

The target distribution expresses an ideal notion of an LM, incorporating human preferences, as probabilities $p(x)$ over texts x according to how well they satisfy the preferences. Formally, $p(x)$ is often defined through a non-negative function $P(x)$ (aka an *energy-based model* or EBM (LeCun et al., 2006)) such that $p(x) \propto P(x)$. The model $P(x)$ (and $p(x)$ after normalization) can be used to score samples, but not to directly produce them because it lacks an autoregressive form. In the rest of the paper, we will focus on target distributions modeling three types of preferences prominently employed in recent literature about GDC (Khalifa et al., 2021) and RLHF (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Menick et al., 2022; Bai et al., 2022a).

Binary preferences For human preferences naturally expressible as a binary constraint $b(x) \in \{0, 1\}$ (e.g. a sample x must never contain a curse word), Khalifa et al. (2021) proposed the following target distribution:

$$p_{\text{GDC\_bin}}(x) \propto a(x)b(x), \tag{1}$$

where a is a pretrained LM and $b(x) = 0$ if x contains a curse and $b(x) = 1$ otherwise. $p_{\text{GDC\_bin}}$ is the distribution enforcing that all samples match the binary constraint, which deviates minimally from a as measured by $\text{KL}(p_{\text{GDC\_bin}}||a)$.

Scalar preferences Some human preferences, such as helpfulness, are more naturally expressed as scalar scores. Alignment with respect to these is typically addressed with RLHF (Stiennon et al., 2020; Ziegler et al., 2019; Ouyang et al., 2022), which consists of, first, capturing human preferences as a reward function $r(x)$ (e.g. scores given a reward model trained to predict human preferences) and second, applying RL with KL penalties (Todorov, 2006a; Kappen et al., 2012; Jaques et al., 2017; 2019) to maximize this reward while penalizing departure from $a(x)$:

$$J_{\text{RLKL}}(\theta) = \mathbb{E}_{x\sim\pi_\theta}\left[r(x) - \beta \log \frac{\pi_\theta(x)}{a(x)}\right]. \tag{2}$$

This objective can be equivalently framed as minimizing the reverse KL, $\text{KL}(\pi_\theta||p_{\text{RLKL}})$, where the target distribution p_{RLKL} is defined as:

$$p_{\text{RLKL}}(x) \propto a(x) \exp(r(x)/\beta), \tag{3}$$

where β is a hyperparameter (Korbak et al., 2022b).

Distributional preferences Finally, there is a class of distributional preferences (Weidinger et al., 2021) that cannot be expressed as a function of a single sample x but depend on the entire distribution, e.g. a particular gender distribution of persons mentioned in LM samples. Khalifa et al. (2021) model such preferences through distributional constraints using the following exponential family target distribution

$$p_{\text{GDC\_dist}}(x) \propto a(x) \exp\Big[\sum_i \lambda_i \phi_i(x)\Big], \tag{4}$$

where ϕ_i are features defined over texts (e.g. the most frequent gender of people mentioned in x) and λ_i are coefficients chosen so that the expected values $\mathbb{E}_{x\sim p}\left[\phi_i(x)\right]$ match some desired values $\bar{\mu}_i$ (e.g., 50% gender balance). The resulting distribution $p_{\text{GDC-d}}$ matches the target feature moments, while deviating minimally from a as measured by $\text{KL}(p_{\text{GDC\_dist}}||a)$.

2.2. Approximating the target distribution

Drawing samples from a target distribution p constitutes the inference problem. There are broadly two approaches to this problem: (i) augmenting decoding from a at inference time to obtain samples from p and (ii) training a new parametric model π_θ to approximate p which can then be sampled from directly. The first family of approaches includes guided decoding methods (Dathathri et al., 2020; Qin et al., 2022), Monte Carlo sampling techniques such as rejection sampling to sample from simple distributions like $p_{\text{GDC\_bin}}$ (Roller et al., 2021; Ziegler et al., 2022), and Quasi Rejection Sampling (QRS) (Eikema et al., 2022) or MCMC techniques (Miao et al., 2019; Goyal et al., 2022) to sample from more complex distributions, such as $p_{\text{GDC\_dist}}$. In the rest of the paper, we will focus on the second family: methods that train a new model π_θ to approximate p by minimizing a divergence measure from p, $D(\pi_\theta||p)$. Khalifa et al. (2021) uses Distributional Policy Gradients (DPG; Parshakova et al., 2019) to approximate the target distribution by minimizing $\text{KL}(p||\pi_\theta)$, or equivalently, $\text{CE}(p, \pi_\theta)$:

$$\nabla_\theta \text{CE}(p, \pi_\theta) = -\mathbb{E}_{x\sim\pi_\theta} \frac{p(x)}{\pi_\theta(x)} \nabla_\theta \log \pi_\theta(x). \tag{5}$$

3. Formal Aspects

In this section, we describe the f-divergence family, and introduce a generic technique, f-DPG, for minimizing the f-divergence between a target distribution p and a model π_θ. We then describe the application of f-DPG to aligning language models with human preferences.

3.1. f-divergences

Consider a convex function $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$. Let $f(0) \doteq \lim_{t\to 0} f(t)$ and $f^{'}(\infty) \doteq \lim_{t\to 0} tf(\frac{1}{t})$.[1] Let p_1, p_2 be two distributions over a discrete set $\mathcal{X}$. The f-divergence between p_1 and p_2 can be defined as

$$D_f(p_1||p_2) \doteq \mathbb{E}_{x\sim p_2}\left[f\left(\frac{p_1(x)}{p_2(x)}\right)\right] + f^{'}(\infty)\, p_1(p_2 = 0) \tag{6}$$

where $p_1(p_2 = 0)$ is the p_1-mass of the set $\{x \in \mathcal{X} : p_2(x) = 0\}$ (Polyanskiy, 2019; Liese & Vajda, 2006). The function f is called a generator of D_f. By convention, if $p_1(p_2 = 0) = 0$, the last term of Eq. (6) is set to 0 regardless of the value of $f^{'}(\infty)$ (which can be infinite).[2] It can be

[1] The limits are well-defined and take values in $(-\infty, \infty]$. The convention for $f^{'}(\infty)$ is motivated by the fact that $\lim_{t\to\infty} f'(t) = \lim_{t\to 0} tf(\frac{1}{t})$ (Hiriart-Urruty & Lemaréchal, 2013).

[2] Based on the commonly made assumption that the support of p_1 is dominated by the support of p_2 ($Supp(p_1) \subset Supp(p_2)$), Eq. (6) simplifies to $D_f(p_1||p_2) = \mathbb{E}_{x\sim p_2}\left[f\left(\frac{p_1(x)}{p_2(x)}\right)\right]$.

shown that $D_f(p_1||p_2) \geq 0$ for any p_1 and p_2, with equality if $p_1 = p_2$; conversely, if $D_f(p_1||p_2) = 0$ and f is strictly convex at 1, then $p_1 = p_2$.

The f-divergence family includes many important divergence measures, in particular KL divergence $\mathrm{KL}(p_1||p_2)$, reverse KL divergence $\mathrm{KL}(p_2||p_1)$, Jensen-Shannon divergence, and Total Variation distance. We list these f-divergences and their generators in Tab. 1. For more details about notations and properties of f-divergences, see App. A.1 and also Liese & Vajda (2006); Polyanskiy (2019); Sason & Verdú (2016); Sason (2018).

3.2. Distributional alignment with f-divergences

Let $\mathcal{X}$ be a discrete countable or finite set, in our case a set of texts. Given a target probability distribution $p(x)$ over elements $x \in \mathcal{X}$, our goal is to approximate p with a generative model (aka policy) π_θ. On the other hand, the generative model π_θ is a parametric model, typically an autoregressive neural network, from which we can (i) directly sample and (ii) evaluate probabilities $\pi_\theta(x)$.

We approach this problem by attempting to minimize the f-divergence of π_θ to p:[3]

$$\min_{\theta \in \Theta} D_f(\pi_\theta||p), \tag{7}$$

where θ varies inside the parametric family Θ. Note that when the family $\pi_\theta, \theta \in \Theta$ is "well-specified", i.e., when $\exists \theta_0$ s.t. $p = \pi_{\theta_0}$, the true minimum of Eq (7) is 0, attained at θ_0, whatever divergence D_f is chosen. In contrast, when the family is "mis-specified" i.e. does not include p, the distribution π_θ with minimal divergence can be strongly dependent on the chosen divergence D_f.

Eq. (7) might be solved approximately using stochastic optimization with samples drawn from the distribution p, as the definition of $D_f(\pi_\theta||p)$ involves taking the expectation with respect to p. However, it is often not possible to sample directly from p, while it is possible to sample from π_θ. Our optimization technique is then based on the following core result, which we prove in App. A.3.

Theorem 1. *Let p and π_θ be distributions over a discrete set $\mathcal{X}$ such that at least one of the following conditions holds: (i) $\forall \theta \in \Theta$, Supp$(p) \subset$ Supp(π_θ), or (ii) Supp(π_θ) does not depend on θ. Then:*

$$\nabla_\theta D_f(\pi_\theta||p) = \mathbb{E}_{x\sim\pi_\theta}\left[f'\left(\frac{\pi_\theta(x)}{p(x)}\right)\nabla_\theta \log \pi_\theta(x)\right]. \tag{8}$$

[3]We could have chosen to do $\min_{\theta\in\Theta} D_f(p||\pi_\theta)$. However the *perspective transform* $f^*(t) \doteq t\, f(\frac{1}{t})$ allows interchangeability of arguments: $D_f(\pi_\theta||p) = D_{f^*}(p||\pi_\theta)$, making either form possible. The form in Eq. (7) permits a simpler statement of our main theorem. See App. A.1, A.3 for details.

Note that it may happen in Eq 8 that $p(x) = 0$ and $\pi_\theta(x) > 0$, hence $\frac{\pi_\theta(x)}{p(x)} = \infty$, in which case the expression $f'\left(\frac{\pi_\theta(x)}{p(x)}\right)$ should be understood as denoting the value $f'(\infty)$ as defined earlier.[4]

In the context of LMs, our domain of application, we will use Thm. 1 in situations where π_θ, being a standard softmax-based autoregressive model, has full support over $\mathcal{X}$ (i.e. Supp$(\pi_\theta) = \mathcal{X}$) for all θ's, while the support of p might be strictly included in $\mathcal{X}$ in some experiments (Sec. 4.2, 4.4).

It is instructive to consider Thm. 1 in relation to rewards in RL. In the standard policy gradient algorithm (Williams, 1992), to find the model that maximizes the average reward $\mathbb{E}_{x\sim\pi_\theta}[r(x)]$, one computes the gradient of the loss using the formula $\nabla_\theta \mathbb{E}_{x\sim\pi_\theta}[r(x)] = \mathbb{E}_{x\sim\pi_\theta}[r(x)\nabla_\theta \log \pi_\theta(x)]$. The gradient in Eq. 8 is very similar, with a "pseudo-reward" $r_\theta(x) = -f'(\frac{\pi_\theta(x)}{p(x)})$, one difference being that now r_θ depends on θ (see (Korbak et al., 2022b) for related remarks). We refer to the approach in Eq. 8 under the name *f-DPG*, in reference to the original DPG (Distributional Policy Gradient) approach introduced in (Parshakova et al., 2019), which can be seen as a special case of f-DPG ("KL-DPG") with $D_f(\pi_\theta||p)$ set to $\mathrm{KL}(p||\pi_\theta)$ as discussed in Sec. 3.4.

3.3. Adding a baseline

Based on the similarity to policy gradients, we adopt the widely used *baseline* technique from RL, as previously studied in Williams (1992); Baxter & Bartlett (2001); Schulman et al. (2016) and in the context of DPG in (Korbak et al., 2022b). This technique involves subtracting a constant B from the reward term, and does not introduce bias in the estimate of the gradient at a given θ. In our case, with $r_\theta(x) \doteq -f'(\frac{\pi_\theta(x)}{p(x)})$, we can write $\nabla_\theta D_f(\pi_\theta||p) = \mathbb{E}_{x\sim\pi_\theta} r_\theta(x)\nabla_\theta \log \pi_\theta(x) = \mathbb{E}_{x\sim\pi_\theta}(r_\theta(x) - B)\,\nabla_\theta \log \pi_\theta(x)$, based on the observation that $\mathbb{E}_{x\sim\pi_\theta}\nabla_\theta \log \pi_\theta(x) = 0$ (see also App. A.6).

Fact 1. *Subtracting B from $r_\theta(x)$ does not introduce bias into f-DPG gradient estimates.*

Typically, B is chosen to be the average of the rewards, $B \doteq \mathbb{E}_{x\sim\pi_\theta}[r_\theta(x)]$. In the experiments of Sec. 4, we use the baseline technique where B is an estimate of the average of pseudo-rewards, unless otherwise specified.

[4]The derivative $f'(t)$ of any convex function $f(t)$ is defined almost everywhere, with the possible exception of a countable number of non-differentiable points, at which a subgradient can be used instead (Hiriart-Urruty & Lemaréchal, 2013; Rockafellar, 1970). See also App. A.4.

$D_f(\pi_\theta\|\|p)$	f	f'	$f'\left(\frac{\pi_\theta(x)}{p(x)}\right)$	$f'(\infty)$
Forward KL ($\mathrm{KL}(p\|\|\pi_\theta)$)	$f(t) = -\log t$	$f'(t) = -\frac{1}{t}$	$-\frac{p(x)}{\pi_\theta(x)}$	0
Reverse KL ($\mathrm{KL}(\pi_\theta\|\|p)$)	$f(t) = t\log t$	$f'(t) = \log t + 1$	$-\left(\log\frac{p(x)}{\pi_\theta(x)}\right) + 1$	∞
Total Variation ($\mathrm{TV}(\pi_\theta\|\|p)$)	$f(t) = 0.5\,\lvert 1-t\rvert$	$f'(t) = \begin{cases} 0.5 & \text{for } t > 1 \\ -0.5 & \text{for } t < 1 \end{cases}$	$\begin{cases} 0.5 & \text{for } \frac{\pi_\theta(x)}{p(x)} > 1 \\ -0.5 & \text{for } \frac{\pi_\theta(x)}{p(x)} < 1 \end{cases}$	0.5
Jensen-Shannon ($\mathrm{JS}(\pi_\theta\|\|p)$)	$f(t) = t\log\frac{2t}{t+1} + \log\frac{2}{t+1}$	$f'(t) = \log\frac{2t}{t+1}$	$\log 2 - \log\left(1 + \frac{p(x)}{\pi_\theta(x)}\right)$	$\log 2$

Table 1. Some common f-divergences $D_f(\pi_\theta||p)$. In the convention of this table, the f shown corresponds to the order of arguments $D_f(\pi_\theta||p)$. Thus the forward KL between the target p and the model, $\mathrm{KL}(p||\pi_\theta)$, corresponds to $D_{-\log t}(\pi_\theta||p)$, and similarly for the reverse KL, $\mathrm{KL}(\pi_\theta||p)$, which corresponds to $D_{t\log t}(\pi_\theta||p)$, etc. Note that for symmetric divergences (TV and JS) the order of arguments is indifferent: $\mathrm{TV}(\pi_\theta||p) = \mathrm{TV}(p||\pi_\theta)$, $\mathrm{JS}(\pi_\theta||p) = \mathrm{JS}(p||\pi_\theta)$.

3.4. Recovering Some Existing Methods

Various existing methods for aligning LM with preferences can be included in the f-DPG framework.

GDC In GDC, fitting the policy π_θ to the target p (which is given by either one of Eq. 1 or Eq. 4) is done using DPG (Parshakova et al., 2019), namely by minimizing the **forward KL**, $\mathrm{KL}(p||\pi_\theta)$. In the f-DPG framework, $\mathrm{KL}(p||\pi_\theta) = D_f(\pi_\theta||p)$ with $f(t) = -\log t$, $f'(t) = -1/t$, and Thm. 1 leads to the formula:

$$\nabla_\theta D_f(\pi_\theta||p) = \mathbb{E}_{x\sim\pi_\theta} -\frac{p(x)}{\pi_\theta(x)}\nabla_\theta \log \pi_\theta(x),$$

which is equivalent to Eq. 5.

RL with KL penalties Let's rewrite the target distribution of Eq. (3) as $p(x) \doteq p_{\mathrm{RLKL}}(x) = 1/Z\; a(x)\, e^{r(x)/\beta}$, where Z is a normaliser. Then $\mathrm{KL}(\pi_\theta||p) = D_f(\pi_\theta||p)$, with $f(t) = t\log t$ corresponding to **reverse KL**, and $f'(t) = 1 + \log t$. Thm. 1 implies that:

$$\begin{aligned}
&\nabla_\theta D_f(\pi_\theta||p) \\
&= \mathbb{E}_{x\sim\pi_\theta}\left(1 + \log\frac{\pi_\theta(x)}{Z^{-1}a(x)\exp(r(x)/\beta)}\right)\nabla_\theta\log\pi_\theta(x) \\
&= \mathbb{E}_{x\sim\pi_\theta}\left(-\frac{r(x)}{\beta} + \log\frac{\pi_\theta(x)}{a(x)}\right)\nabla_\theta\log\pi_\theta(x),
\end{aligned}$$

where we have exploited the fact that $1 + \log Z$ is a constant, hence $\mathbb{E}_{x\sim\pi_\theta}(1 + \log Z)\;\nabla_\theta\log\pi_\theta(x) = 0$. Up to the constant factor β, this form recovers the usual formula for estimating the gradient of the loss defined in Eq. (2): $\nabla_\theta J_{\mathrm{RLKL}}(\theta) = \mathbb{E}_{x\sim\pi_\theta}\left(r(x) - \beta\log\frac{\pi_\theta(x)}{a(x)}\right)\nabla_\theta\log\pi_\theta(x)$.

3.5. Estimating Z

The target distribution p is often defined as $p(x) \propto P(x)$, where $P(x)$ is a non-negative function over $\mathcal{X}$. The distribution p can then be computed as $p(x) = 1/Z\; P(x)$, where Z is the normalizing constant (partition function) defined by $\sum_{x\in\mathcal{X}} P(x)$. An estimate of Z can be obtained by importance sampling, using samples from the current π_θ, based on the identity $Z = \mathbb{E}_{\pi_\theta}\frac{P(x)}{\pi_\theta(x)}$. Each such estimate is unbiased, and by averaging the estimates based on different π_θ's, one can obtain a more precise estimate of Z, exploiting *all* the samples obtained so far. For details about the estimate of Z, see Algorithm 1 in App. A.3, as well as the ablation study in App. H.3.

3.6. Conditional Target Distributions

For a conditional task such as machine translation, summarization or dialogue, where π_θ is defined as a conditional distribution $\pi_\theta(x|c)$, we adapt the conditional generalization of DPG introduced in Korbak et al. (2022a). Given a distribution over contexts $\tau(c)$ and a map from a context c to a target distribution p_c, we have (see App. E for details):

Fact 2. *f-DPG is generalized to the conditional case by optimizing the loss*

$$\mathbb{E}_{c\sim\tau(c)}\left[\nabla_\theta D_f(\pi_\theta(\cdot\,|c)||p_c(\cdot\,))\right]. \tag{9}$$

4. Experiments

We study four instantiations of f-DPG, namely KL-DPG, RKL-DPG, TV-DPG and JS-DPG, corresponding to minimizing the forward KL, reverse KL, Total Variation, and Jensen-Shannon divergences, respectively. We use an exponential moving average baseline with weight $\alpha = 0.99$ for all, except for KL-DPG, where we use the analytically computed value of the pseudo-reward expectation, which amounts to 1 (Korbak et al., 2022b). We evaluate them on a diverse array of tasks including imposing sentiment constraints (Sec. 4.1), lexical constraints (Sec. 4.2), debiasing genders' prevalence and religious groups' regard (Sec. 4.3), and context-conditioned tasks, such as enforcing factual consistency in summarization (Sec. 4.4) or compilability of generated code (see App. E.1). Unless specified otherwise, we use a pretrained GPT-2 "small" (Radford et al.,

2019) with 117M parameters for the initial model. Yet, we demonstrate in Sec. 4.5 that the observations continue to hold for models of larger size. Implementation details and hyper-parameters are available in App. C.

Metrics We report the following key metrics. We add task-specific metrics if needed.

1. $D_f(\pi_\theta||p)$, the f-divergence between p and π_θ, with four different f's corresponding to forward KL, $\mathrm{KL}(p||\pi_\theta)$; reverse KL, $\mathrm{KL}(\pi_\theta||p)$; Total Variation, $\mathrm{TV}(\pi_\theta||p)$; and Jensen-Shannon, $\mathrm{JS}(\pi_\theta||p)$. We use importance sampling to estimate these divergences.
2. $\mathrm{KL}(\pi_\theta||a)$, a measure of the divergence from original LM a (Ziegler et al., 2019; Khalifa et al., 2021).
3. Alignment score, measured by moments $\mathbb{E}_{x\sim\pi_\theta}\phi(x)$ of a feature of interest $\phi(x)$.
4. Normalized Entropy (Berger et al., 1996), a measure of diversity in probability distribution normalized by number of tokens.
5. Standard deviation of a minibatch's pseudo-rewards, $\mathrm{std}(r_\theta(x))$, where r_θ is defined as in Sec. 3.3.

4.1. Alignment with Scalar Preferences

Task We begin with the task of maximizing a scalar preference with KL penalties, whose target distribution, p_{RLKL}, is defined in Eq. 3. We set $r(x) = \log\phi(x)$ where $\phi(x)$ is the probability returned by a sentiment classifier fine-tuned from Distil-BERT (HF Canonical Model Maintainers, 2022). This reward function is optimal for modeling a decision-maker which given k different samples $x_1, \ldots, x_k$, will pick x_i with probability proportional to $\phi(x_i)$ (see Appendix F). We set $\beta = 0.1$, which is in line with the range of values explored by Ziegler et al. (2019). Note that applying RKL-DPG on p_{RLKL} is equivalent to the RL with KL penalties method, as described in Sec. 3.4. However, through f-DPG we can explore alternative objectives to approximate the same target.

Results Fig. 2 shows the evolution of the above-mentioned metrics. Further details are given in Fig. 11 in the Appendix. We observe that whereas RKL-DPG achieves by far the best performance in terms of reverse KL, $\mathrm{KL}(\pi_\theta||p)$ (top-right), it fails to minimize all other divergence metrics. This shows that minimizing one divergence does not necessarily imply that other divergences will follow. Notably, RKL-DPG yields the highest value of alignment score $\mathbb{E}_{\pi_\theta}[\phi(x)]$ at the cost of a significant departure from a. We connect this to the strong influence that low values $p(x)$ have on RKL-DPG, which induces a large pseudo-reward for strongly reducing $\pi_\theta(x)$ on those samples (see Sec 5) and produces the spike at the beginning of training in $\mathrm{std}(\mathrm{rewards})$. This can lead $\pi_\theta(x)$ to concentrate on high-probability regions of $p(x)$, at the cost of diversity, which can also be seen in the low entropy of the generated samples. Interestingly, the three remaining variants of DPG (KL, TV and JS) consistently minimize all four tracked divergences, with JS-DPG performing best overall.

In App. D.1, we show additional metrics on generated sentences, which show low diversity but high quality for RKL-DPG, compared to other f-DPGs, suggesting it captures a subset of the target distribution ("mode collapse"), as commonly observed in other generative models (Huszar, 2015; Che et al., 2017; Mescheder et al., 2018).

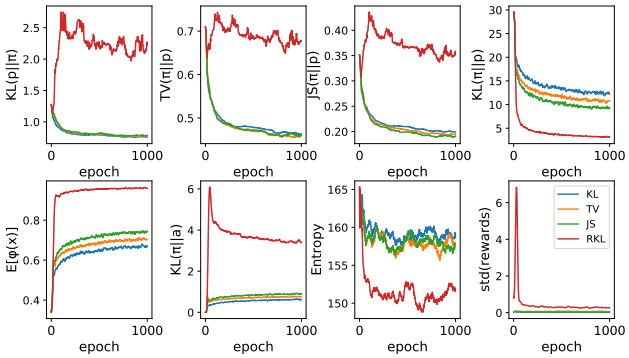


Figure 2. Comparison of f-DPG on sentiment preference. Evaluation metrics: four f-divergences $D_f(\pi_\theta||p)$ (↓ better), alignment score $\mathbb{E}_{\pi_\theta}[\phi(x)]$ (↑ better), entropy (↑ better), standard deviation of pseudo-reward $\mathrm{std}(r_\theta(x))$.

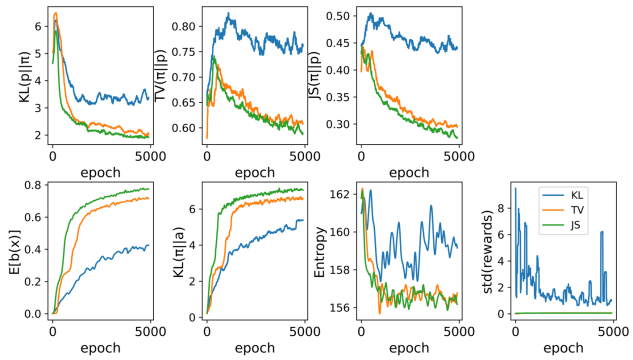
4.2. Alignment with Lexical Constraints

Task In this task, we constrain the presence of a specific word in the generated text. Following Khalifa et al. (2021), we formulate this goal as a binary preference on the LM by using a target distribution $p_{\mathrm{GDC\_bin}}$, where $b(x) = 1$ iff the target word appears in the sequence x, and using a scalar preference target distribution p_{RLKL} where $r(x)$ is set in the same way as $b(x)$ above. Note that in the GDC framework, $p_{\mathrm{GDC\_bin}}(x) = 0$ when $b(x) = 0$, implying that reverse KL, namely $\mathrm{KL}(\pi_\theta||p)$, becomes infinite, so RKL-DPG cannot be used (nor measured) for that target. We use four words with different occurrence frequency: "amazing"($1\cdot 10^{-3}$), "restaurant" ($6\cdot 10^{-4}$), "amusing" ($6\cdot 10^{-5}$), and "Wikileaks" ($8\cdot 10^{-6}$).

Results The aggregated evolution of the metrics for both GDC and RL with KL penalties framework is presented in Fig. 3 (Fig. 1 shows a simplified view of Fig. 3 (a)). Disaggregated results for each task are presented on App. G. We see that all variants of f-DPG reduce the divergence from the target distribution across all measured f-divergences. Furthermore, as expected, convergence to the target is con-

nected with the success ratio in producing the desired word, $\mathbb{E}_{\pi_\theta}[b(x)]$, while balancing it with a moderate divergence from a, $\mathrm{KL}(\pi_\theta||a)$. This reflects that approaching the optimal distribution p translates into metrics in the downstream task. Strinklingly, the original KL-DPG is outperformed by all other variants of f-DPG, even in terms of forward KL. We hypothesize that this is linked to the high variance of the pseudo-rewards in KL-DPG, as visualized in the last panel of Fig. 3 (a) and (b). In Sec. 5, we suggest an interpretation for this. We also observe that RKL-DPG tends to produce distributions with lower normalized entropy. Despite this effect, we found no significant difference in diversity among the generated sentences (see Tab. 4 in App. D.1)

(a) lexical constraint with $p_{\mathrm{GDC\_bin}}$



(b) lexical constraint with p_{RLKL}

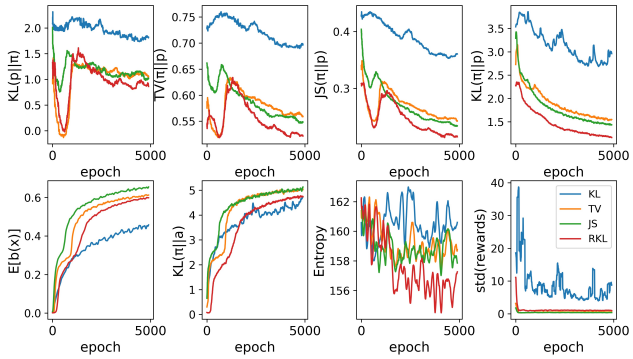


Figure 3. Comparison of f-DPG aggregated on four lexical constraints. Standard deviations are suppressed for clarity. Evaluation metrics: four f-divergences $D_f(\pi_\theta||p)$ (↓ better), alignment score $\mathbb{E}_{\pi_\theta}[b(x)]$ (↑ better), entropy (↑ better), standard deviation of pseudo-reward $\mathrm{std}(r_\theta(x))$.

4.3. Alignment with Distributional Constraints

Task We now investigate enforcing distributional preferences on the LM. We focus on debiasing the pretrained model on two kinds of preferences, namely genders' prevalence (Khalifa et al., 2021) and regard relative to religious groups. The preferences for the genders' debiasing task are defined as $\phi_1(x) = 1$ iff x contains more female than male pronouns, with desired moment $\bar{\mu}_1 = 0.5$ and $\phi_2(x) = 1$ iff x contains at least one of the words in the 'science' word list compiled by Dathathri et al. (2020), with desired moment $\bar{\mu}_2 = 1$. For regard debiasing, we use a single distributional constraint where $0 < \phi(x) < 1$ is a regard score of the sentence when prompted with `Muslims`, evaluated with a pretrained classifier (Sheng et al., 2019). We set the desired moment $\bar{\mu} = 0.568$, the regard score observed `Christians`. The initial average regard score given `Muslims` is 0.385. For the first experiment, we use GPT-2 small as the initial model a, additionally fine-tuned on the WikiBio dataset (Lebret et al., 2016), whereas for the last one we use vanilla GPT-2 small.

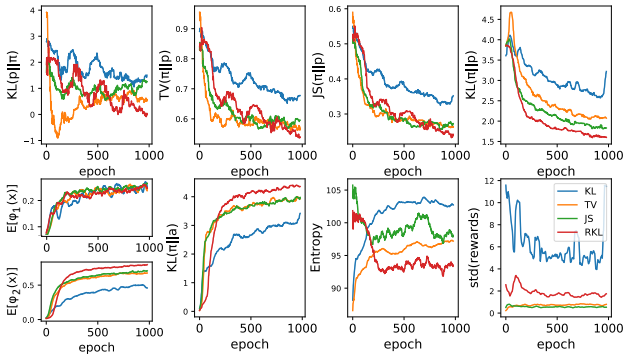
Results We report the results of both experiments on Fig. 4. For the regard score rebalancing, we considerably reduce bias in the regard score for two different demographic groups, from initial regard score ratio $\mathbb{E}[\phi(x)|\,\texttt{Christians}] : \mathbb{E}[\phi(x)|\,\texttt{Muslims}] = 1 : 0.677$ to $\mathbb{E}[\phi(x)|\,\texttt{Christians}] : \mathbb{E}[\phi(x)|\,\texttt{Muslims}] = 1 : 0.801$ on average. Interestingly, this task showcases a weakness of TV-DPG: Because the original distribution is already close to the target, the hard-thresholded pseudo-reward has a large variance (last panel of Fig 4(b)), inducing noisy gradient estimates and, consequently, sub-optimal convergence. Concerning the gender debiasing experiments, we can see that all other variants of f-DPG outperform the original KL-DPG explored in Khalifa et al. (2021), with RKL-DPG giving the best results and better matching the pointwise constraint although seemingly at the cost of lower diversity as measured by the entropy.

4.4. Alignment with Conditional Constraints

Task We adopt the conditional task from Korbak et al. (2022a), which aims to constrain the T5 (Raffel et al., 2020) language model to generate more factually faithful summaries (Maynez et al., 2020; Nan et al., 2021). Specifically, let $\mathrm{NER}(\cdot)$ denote the set of named entities found in a text. Then, $b(x, c) = 1$ iff $[\mathrm{NER}(x) \subseteq \mathrm{NER}(c)] \wedge [|\mathrm{NER}(x)| \geq 4]$, and 0 otherwise. Following the authors, we sample source documents from the the CNN/Daily Mail dataset (Nallapati et al., 2016), i.e. $\tau(c)$ is a uniform distribution over a given subset of source documents. In addition to the divergences, we evaluate the performance using Rouge (Lin, 2004), a measure of summarization quality in terms of unigram overlap between the source document and ground truth summary (See App. E for additional metrics and more experiments on code generation with compilability preferences).

Results We present the evolution of metrics in Fig. 5. The results show that f-DPG increases the fraction of consistent named entities in summarization, and interestingly, this

(a) distributional constraint for gender prevalence



(b) distributional constraint for regard rebalancing

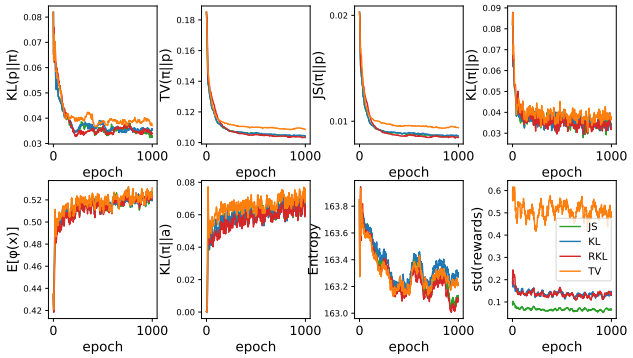


Figure 4. Comparison of f-DPG aggregated on distributional constraints. Evaluation metrics: four f-divergences $D_f(\pi_\theta||p)$ (↓ better), alignment score $\mathbb{E}_{\pi_\theta}[\phi(x)]$ (↑ better), entropy (↑ better), standard deviation of pseudo-reward $\mathrm{std}(r_\theta(x))$.

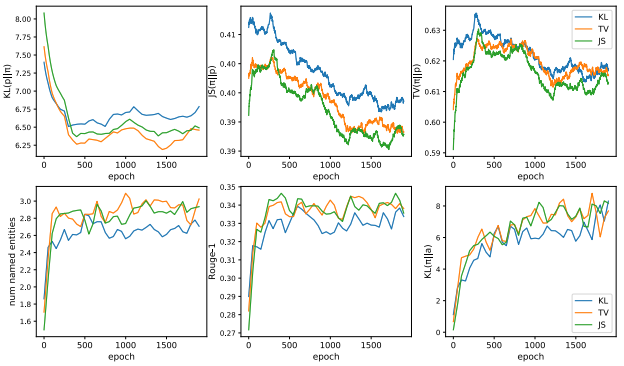


Figure 5. Comparison of f-DPG on factual summarization. Evaluation metrics: 3 f-divergences $D_f(\pi_\theta||p)$ (↓ better), number of named entities (↑ better), Rouge (↑ better).

also leads to indirect improvement in the overall quality of generated summaries compared to ground truth, even though ground truth summaries are not used in training. As also observed in Sec. 4.2, JS-DPG leads to better convergence to p than KL-DPG as used in Korbak et al. (2022a).

4.5. Scaling Trends of f-DPG

We conduct experiments to investigate the effect of model size on our approach using the scalar preference task described in Sec. 4.1. Specifically, we gradually increase the model size from GPT-2 "small" (117M parameters) to "xl" (1.5B parameters) while tracking two important metrics: alignment score, which is measured by the expected reward $\mathbb{E}_{\pi_\theta}[\phi(x)]$, and diversity, which is measured by the entropy. Figure 6 demonstrates that the alignment score steadily improves as the model size increases. However, we observe persistent differences between the divergence objectives for different f-DPGs, leaving the general order between f-DPGs intact with increasing model size (See Fig. 16 in App. G for evolution of metrics through training epochs). The scaling trend of LM alignment, characterized by a gradual and predictable increase without sudden shifts in performance, aligns with previous findings in the literature (Bai et al., 2022a). Nonetheless, our study further emphasizes the importance of proper divergence objectives, as increasing model size alone does not necessarily bridge the gap between optimal and suboptimal objectives. The smooth and gradual increase of the alignment score as a function of model size suggests that our findings will generalize to even larger LMs.

4.6. Ablation Study

This section presents just the key findings of our study. Full results and detailed discussions can be found in App. H.

Effect of parameter family capacity All experiments presented so far correspond to possibly mis-specified target distributions. To understand whether the observed behavior of different variants f-DPG is affected by this factor, we used pre-trained models with the same architecture as π_θ and p. We found that KL-DPG again lags considerably in terms of divergence, while presenting a high variance of in the pseudo-reward. RKL-DPG shows a significant drop of entropy in the initial phase, but with full capacity of parameter family, the model can recover, and cover the rest of the distribution. Additionally, applying zero-shot the fine-tuned LMs to a summarization task, following Radford et al. (2019), we found that the they recover to a large extent the quality of the target distribution.

Effect of training scheme We examined different training schemes for the lexical constraint on "amazing" from Sec. 4.2. We saw that the use of a baseline technique improves the performance of the f-DPG method, with RKL-DPG showing the greatest benefit. Additionally, we found that even though a large batch size is effective at reducing the variance of KL-DPG, we still observe KL-DPG to

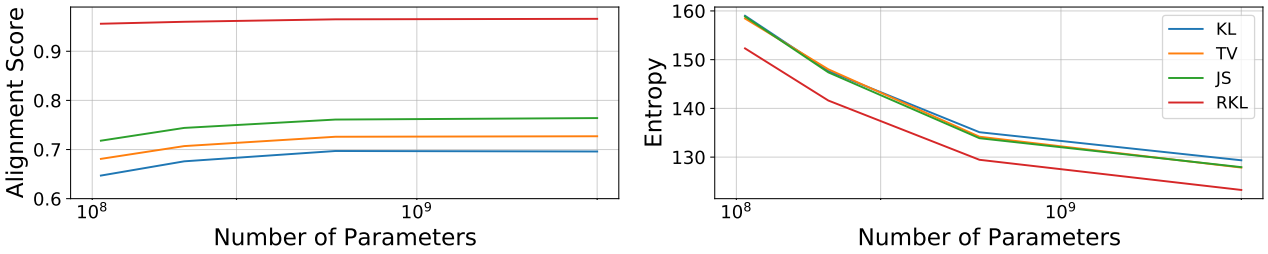


Figure 6. The scaling trend of f-DPG on sentiment preference. The x-axis denotes number of parameters of the LM π_θ and the y-axis denotes the alignment score and diversity measured by the expected reward $\mathbb{E}_{\pi_\theta}[\phi(x)]$ and by entropy, respectively.

perform comparatively worse than other divergences. Finally, we observe that our importance sampling estimates converged to the true value of Z.

5. Discussion and Conclusion

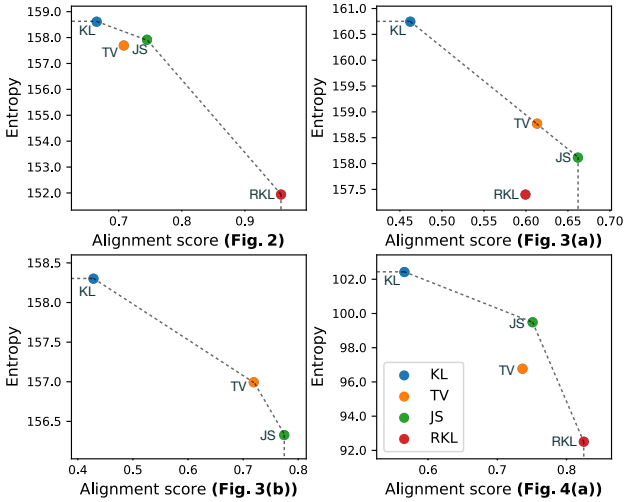


Figure 7. Pareto frontier of f-DPG for different alignment tasks; sentiment preference (Fig. 2), lexical constraints (Fig. 3(a), (b)), and distributional constraint for gender prevalence (Fig. 4(a))

A plausible hypothesis would have been that each variant of f-DPG is comparatively better at least in terms of the f-divergence objective being optimized. Surprisingly, we found that, save for a few exceptions (Sec. 4.1), for a given target there is one or a few variants that are the best across all measured divergences. Furthermore, we observed that divergence measures can have a significant impact on the performance of the model depending on the target distribution. Fig. 7 summarizes the Pareto frontier of the alignment-diversity trade-off of the f-DPG method. The results demonstrate that RKL-DPG and KL-DPG consistently represent two contrasting extremes: RKL-DPG shows high alignment but limited diversity, whereas KL-DPG exhibits low alignment but high diversity. JS-DPG shows a balanced trade-off between alignment and diversity and consistently appeared

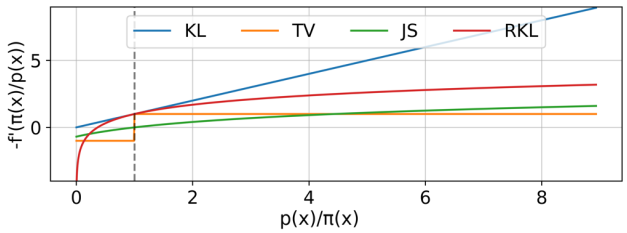


Figure 8. Pseudo-rewards for various f-divergences. The x-axis denotes $\frac{p(x)}{\pi_\theta(x)}$ and the y-axis denotes the pseudo-reward. The dotted line denotes the point where $p(x) = \pi_\theta(x)$.

on the Pareto frontier across all experiments we conducted.

Fig. 8 illustrates the differences between pseudo-rewards for distinct f-divergences, giving a plausible explanation for the observed differences. The forward KL loss aims to ensure coverage of the subset where $p(x) > 0$, giving a large pseudo-reward for samples with $p(x) >> \pi(x)$. However, the optimization can be sensitive to sampling noise in the finite sample approximation (see, e.g., Sec. 4.2). Conversely, the reverse KL loss results in extreme negative rewards for samples with $p(x) << \pi_\theta(x)$, leading π_θ to avoid such regions and resulting in distributional collapse (Sec. 4.1). Total Variation loss is robust to outliers thanks to its hard-thresholded pseudo-reward, however it can lead to high variance behavior when $\pi_\theta \approx p$ (Sec. 4.3). On the other hand, the Jensen-Shannon loss gives smooth and robust rewards in both directions and prevents π_θ from heavily relying on a single direction, making it a reasonable default choice as confirmed by our experiments.

To conclude, we propose a flexible framework for approximating a target distribution by minimizing any f-divergence, unifying earlier approaches for aligning language models. Our results on a diverse array of tasks show that minimizing well-chosen f-divergences leads to significant gains over previous work. The fact that increasing the model size improves the alignment score but does not inherently bridge the gap between objectives underscores the importance of selecting appropriate divergence objectives.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W. (eds.), *Proc. of ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 2017. URL http://proceedings.mlr.press/v70/arjovsky17a.html.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021. URL https://arxiv.org/abs/2112.00861.

Bahdanau, D., Brakel, P., Xu, K., Goyal, A., Lowe, R., Pineau, J., Courville, A. C., and Bengio, Y. An actor-critic algorithm for sequence prediction. In *Proc. of ICLR*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJDaqqveg.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022a. URL https://arxiv.org/abs/2204.05862.

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

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Berger, A. L., Della Pietra, S. A., and Della Pietra, V. J. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996. URL https://aclanthology.org/J96-1002.

Black, S., Gao, L., Wang, P., Leahy, C., and Biderman, S. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, 2021. URL https://doi.org/10.5281/zenodo.5297715.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Proc. of NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Cao, Y., Sotnikova, A., Daumé III, H., Rudinger, R., and Zou, L. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proc. of NAACL-HLT*, pp. 1276–1295, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.92. URL https://aclanthology.org/2022.naacl-main.92.

Che, T., Li, Y., Jacob, A. P., Bengio, Y., and Li, W. Mode regularized generative adversarial networks. In *Proc. of ICLR*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJKkY35le.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *ArXiv preprint*, abs/2210.11416, 2022. URL https://arxiv.org/abs/2210.11416.

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., and Liu, R. Plug and play language models: A simple approach to controlled text generation. In *Proc. of ICLR*. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1edEyBKDS.

Dohan, D., Xu, W., Lewkowycz, A., Austin, J., Bieber, D., Lopes, R. G., Wu, Y., Michalewski, H., Saurous, R. A., Sohl-Dickstein, J., et al. Language model cascades. *ArXiv preprint*, abs/2207.10342, 2022. URL https://arxiv.org/abs/2207.10342.

Eikema, B., Kruszewski, G., Dance, C. R., Elsahar, H., and Dymetman, M. An approximate sampler for energy-based models with divergence diagnostics. *Transactions of Machine Learning Research*, 2022. URL https://openreview.net/forum?id=VW4IrC0n0M.

Fan, A., Lewis, M., and Dauphin, Y. Hierarchical neural story generation. In *Proc. of ACL*, pp. 889–898, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*,

pp. 3356–3369, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301.

Ghasemipour, S. K. S., Zemel, R., and Gu, S. A divergence minimization perspective on imitation learning methods. In Kaelbling, L. P., Kragic, D., and Sugiura, K. (eds.), *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 1259–1277. PMLR, 2020. URL https://proceedings.mlr.press/v100/ghasemipour20a.html.

Glaese, A., McAleese, N., Trebacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., Campbell-Gillingham, L., Uesato, J., Huang, P., Comanescu, R., Yang, F., See, A., Dathathri, S., Greig, R., Chen, C., Fritz, D., Elias, J. S., Green, R., Mokrá, S., Fernando, N., Wu, B., Foley, R., Young, S., Gabriel, I., Isaac, W., Mellor, J., Hassabis, D., Kavukcuoglu, K., Hendricks, L. A., and Irving, G. Improving alignment of dialogue agents via targeted human judgements. *CoRR*, abs/2209.14375, 2022. doi: 10.48550/arXiv.2209.14375. URL https://doi.org/10.48550/arXiv.2209.14375.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Commun. ACM*, 63(11): 139–144, 2020. ISSN 0001-0782. doi: 10.1145/3422622. URL https://doi.org/10.1145/3422622.

Goyal, K., Dyer, C., and Berg-Kirkpatrick, T. Exposing the implicit energy networks behind masked language models via metropolis–hastings. In *Proc. of ICLR*. OpenReview.net, 2022. URL https://openreview.net/forum?id=6PvWo1kEvlT.

HF Canonical Model Maintainers. distilbert-base-uncased-finetuned-sst-2-english (revision bfdd146), 2022. URL https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english.

Hiriart-Urruty, J.-B. and Lemaréchal, C. *Convex analysis and minimization algorithms I: Fundamentals*, volume 305. Springer science & business media, 2013.

Huszar, F. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *ArXiv preprint*, abs/1511.05101, 2015. URL https://arxiv.org/abs/1511.05101.

Jaques, N., Gu, S., Bahdanau, D., Hernández-Lobato, J. M., Turner, R. E., and Eck, D. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. In Precup, D. and Teh, Y. W. (eds.), *Proc. of ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1645–1654. PMLR, 2017. URL http://proceedings.mlr.press/v70/jaques17a.html.

Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, À., Jones, N., Gu, S., and Picard, R. W. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *ArXiv preprint*, abs/1907.00456, 2019. URL https://arxiv.org/abs/1907.00456.

Kappen, H. J., Gómez, V., and Opper, M. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.

Kappen, H. J., Gómez, V., and Opper, M. Optimal control as a graphical model inference problem. In Borrajo, D., Kambhampati, S., Oddi, A., and Fratini, S. (eds.), *Proceedings of the Twenty-Third International Conference on Automated Planning and Scheduling, ICAPS 2013, Rome, Italy, June 10-14, 2013*. AAAI, 2013. URL http://www.aaai.org/ocs/index.php/ICAPS/ICAPS13/paper/view/6012.

Ke, L., Choudhury, S., Barnes, M., Sun, W., Lee, G., and Srinivasa, S. S. Imitation learning as f-divergence minimization. In LaValle, S. M., Lin, M., Ojala, T., Shell, D. A., and Yu, J. (eds.), *Algorithmic Foundations of Robotics XIV, Proceedings of the Fourteenth Workshop on the Algorithmic Foundations of Robotics, WAFR 2021, Oulu, Finland, June 21-23, 2021*, volume 17 of *Springer Proceedings in Advanced Robotics*, pp. 313–329. Springer, 2021. doi: 10.1007/978-3-030-66723-8_19. URL https://doi.org/10.1007/978-3-030-66723-8_19.

Khalifa, M., Elsahar, H., and Dymetman, M. A distributional approach to controlled text generation. In *Proc. of ICLR*. OpenReview.net, 2021. URL https://openreview.net/forum?id=jWkw45-9AbL.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *Proc. of ICLR*, 2015. URL http://arxiv.org/abs/1412.6980.

Korbak, T., Elsahar, H., Kruszewski, G., and Dymetman, M. Controlling conditional language models without catastrophic forgetting. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proc. of ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11499–11528. PMLR, 2022a. URL https://proceedings.mlr.press/v162/korbak22a.html.

Korbak, T., Elsahar, H., Kruszewski, G., and Dymetman, M. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Proc. of NeurIPS*, 2022b. URL https://openreview.net/forum?id=XvI6h-s4un.

Korbak, T., Perez, E., and Buckley, C. L. RL with KL penalties is better viewed as bayesian inference. *CoRR*, abs/2205.11275, 2022c. doi: 10.48550/arXiv.2205.11275. URL https://doi.org/10.48550/arXiv.2205.11275.

Lebret, R., Grangier, D., and Auli, M. Neural text generation from structured data with application to the biography domain. In *Proc. of EMNLP*, pp. 1203–1213, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1128. URL https://aclanthology.org/D16-1128.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Levine, S. Reinforcement learning and control as probabilistic inference: Tutorial and review. *ArXiv preprint*, abs/1805.00909, 2018. URL https://arxiv.org/abs/1805.00909.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. A diversity-promoting objective function for neural conversation models. In *Proc. of NAACL-HLT*, pp. 110–119, San Diego, California, 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014.

Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M., and Gao, J. Deep reinforcement learning for dialogue generation. In *Proc. of EMNLP*, pp. 1192–1202, Austin, Texas, 2016b. Association for Computational Linguistics. doi: 10.18653/v1/D16-1127. URL https://aclanthology.org/D16-1127.

Liese, F. and Vajda, I. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In *Proc. of ACL*, pp. 3214–3252, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proc. of EMNLP*, pp. 2122–2132, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1230. URL https://aclanthology.org/D16-1230.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proc. of ACL*, pp. 142–150, Portland, Oregon, USA, 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. On faithfulness and factuality in abstractive summarization. In *Proc. of ACL*, pp. 1906–1919, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173.

Menick, J., Trebacz, M., Mikulik, V., Aslanides, J., Song, F., Chadwick, M., Glaese, M., Young, S., Campbell-Gillingham, L., Irving, G., and McAleese, N. Teaching language models to support answers with verified quotes, 2022. URL https://arxiv.org/abs/2203.11147.

Mescheder, L. M., Geiger, A., and Nowozin, S. Which training methods for gans do actually converge? In Dy, J. G. and Krause, A. (eds.), *Proc. of ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3478–3487. PMLR, 2018. URL http://proceedings.mlr.press/v80/mescheder18a.html.

Miao, N., Zhou, H., Mou, L., Yan, R., and Li, L. CGMH: constrained sentence generation by metropolis-hastings sampling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 6834–6842. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33016834. URL https://doi.org/10.1609/aaai.v33i01.33016834.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. A. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013. URL http://arxiv.org/abs/1312.5602.

Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., and Xiang, B. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1028. URL https://aclanthology.org/K16-1028.

Nan, F., Nallapati, R., Wang, Z., Nogueira dos Santos, C., Zhu, H., Zhang, D., McKeown, K., and Xiang, B. Entity-level factual consistency of abstractive text summarization. In *Proc. of EACL*, pp. 2727–2733, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.235. URL https://aclanthology.org/2021.eacl-main.235.

Ngo, H., Raterink, C., Araújo, J. G., Zhang, I., Chen, C., Morisot, A., and Frosst, N. Mitigating harm in language models with conditional-likelihood filtration. *ArXiv preprint*, abs/2108.07790, 2021. URL https://arxiv.org/abs/2108.07790.

Norouzi, M., Bengio, S., Chen, Z., Jaitly, N., Schuster, M., Wu, Y., and Schuurmans, D. Reward augmented maximum likelihood for neural structured prediction. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Proc. of NeurIPS*, pp. 1723–1731, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/2f885d0fbe2e131bfc9d98363e55d1d4-Abstract.html.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Proc. of NeurIPS*, pp. 271–279, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/cedebb6e872f539bef8c3f919874e9d7-Abstract.html.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Gray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Proc. of NeurIPS*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Parshakova, T., Andreoli, J.-M., and Dymetman, M. Distributional reinforcement learning for energy-based sequential models. *ArXiv preprint*, abs/1912.08517, 2019. URL https://arxiv.org/abs/1912.08517.

Pasunuru, R. and Bansal, M. Reinforced video captioning with entailment rewards. In *Proc. of EMNLP*, pp. 979–985, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1103. URL https://aclanthology.org/D17-1103.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Proc. of NeurIPS*, pp. 8024–8035, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Paulus, R., Xiong, C., and Socher, R. A deep reinforced model for abstractive summarization. In *Proc. of ICLR*. OpenReview.net, 2018. URL https://openreview.net/forum?id=HkACIQgA-.

Peters, J. and Schaal, S. Reinforcement learning by reward-weighted regression for operational space control. In Ghahramani, Z. (ed.), *Proc. of ICML*, volume 227 of *ACM International Conference Proceeding Series*, pp. 745–750. ACM, 2007. doi: 10.1145/1273496.1273590. URL https://doi.org/10.1145/1273496.1273590.

Polyanskiy, Y. f-divergences, 2019. URL https://people.lids.mit.edu/yp/homepage/data/LN_fdiv.pdf.

Qin, L., Welleck, S., Khashabi, D., and Choi, Y. Cold decoding: Energy-based constrained text generation with langevin dynamics. *ArXiv preprint*, abs/2202.11705, 2022. URL https://arxiv.org/abs/2202.11705.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. Sequence level training with recurrent neural networks. In Bengio, Y. and LeCun, Y. (eds.), *Proc. of ICLR*, 2016. URL http://arxiv.org/abs/1511.06732.

Raychev, V., Bielik, P., and Vechev, M. Probabilistic model for code with decision trees. *ACM SIGPLAN Notices*, 51 (10):731–747, 2016.

Rockafellar, R. T. *Convex analysis*, volume 18. Princeton university press, 1970.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., and Weston, J. Recipes for building an open-domain chatbot. In *Proc. of EACL*, pp. 300–325, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.24. URL https://aclanthology.org/2021.eacl-main.24.

Sason, I. On f-divergences: Integral representations, local behavior, and inequalities. *Entropy*, 20(5):383, 2018.

Sason, I. and Verdú, S. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., and Leike, J. Self-critiquing models for assisting human evaluators. *ArXiv preprint*, abs/2206.05802, 2022. URL https://arxiv.org/abs/2206.05802.

Scheurer, J., Campos, J. A., Chan, J. S., Chen, A., Cho, K., and Perez, E. Training language models with natural language feedback. *ArXiv preprint*, abs/2204.14146, 2022. URL https://arxiv.org/abs/2204.14146.

Schulman, J., Moritz, P., Levine, S., Jordan, M. I., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. In Bengio, Y. and LeCun, Y. (eds.), *Proc. of ICLR*, 2016. URL http://arxiv.org/abs/1506.02438.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *ArXiv preprint*, abs/1707.06347, 2017. URL https://arxiv.org/abs/1707.06347.

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. The woman worked as a babysitter: On biases in language generation. In *Proc. of EMNLP*, pp. 3407–3412, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1339. URL https://aclanthology.org/D19-1339.

Solaiman, I. and Dennison, C. Process for adapting language models to society (palms) with values-targeted datasets. *Proc. of NeurIPS*, 34:5861–5873, 2021.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Rahane, A., Iyer, A. S., Andreassen, A., Santilli, A., Stuhlmüller, A., Dai, A. M., La, A., Lampinen, A. K., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakas, A., and et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615, 2022. doi: 10.48550/arXiv.2206.04615. URL https://doi.org/10.48550/arXiv.2206.04615.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize from human feedback. In *Proc. of NeurIPS*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Tambwekar, P., Dhuliawala, M., Martin, L. J., Mehta, A., Harrison, B., and Riedl, M. O. Controllable neural story plot generation via reward shaping. In Kraus, S. (ed.), *Proc. of IJCAI*, pp. 5982–5988. ijcai.org, 2019. doi: 10.24963/ijcai.2019/829. URL https://doi.org/10.24963/ijcai.2019/829.

Theis, L., van den Oord, A., and Bethge, M. A note on the evaluation of generative models. In Bengio, Y. and LeCun, Y. (eds.), *Proc. of ICLR*, 2016. URL http://arxiv.org/abs/1511.01844.

Thoppilan, R., Freitas, D. D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhou, Y., Chang, C., Krivokon, I., Rusch, W., Pickett, M., Meier-Hellstern, K. S., Morris, M. R., Doshi, T., Santos, R. D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E. H., and Le, Q. Lamda: Language models for dialog applications. *ArXiv preprint*, abs/2201.08239, 2022. URL https://arxiv.org/abs/2201.08239.

Todorov, E. Linearly-solvable markov decision problems. In Schölkopf, B., Platt, J. C., and Hofmann, T. (eds.), *Proc. of NeurIPS*, pp. 1369–1376. MIT Press, 2006a. URL https://proceedings.neurips.cc/paper/2006/hash/

d806ca13ca3449af72a1ea5aedbed26a-Abstract.html.

Todorov, E. Linearly-solvable markov decision problems. In Schölkopf, B., Platt, J. C., and Hofmann, T. (eds.), *Proc. of NeurIPS*, pp. 1369–1376. MIT Press, 2006b. URL https://proceedings.neurips.cc/paper/2006/hash/d806ca13ca3449af72a1ea5aedbed26a-Abstract.html.

Wang, D., Liu, H., and Liu, Q. Variational inference with tail-adaptive f-divergence. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Proc. of NeurIPS*, pp. 5742–5752, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/1cd138d0499a68f4bb72bee04bbec2d7-Abstract.html.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W. S., Legassick, S., Irving, G., and Gabriel, I. Ethical and social risks of harm from language models. *ArXiv preprint*, abs/2112.04359, 2021. URL https://arxiv.org/abs/2112.04359.

Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., Anderson, K., Kohli, P., Coppin, B., and Huang, P.-S. Challenges in detoxifying language models. In *Findings of EMNLP*, pp. 2447–2469, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.210. URL https://aclanthology.org/2021.findings-emnlp.210.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. Transformers: State-of-the-art natural language processing. In *Proc. of EMNLP*, pp. 38–45, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Xu, J., Ju, D., Li, M., Boureau, Y.-L., Weston, J., and Dinan, E. Bot-adversarial dialogue for safe conversational agents. In *Proc. of NAACL-HLT*, pp. 2950–2968, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.235. URL https://aclanthology.org/2021.naacl-main.235.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. STar: Bootstrapping reasoning with reasoning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Proc. of NeurIPS*, 2022. URL https://openreview.net/forum?id=_3ELRdg2sgI.

Zhao, J., Khashabi, D., Khot, T., Sabharwal, A., and Chang, K.-W. Ethical-advice taker: Do language models understand natural language interventions? In *Findings of ACL*, pp. 4158–4164, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.364. URL https://aclanthology.org/2021.findings-acl.364.

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., and Yu, Y. Texygen: A benchmarking platform for text generation models. In Collins-Thompson, K., Mei, Q., Davison, B. D., Liu, Y., and Yilmaz, E. (eds.), *Proc. of SIGIR*, pp. 1097–1100. ACM, 2018. doi: 10.1145/3209978.3210080. URL https://doi.org/10.1145/3209978.3210080.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *ArXiv preprint*, abs/1909.08593, 2019. URL https://arxiv.org/abs/1909.08593.

Ziegler, D. M., Nix, S., Chan, L., Bauman, T., Schmidt-Nielsen, P., Lin, T., Scherlis, A., Nabeshima, N., Weinstein-Raun, B., de Haas, D., Shlegeris, B., and Thomas, N. Adversarial training for high-stakes reliability. *CoRR*, abs/2205.01663, 2022. doi: 10.48550/arXiv.2205.01663. URL https://doi.org/10.48550/arXiv.2205.01663.

A. Complements on Formal Aspects and Proofs

A.1. Equivalent definitions for f-divergences

The definition of f-divergences of Eq. 6 is equivalent to a second definition, in a more "symmetrical" format, following (Liese & Vajda, 2006), which will help in some derivations, in particular in the proof of Theorem 1.

Definition (f-divergence: "symmetrical" format)**.** *The f-divergence $D_f(p||q)$, where p and q are distributions over a discrete set $\mathcal{X}$ can be defined as*

$$D_f(p||q) \doteq \sum_{\{x:\ p(x)>0,\ q(x)>0\}} q(x)\ f(\frac{p(x)}{q(x)}) \quad + f(0)\ q(p=0) \quad + f^*(0)\ p(q=0), \tag{10}$$

where the generator function $f : (0, \infty) \to \mathbb{R}$ is a convex function satisfying $f(1) = 0$. We denote by $q(p = 0)$ the q-mass of the set $\{x : p(x) = 0\}$, i.e. $q(p = 0) = \sum_{\{x:p(x)=0\}} q(x)$ and similarly for $p(q = 0)$.

In this definition, the function $f^*(t)$ is the so-called *perspective transform* of f defined by $f^*(t) = t\ f(\frac{1}{t})$. It can be shown to be also a convex function $f^* : (0, \infty) \to \mathbb{R}$ with $f^*(1) = 0$ and $f^{**} = f$. As we mentioned in the main text, we also have the following important "swapping" property: $D_f(p, q) = D_{f^*}(q, p)$.

Following Liese & Vajda (2006); Polyanskiy (2019), we use the conventions:

$$f(0) \doteq \lim_{t\to 0} f(t), \quad f^*(0) = \lim_{t\to 0} f^*(t) = \lim_{t\to 0} t\ f(\frac{1}{t}), \tag{11}$$

$$0\ f(0) \doteq 0, \quad 0\ f^*(0) \doteq 0, \qquad \text{including when } f(0) = \infty \text{ and } f^*(0) = \infty, \tag{12}$$

$$f'(\infty) \doteq f^*(0) = \lim_{t\to 0} t\ f(\frac{1}{t}). \tag{13}$$

For the existence of the limits in these equations, where $f(0)$ and $f^*(0)$ can take values in $\mathbb{R} \cup \{\infty\}$, as well as for the motivation for defining $f'(\infty) \doteq \lim_{t\to 0} t\ f(\frac{1}{t})$, one may refer to (Liese & Vajda, 2006) and (Hiriart-Urruty & Lemaréchal, 2013, §2.3).

Equivalence of definitions 6 and 10 In order to prove this equivalence, after noting that $f'(\infty) = f^*(0)$, it remains to show that $\mathbb{E}_{x\sim q} f\left(\frac{p(x)}{q(x)}\right)$ is equal to $\sum_{\{x:\ p(x)>0,\ q(x)>0\}} q(x)\ f(\frac{p(x)}{q(x)}) + f(0)\ q(p=0)$. We have:

$$\begin{aligned}
\mathbb{E}_{x\sim q}\ f(\frac{p(x)}{q(x)}) &= \sum_{\{x:\ q(x)>0\}} q(x)\ f(\frac{p(x)}{q(x)}) \\
&= \sum_{\{x:\ q(x)>0,\ p(x)>0\}} q(x)\ f(\frac{p(x)}{q(x)}) \quad + \sum_{\{x:\ q(x)>0,\ p(x)=0\}} q(x)\ f(0) \\
&= \sum_{\{x:\ q(x)>0,\ p(x)>0\}} q(x)\ f(\frac{p(x)}{q(x)}) \quad + f(0)\ q(p=0),
\end{aligned}$$

which concludes the proof.

A.2. Illustrations of a few f-divergences

Let's now see how the notion of f-divergence can be applied to a few common cases.

Forward and reverse KL By the standard definition for KL divergence, we have, for $\mathrm{KL}(p||\pi)$, the "forward KL" from a model π to a target p:

$$\mathrm{KL}(p||\pi) = \begin{cases} \mathbb{E}_{x\sim p}\ \log \frac{p(x)}{\pi(x)} & \text{if Supp}(p) \subset \text{Supp}(\pi), \\ \infty, & \text{otherwise.} \end{cases} \tag{14}$$

If we take $f(t) = -\log t$, as in Table 1, then we have $f(0) = \infty$. On the other hand we see that $f^*(t) = t\ \log t$ and $f^*(0) = 0$. We can then write, using (10):

$$\begin{aligned} D_f(\pi||p) &= \sum_{\{x:\ \pi(x)>0,\ p(x)>0\}} -p(x)\ \log(\frac{\pi(x)}{p(x)}) \quad + \infty\ p(\pi=0) + 0\ \pi(p=0) \\ &= \sum_{\{x:\ \pi(x)>0,\ p(x)>0\}} p(x)\ \log(\frac{p(x)}{\pi(x)}) \quad + \infty\ p(\pi=0), \end{aligned}$$

where $\infty\ p(\pi = 0)$ is null for $\text{Supp}(p) \subset \text{Supp}(\pi)$ and infinite otherwise. Hence $D_f(\pi||p) = \text{KL}(p||\pi)$, the forward KL from π to p.

Now, consider the "reverse KL" from π to p, namely $\text{KL}(\pi||p)$. Based on the previous derivation, and with the same $f(t) = -\log t$ we can write it as $\text{KL}(\pi||p) = D_f(p||\pi)$, but using the perspective function $f^*(t) = t \log t$, we can also write it (as we actually do in Table 1) as $D_{f^*}(\pi||p) = D_{t\ \log t}(\pi||p)$.

Total Variation divergence The Total Variation divergence between p and π is standardly defined as $\text{TV}(p||\pi) = \frac{1}{2}\ \sum_{x\in\mathcal{X}} |p(x) - \pi(x)|$. We then have $\text{TV}(p||\pi) = \text{TV}(\pi||p)$. Let's then define $f(t) = \frac{1}{2}|1 - t|$. We have $f(0) = 1/2$, $f^*(t) = f(t)$, and $f^*(0) = 1/2$. Then, using (10):

$$\begin{aligned} D_f(\pi||p) &= \sum_{\{x:\ \pi(x)>0,\ p(x)>0\}} \frac{1}{2}\ p(x)\ \left|1 - \frac{\pi(x)}{p(x)}\right| \quad + \frac{1}{2}\ p(\pi=0) + \frac{1}{2}\ \pi(p=0) \\ &= \sum_{\{x:\ \pi(x)>0,\ p(x)>0\}} \frac{1}{2}\ |p(x) - \pi(x)| \quad + \frac{1}{2}\ p(\pi=0) + \frac{1}{2}\ \pi(p=0) \\ &= \sum_{\{x:\ \pi(x)>0,\ p(x)>0\}} \frac{1}{2}\ |p(x) - \pi(x)| \quad + \frac{1}{2} \sum_{\{x:\ \pi(x)=0,\ p(x)>0\}} |p(x) - \pi(x)| \\ &\qquad\qquad + \frac{1}{2} \sum_{\{x:\ \pi(x)>0,\ p(x)=0\}} |p(x) - \pi(x)| \\ &= \frac{1}{2} \sum_{x\in\mathcal{X}} |p(x) - \pi(x)|\,, \end{aligned}$$

and therefore $\text{TV}(p||\pi) = D_f(\pi||p)$, and also $\text{TV}(p||\pi) = \text{TV}(\pi||p) = D_{f^*}(p||\pi) = D_f(p||\pi)$.

A.3. Proof of Theorem 1

We restate the theorem here for convenience.

Theorem (Theorem 1)**.** *Let p and π_θ be distributions over a discrete set $\mathcal{X}$ such that at least one of the following conditions holds: (i) $\forall\theta \in \Theta$, Supp$(p) \subset$ Supp(π_θ), or (ii) Supp(π_θ) does not depend on θ. Then:*

$$\nabla_\theta D_f(\pi_\theta||p) = \mathbb{E}_{x\sim\pi_\theta}\left[f'\left(\frac{\pi_\theta(x)}{p(x)}\right)\nabla_\theta \log \pi_\theta(x)\right]. \tag{15}$$

Proof. Based on definition (10) we have:

$$\begin{aligned}
\nabla_\theta D_f(\pi_\theta||p) &= \sum_{\{x:p(x)>0,\pi_\theta(x)>0\}} p(x)\ \nabla_\theta f(\frac{\pi_\theta(x)}{p(x)}) + f'(\infty)\nabla_\theta\pi_\theta(p=0) + f(0)\nabla_\theta p(\pi_\theta=0) \\
&= \sum_{\{x:p(x)>0,\pi_\theta(x)>0\}} p(x)\ f'(\frac{\pi_\theta(x)}{p(x)})\ \nabla_\theta \frac{\pi_\theta(x)}{p(x)} + f'(\infty)\nabla_\theta\pi_\theta(p=0) \\
&= \sum_{\{x:p(x)>0,\pi_\theta(x)>0\}} \pi_\theta(x) f'(\frac{\pi_\theta(x)}{p(x)})\nabla_\theta \log \pi_\theta(x) + f'(\infty)\nabla_\theta\pi_\theta(p=0) \\
&= \sum_{\{x:p(x)>0,\pi_\theta(x)>0\}} \pi_\theta(x) f'(\frac{\pi_\theta(x)}{p(x)})\nabla_\theta \log \pi_\theta(x) + f'(\infty)\nabla_\theta \left[\sum_{\{x:p(x)=0,\pi_\theta(x)>0\}} \pi_\theta(x)\right] \\
&= \sum_{\{x:p(x)>0,\pi_\theta(x)>0\}} \pi_\theta(x) f'(\frac{\pi_\theta(x)}{p(x)})\nabla_\theta \log \pi_\theta(x) + \sum_{\{x:p(x)=0,\pi_\theta(x)>0\}} \pi_\theta(x) f'(\infty)\nabla_\theta \log \pi_\theta(x) \\
&= \sum_{\{x:\pi_\theta(x)>0\}} \pi_\theta(x) f'(\frac{\pi_\theta(x)}{p(x)})\nabla_\theta \log \pi_\theta(x) \\
&= \mathbb{E}_{x\sim\pi_\theta}\ f'(\frac{\pi_\theta(x)}{p(x)})\ \nabla_\theta \log \pi_\theta(x).
\end{aligned}$$

In the first line of this derivation, we use the previously introduced notation $f'(\infty) \doteq f^*(0)$, employed in particular by (Polyanskiy, 2019), which is motivated by the fact that $\lim_{t\to\infty} f^{'}(t) = \lim_{t\to\infty} \frac{1}{t} f(t) = f^*(0)$ (See (Hiriart-Urruty & Lemaréchal, 2013)). In the second line, we employ a variant of the chain-rule for derivatives of multivariate functions. We also exploit the fact that the condition (i) stating that the support of p is contained in the support of π_θ for all $\theta \in \Theta$ implies that $\nabla_\theta p(\pi_\theta = 0) = \nabla_\theta 0 = 0$, and that the condition (ii) that the support of π_θ does not depend on θ also implies that $\nabla_\theta p(\pi_\theta = 0) = 0$. In the fourth line, we write $\pi_\theta(p = 0)$ as a sum. In the sixth line, we allow the notation $f'(\frac{\pi_\theta(x)}{p(x)})$ instead of $f'(\infty)$ when $p(x) = 0$ and $\pi_\theta(x) > 0$. □

Working with the opposite divergence $D_f(p||\pi_\theta)$ In case one may prefer to work with a divergence $D_f(p||\pi_\theta)$ having the opposite argument order, then one can use the identity $D_f(p||\pi_\theta) = D_{f^*}(\pi_\theta||p)$ to conclude that under the exact same conditions (i) or (ii) as previously, we have:

$$\nabla_\theta D_f(p||\pi_\theta) = \nabla_\theta D_{f^*}(\pi_\theta||p) = \mathbb{E}_{x\sim\pi_\theta}\left[f^{*'}\left(\frac{\pi_\theta(x)}{p(x)}\right)\ \nabla_\theta \log \pi_\theta(x)\right],$$

where the derivative is applied to the perspective transform of f.

A.4. About non-differentiability of f

In practice when sampling from π_θ in Eq (8), the problem of non-differentiability can be neglected, and recourse to subgradients is typically unnecessary, even for f's that have non-differentiability points (such as e.g. the generator $f(t) = 0.5|1-t|$ for the Total Variation divergence). Indeed, let $T_{nd} \doteq \{t : f(t) \text{ is non differentiable at } t\}$, and let $\Theta_{nd} \doteq \{\theta : \exists x \in \mathcal{X} : \frac{\pi_\theta(x)}{p(x)} \in T_{nd}\}$ be the set of θ's for which $f^{'}\left(\frac{\pi_\theta(x)}{p(x)}\right)$ is undefined on at least one x. Then $\Theta_{nd} \subset \mathbb{R}^d$ (with d the parameter dimension) is the countable union of countable sets, hence is countable, and therefore of null measure inside $\mathbb{R}^d$. This means that, almost surely over θ, the RHS of Eq 8 is well-defined for all x's.

A.5. f-DPG algorithm

A.6. Baseline: alternative derivation

The generator function is not uniquely determined for a given f-divergence:

Fact 3. *For generators f, g such that $f(t) = g(t) + c(t-1)$, $c \in \mathbb{R}$, $D_f(p_1||p_2) = D_g(p_1||p_2)$.*

Algorithm 1 f-DPG

Input: unnormalized target distribution $P(\cdot)$, initial model $a(\cdot)$, D_f generator $f(\cdot)$
Initialize: $\pi_\theta(\cdot) \leftarrow a(\cdot)$, $Z \leftarrow 0$, $N \leftarrow 0$ {initialize model π_θ, partition Z, sample size N for moving average}
for each iteration **do**
  for each episode **do**
    sample x from $\pi_\theta(\cdot)$
    $N \leftarrow N + 1$
    $Z \leftarrow \frac{(N-1)Z + (P(x)/\pi_\theta(x))}{N}$ {Estimate Z with historical samples, using a moving average}
    $p(\cdot) \leftarrow P(\cdot)/Z$
    $\theta \leftarrow \theta + \alpha^{(\theta)} f^{'}\left(\frac{\pi_\theta(x)}{p(x)}\right) \nabla_\theta \log \pi_\theta(x)$ {Update π_θ according to Thm. 1}
  end for
end for
Output: π_θ

We provide here an alternative way to introducing baselines, based on a change of generator.

Theorem (Baseline based on change of generator)**.** *If $D_f(\pi_\theta||p)$ is a divergence with any generator f, and $B \in \mathbb{R}$, there exists a generator g with the same divergence $D_f(\pi_\theta||p) = D_g(\pi_\theta||p)$ such that*

$$\begin{aligned}\nabla_\theta D_g(\pi_\theta||p) &= \mathbb{E}_{x\sim\pi_\theta}\left[\left(f^{'}\left(\frac{\pi_\theta(x)}{p(x)}\right) - B\right)\nabla_\theta \log \pi_\theta(x)\right] \\ &= \nabla_\theta D_f(\pi_\theta||p).\end{aligned}$$

Proof. Recall that $D_f(\pi_\theta||p) = D_g(\pi_\theta||p)$ when $g(x) = f(x) - B(x-1)$. Therefore, $\nabla_\theta D_f(\pi_\theta||p) = \nabla_\theta D_g(\pi_\theta||p)$ with $g^{'}\left(\frac{\pi_\theta(x)}{p(x)}\right) = f^{'}\left(\frac{\pi_\theta(x)}{p(x)}\right) - B$. □

B. Extended Related Work

RL for LMs There is a large reinforcement learning inspired literature about steering an autoregressive sequential model towards optimizing some global reward over the generated text. This includes REINFORCE (Williams, 1992) for Machine Translation (Ranzato et al., 2016), actor critic for Abstractive Summarization (Paulus et al., 2018), Image-to-Text (Liu et al., 2016), Dialogue Generation (Li et al., 2016b), and Video Captioning (Pasunuru & Bansal, 2017). With respect to rewards, some approaches for Machine Translation and Summarization (Ranzato et al., 2016; Bahdanau et al., 2017) directly optimize end task rewards such as BLEU and ROUGE at training time to compensate for the mismatch between the perplexity-based training of the initial model and the evaluation metrics used at test time. Some others use heuristic rewards as in (Li et al., 2016b; Tambwekar et al., 2019), in order to improve certain a priori desirable features of generated stories or dialogues.

Several studies, have considered incorporating a distributional term inside the reward to be maximized. In particular Jaques et al. (2017; 2019); Ziegler et al. (2019); Stiennon et al. (2020) have applied variations of KL-control (Todorov, 2006b; Kappen et al., 2013) which adds a penalty term to the reward term so that the resulting policy does not deviate too much from the original one in terms of KL-divergence. The overall objective with the KL-penalty is maximized using an RL algorithm of choice including: PPO (Schulman et al., 2017) as in Ziegler et al. (2019) or Q-learning (Mnih et al., 2013) as in Jaques et al. (2017). This approach recently get a huge attention with its impact with using the human data to train aligned language models in LaMDA (Thoppilan et al., 2022), InstructGPT (Ouyang et al., 2022), Sparrow (Glaese et al., 2022), and CAI (Bai et al., 2022b). Similar work involving model self-critique and natural language feedback includes (Zhao et al., 2021; Scheurer et al., 2022; Saunders et al., 2022)

f-divergence objectives for generative models In the literature, there have been several studies exploring the use of f-divergences in generative models. Goodfellow et al. (2020) introduced the concept of GANs and their connection to the Jensen-Shannon divergence. Nowozin et al. (2016) proposed a variational expression of f-divergences as a loss function for GANs. Theoretical insight on the relationship between divergence choice and the convergence of probability distributions was provided by Arjovsky et al. (2017). Additionally, Theis et al. (2016) discussed potential drawbacks of forward KL

Experiment	Hyperparameters
Common	batch size = 258, optimizer = Adam, learning rate schedule = constant with warmup (100 epochs)
Sentiment preference	original model = gpt2, learning rate = 1×10^{-5} maximum length = 40, batch size = 2048, total epochs=1000
Lexical(RLKL)	original model = gpt2, learning rate = 1×10^{-5}, maximum length = 40, total epochs=5000
Lexical(GDC)	original model = gpt2, learning rate = 1.41×10^{-5}, maximum length = 40, total epochs=5000
Female50% Science100%	original model = mkhalifa/gpt2-biographies, learning rate = 1.41×10^{-5}, maximum length = 40, total epochs=1000
Regard balancing	original model = gpt2, learning rate = 5×10^{-6}, maximum length = 40, batch size = 2048, total epochs=1000
Summarization	original mode=t5-small, learning rate = 1×10^{-4}, maximum length = 128, total epochs=2000
Code generation	original mode=gpt-neo-125M, learning rate = 1×10^{-4}, maximum length = 128, total epochs=2000
GPT2 approximation	original model = lvwerra/gpt2-imdb, learning rate = 5×10^{-6} maximum length = 40, total epochs=8000

Table 2. Hyperparameters used throughout all experiments

divergence in generative models and Huszar (2015) proposed a generalization of Jensen-Shannon divergence that interpolates between KL and reverse KL and has Jensen-Shannon as its midpoint.

The connections between RL and divergence minimization have also been explored, with studies showing that entropy regularization in RL can be viewed as minimizing reverse KL divergence between reward-weighted trajectory and policy trajectory distributions (Kappen et al., 2013; Levine, 2018). Other studies have also explored the use of forward KL divergence in RL (Peters & Schaal, 2007; Norouzi et al., 2016). Additionally, a unified probabilistic perspective on f-divergence minimization in imitation learning has been presented for both discrete and continuous control environments (Ke et al., 2021; Ghasemipour et al., 2020).

Wang et al. (2018) introduced variational inference with adaptive f-divergences and demonstrated its effectiveness in RL, with focus on continuous sample spaces. Their Proposition 4.2.1 is similar to our theorem 1. However, our result exhibits greater generality by defining $D_f(\pi_\theta||p)$ without requirements of absolute continuity in either direction (Polyanskiy, 2019; Liese & Vajda, 2006). We note that this generalization is crucial for LM alignment, as the case of $p(x) = 0,\ \pi_\theta(x) > 0$ can easily occur.

C. Implementation Details

All models were implemented using PyTorch (Paszke et al., 2019) and HuggingFace Transformers (Wolf et al., 2020) with the Adam optimizer (Kingma & Ba, 2015). Training was performed on Nvidia V100 GPU, with the longest run taking approximately 2 days. Hyperparameter details are listed in Tab. 2. Pretrained models are available on the Huggingface Model Hub under the specified model names. We focused on searching for hyperparameters based on KL-DPG, which served as the baseline method we aimed to improve upon, providing it with an initial advantage. To ensure that all methods were evaluated under comparable settings, we tuned the hyperparameters once for all f-DPG methods.

D. Additional Experiments

D.1. Generation Quality

Metrics To see if different objective affects the quality of the generated sentences, we report the following metrics on experiment in Sec. 4.1, Sec. 4.2.

Loss	Entropy	Self-BLEU-5	Dist-1	Perplexity
KL	159.09 (9.58)	0.62 (0.01)	0.88 (0.01)	58.87 (7.48)
TV	157.60 (8.91)	0.65 (0.01)	0.88 (0.01)	59.48 (5.25)
JS	158.04 (8.62)	0.64 (0.01)	0.88 (0.01)	59.67 (6.23)
RKL	151.04 (7.99)	0.70 (0.01)	0.87 (0.01)	53.15 (4.14)

Table 3. Quality of the generated text metrics for the experiment on scalar preferences (Sec. 4.1). entropy (↑ better), Self-BLEU-5 (↓ better), Distinct-1 (↑ better), and Perplexity (↓ better).

1. Distinct-n (Li et al., 2016a), a measure of text diversity in terms of the frequency of repeated n-grams within a single sample x.
2. Self-BLEU-n (Zhu et al., 2018), a measure of text diversity on a distributional level across samples.
3. Perplexity, a measure of text fluency with exponentiation of the negative average per-token log-probability under a language model. We use a separate model Distil-GPT-2 (Wolf et al., 2020) to calculate perplexity to avoid inflated estimates (Liu et al., 2016).

Results Tab. 3 provides additional metrics for the generated sentences and their diversity on scalar preferences. The notably low entropy and high Self-BLEU of RKL-DPG again indicate low diversity of RKL-DPG at the distributional level, whereas other f-DPGs have similar values to each other. On the other hand, in quality for individual samples as measured by the perplexity metric, RKL-DPG shows better quality, which suggests that RKL-DPG captures a subset of the target distribution, an observation that is frequently discussed in other generative models (Huszar, 2015; Che et al., 2017; Mescheder et al., 2018). We provide metrics for the generated sentences aggregated on lexical constraint in Tab. 4. We found no significant difference in diversity among the generated sentences.

Loss	$\mathbb{E}\left[b(x)\right]$	Self-BLEU-5	Dist-1	Perplexity
KL	0.45 (0.09)	0.66 (0.02)	0.96 (0.00)	90.59 (11.74)
TV	0.60 (0.12)	0.67 (0.01)	0.96 (0.01)	80.52 (8.79)
JS	0.66 (0.14)	0.67 (0.01)	0.95 (0.01)	79.53 (8.80)
RKL	0.60 (0.20)	0.66 (0.02)	0.95 (0.01)	79.49 (7.79)

Table 4. Quality of the generated text metrics for the experiment on lexical constraint (Sec. 4.2). $\mathbb{E}_{\pi_\theta}\left[b(x)\right]$ (↑ better), Self-BLEU-5 (↓ better), Distinct-1 (↑ better), and Perplexity (↓ better).

E. f-DPG on Conditional Target Distributions

Let C be a discrete (potentially infinite) set of conditions c. The problem of fine-tuning a pretrained model $a(x|c)$ to satisfy a control objective (e.g. generating factually correct summaries) can be seen as a constraint satisfaction problem: finding a model $p_c(x)$ that meets the demands of the control objective but at the same time stays as close as possible to the original pretrained model $a(x|c)$. A control objective can be defined in terms of a binary scorer $b(x, c)$ such that $b(x, c) = 1$ if a sample (c, x) satisfies a constraint given by a control objective (e.g. x is factually correct with respect to c) and $b(x, c) = 0$ otherwise.

For each $c \in C$, we can frame the problem of finding the unique model $p_c(x)$ such that (i) $b(x, c) = 1$ for all samples $x \sim p_c(x)$, and (ii) $p_c(\cdot)$ has minimal KL divergence from $a(\cdot|c)$ as an instance of the unconditional case already considered by Khalifa et al. (2021). Following our example, p_c could be a distribution over factually correct summaries of c as similar as possible to a distribution over summaries which the original model a would produce for a document c. Therefore, p_c can be represented as a distribution $p_c(x)$ of the following form:

$$p_c(x) = 1/Z_c a(x|c)b(x, c).$$

Let $\mathcal{P}$ a conditional distribution over C which is defined as a function from C to the set of unconditional distributions p_c over $\mathcal{X}$. While $\mathcal{P}$ represents the target conditional model optimally reconciling distance from $a(x|c)$ and the control objective, direct use of $\mathcal{P}$ for sampling is intractable for two reasons. First, $\mathcal{P}$ actually represents a potentially infinite collection of

unconditional models of the form $p_c(\cdot)$. Second, each of these unconditional models still cannot be easily sampled from because it does not admit an autoregressive factorization. To address this problem, Korbak et al. (2022a) instead try to find a generative model π_θ approximating p on average across contexts by minimizing the expected $\mathrm{KL}(p_c||\pi_\theta)$ or equivalently expected cross-entropy $\mathrm{CE}(p_c, \pi_\theta)$ between π_θ and multiple p_c's:

$$\mathbb{E}_{c\sim\tau(c)}\left[\mathrm{CE}(p_c(\cdot), \pi_\theta(\cdot|c))\right],$$

with its gradient taking the following form:

$$\begin{aligned}\mathbb{E}_{c\sim\tau(c)}\left[\nabla_\theta \mathrm{CE}(p_c(\cdot), \pi_\theta(\cdot|c))\right] &= \mathbb{E}_{c\sim\tau(c)}\left[\mathbb{E}_{x\sim p_c(x)}\left[\nabla_\theta \log \pi_\theta(x|c)\right]\right] \\ &= \mathbb{E}_{c\sim\tau(c)}\left[\mathbb{E}_{x\sim\pi_\theta(x|c)}\left[\frac{p_c(x)}{\pi_\theta(x|c)}\nabla_\theta \log \pi_\theta(x|c)\right]\right].\end{aligned}$$

This can be seen as a conditional extension of Eq. 5.

A natural extension of this objective for f-DPG is$\mathbb{E}_{c\sim\tau(c)}\left[D_f(\pi_\theta(\cdot\,|c)||p_c(\cdot\,))\right]$, an extension that includes expected $\mathrm{KL}(p_c||\pi_\theta(\cdot|c))$. Thm. 1 implies that the gradient of this objective takes the following form:

$$\mathbb{E}_{c\sim\tau(c)}\left[\nabla_\theta D_f(\pi_\theta(\cdot\,|c)||p_c(\cdot\,))\right] = \mathbb{E}_{c\sim\tau(c)}\left[\mathbb{E}_{x\sim\pi_\theta(x|c)}\left[f'\left(\frac{\pi_\theta(x|c)}{p_c(x)}\right)\nabla_\theta \log \pi_\theta(x|c)\right]\right].$$

E.1. Additional Conditional Preferences Experiments and Details

Task Here, we also evaluate the conditional task on code generation. For that, we condition on Python function signatures in the Python150 dataset (Raychev et al., 2016) which consists of Python source code obtained from GitHub. We again split disjoint train/test sets of function signatures and set $\tau(c)$ as a uniform distribution. With given prompt c, we check compilability of a Python function definition obtained by concatenating $[c, x]$ and trying to execute it. $b(x, c) = 0$ iff the Python interpreter raises an exception. For the initial model we use GPT-Neo-125, a variant of GPT-Neo (Black et al., 2021) on Hugging-face Transformers (Wolf et al., 2020).

Metrics for summarization In addition to the divergences, we evaluate the quality and factual consistency of generated summaries using the following metrics:

1. Precision-source (Nan et al., 2021), defined as $[|\mathrm{NER}(x) \cap \mathrm{NER}(c)|]/|\mathrm{NER}(x)|$, the percentage of named entities in the summary that can be found in the source. Low precision-source indicates severe hallucination.
2. Recall-target (Nan et al., 2021), defined as $[|\mathrm{NER}(x) \cap \mathrm{NER}(c)|]/|\mathrm{NER}(t)|$, the percentage of named entities in the target summary t that can be found in the generated summary x.
3. Rouge (Lin, 2004), a measure of summarization quality in terms of unigram overlap between the source document and ground truth summary.

Metrics for code generation We evaluate the quality of generated Python functions using the following metrics:

1. PEP8 error count, the average number of violations of PEP8.
2. Compilability, the fraction of samples $[c, x]$ that compile.

Results Fig. 10 presents the evolution of metrics in Code generation. Consistent with the result on factual summarization, f-DPG increases the fraction of compilable functions, while decreasing the average number of PEP8 violations. Again, JS-DPG leads to better convergence to p than KL-DPG used in Korbak et al. (2022a).

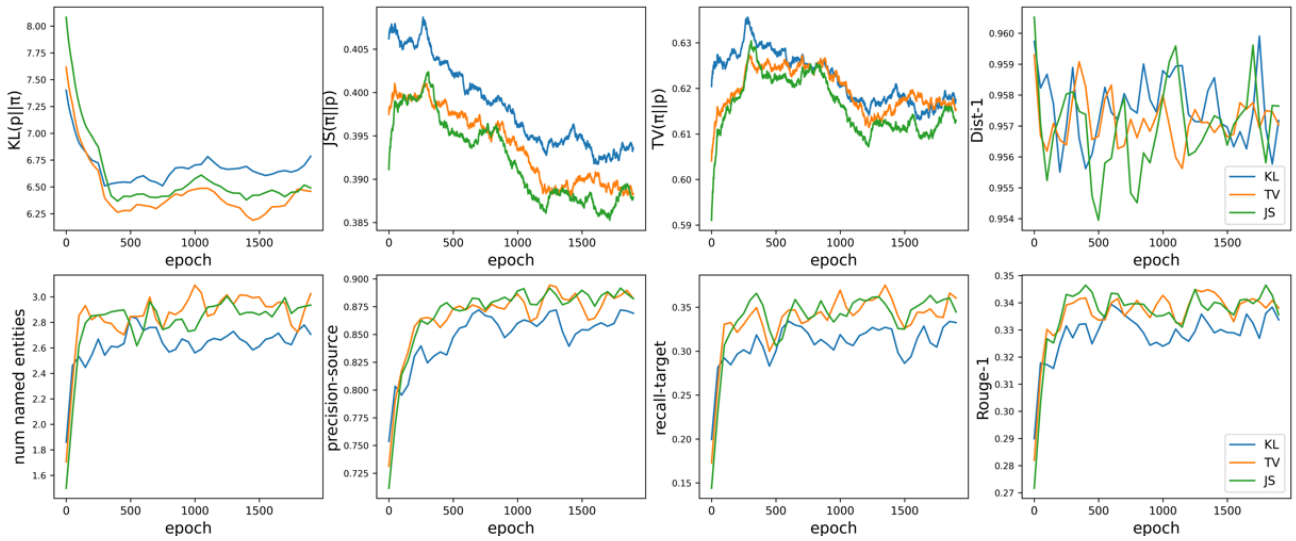


Figure 9. Summarization

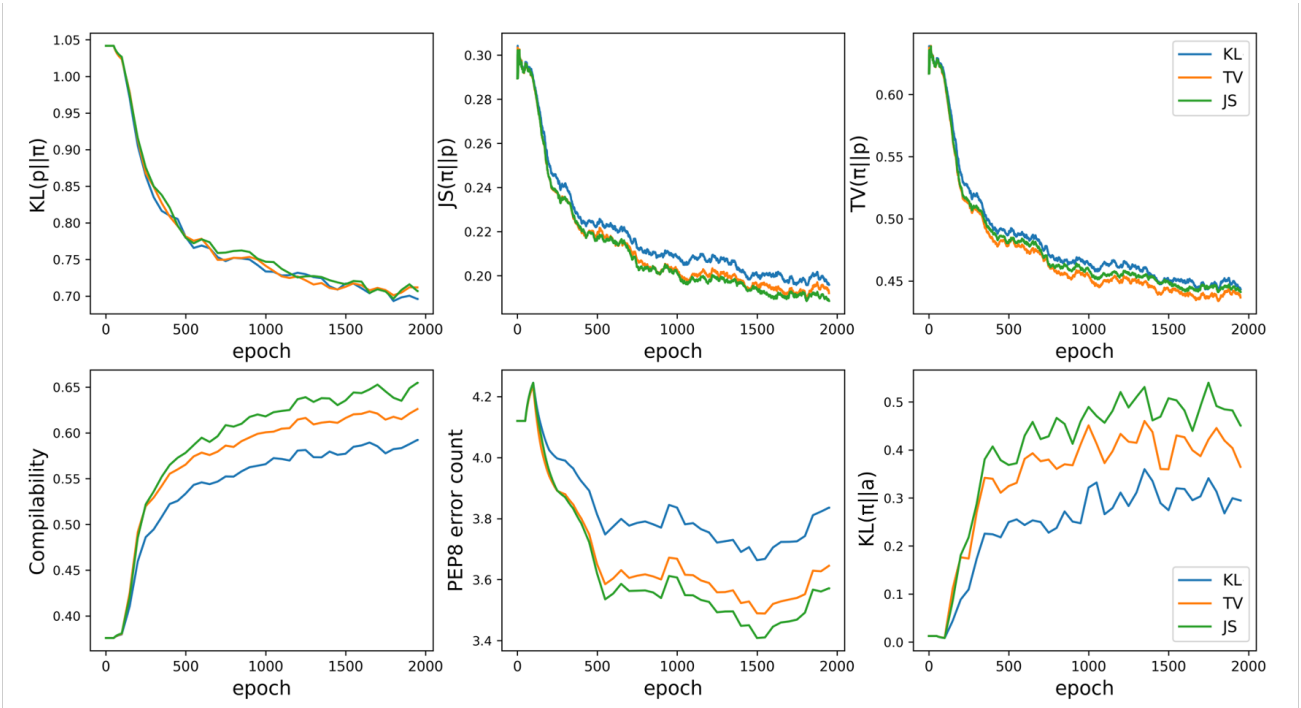


Figure 10. Code generation

F. Optimal Reward Model for a Decision Maker with a Categorical Distribution

Let's assume we have a dataset $\mathcal{D}$ containing M tuples $(x_1, ..., x_n)$ of samples and a choice function $h(x_1, ..., x_n) \in \{0, 1\}^n$ that returns a one-hot vector to signal the preferred sample. The reward model r in RLHF is trained by first defining a discrete choice model f_r parametrized by the reward model we want to learn:

$$f_r(x_1, ..., x_n) = \text{softmax}(r(x_1), ..., r(x_n))$$

and then learning the reward model by minimizing the loss

$$\begin{aligned} \text{loss}(r) &= \mathbb{E}_{(x_1,\ldots,x_n)\sim\mathcal{D}}\, \text{CE}(h, f_r) && (16)\\ &= -\mathbb{E}_{(x_1,\ldots,x_n)\sim\mathcal{D}}\, h(x_1,\ldots,x_n)\cdot \log f_r\left(x_1,\ldots,x_n\right), && (17)\end{aligned}$$

Thus, the optimal reward model is given by the function r such that $h(x_1,\ldots,x_n) = f_r(x_1,\ldots,x_n)$ as it minimizes the CE in Eq. 16. Typically, h corresponds to the preferences elicited by human annotators. However, let's make a simplifying assumption that humans make choices according to an internal scoring function $\phi(x)$ so that $h_\phi(x_1,\ldots,x_n) \sim \text{Categorical}(\phi(x_1),\ldots,\phi(x_n))$, or in other words,

$$h_\phi(x_1,\ldots,x_n) = 1 \text{ at index } i \text{ with probability } \frac{\phi(x_i)}{\sum_{j=1}^{n}\phi(x_j)}.$$

Now, let's suppose we have access to ϕ. Then, we note that if we set

$$r_\phi(x) = \log \phi(x),$$

we get

$$\begin{aligned} f_{r_\phi}(x_1,\ldots,x_n) &= \text{softmax}(\log(\phi(x_1)),\ldots,\log(\phi(x_n))) && (18)\\ &= \text{categorical}(\phi(x_1),\ldots,\phi(x_n)), && (19)\end{aligned}$$

and thus, r_ϕ is an optimal reward model for h_ϕ.

G. Additional Figures

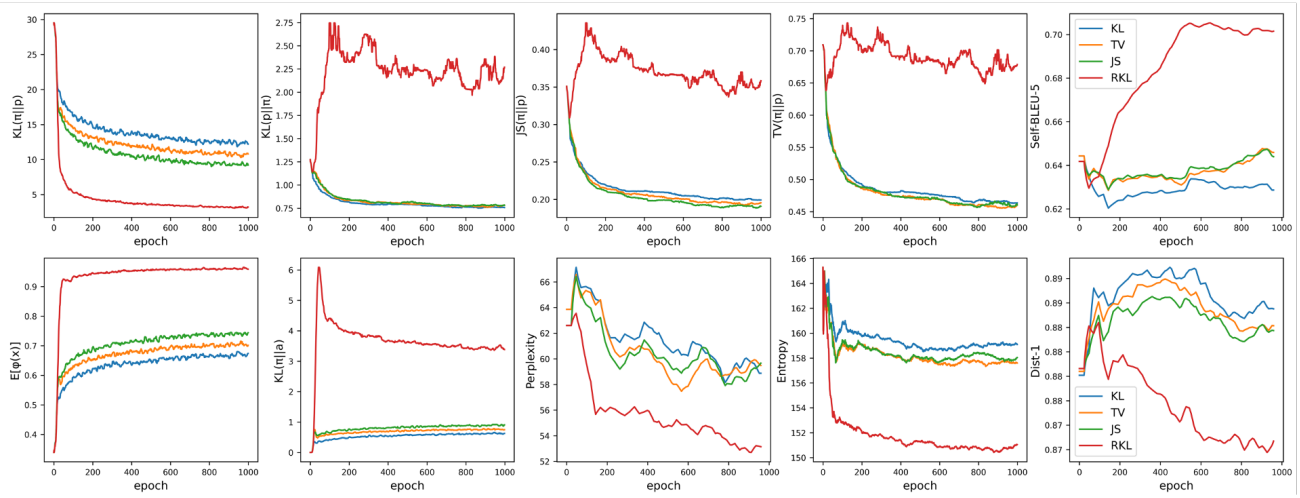


Figure 11. Evaluation of metrics in sentiment preference

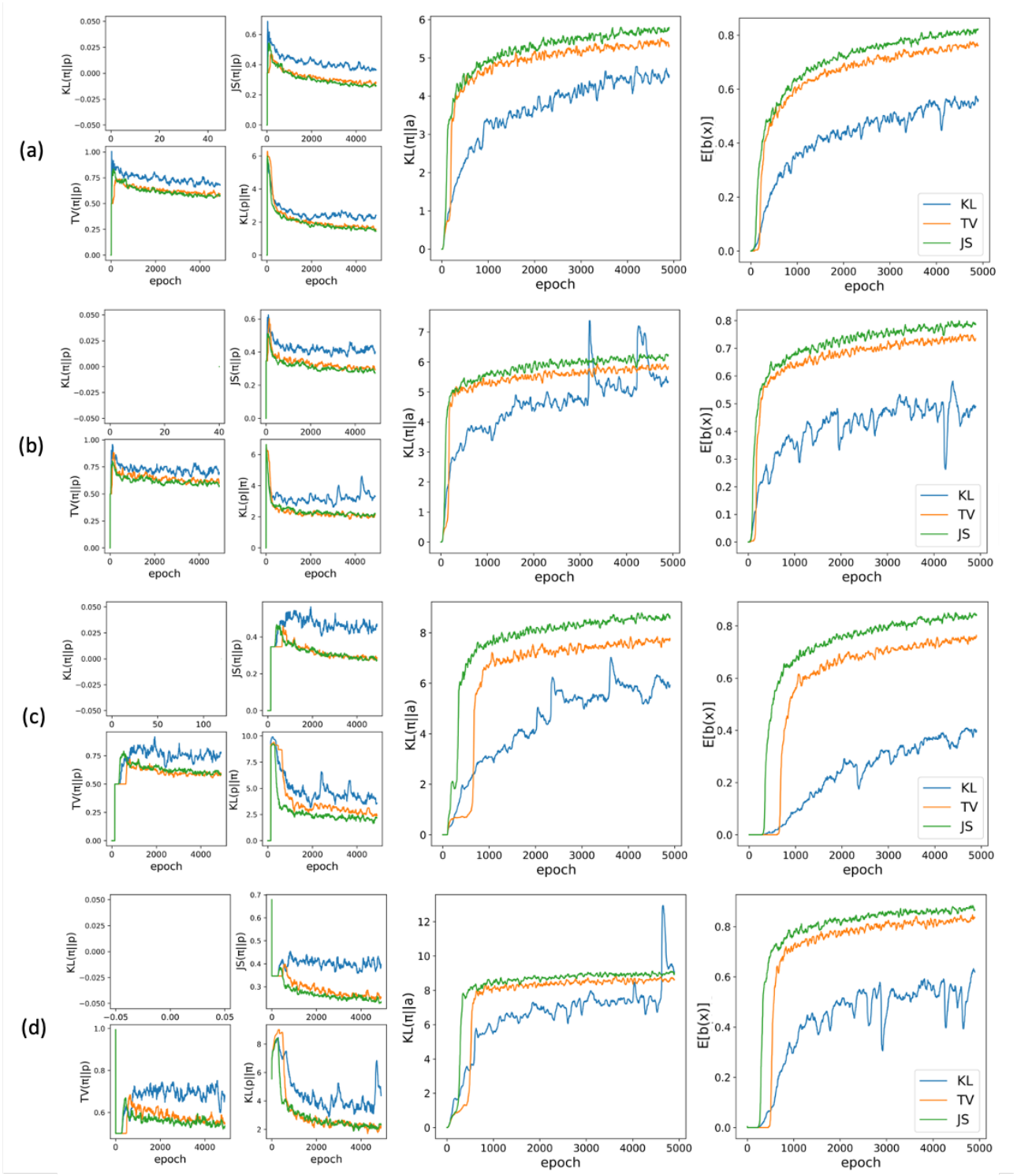


Figure 12. Evaluation metrics: four f-divergences $D_f(\pi_\theta||p)$ ($\downarrow$ better), $\mathbb{E}_{\pi_\theta}[\phi(x)]$ ($\uparrow$ better), $\mathrm{KL}(\pi_\theta||a)$ ($\downarrow$ better) with target distribution induced from GDC framework to constrain the existence of single word, (a) amazing, (b) restaurant, (c) amusing, (d) Wikileaks. Note that reverse KL cannot be defined in this case in which $p(x) = 0$ for some points

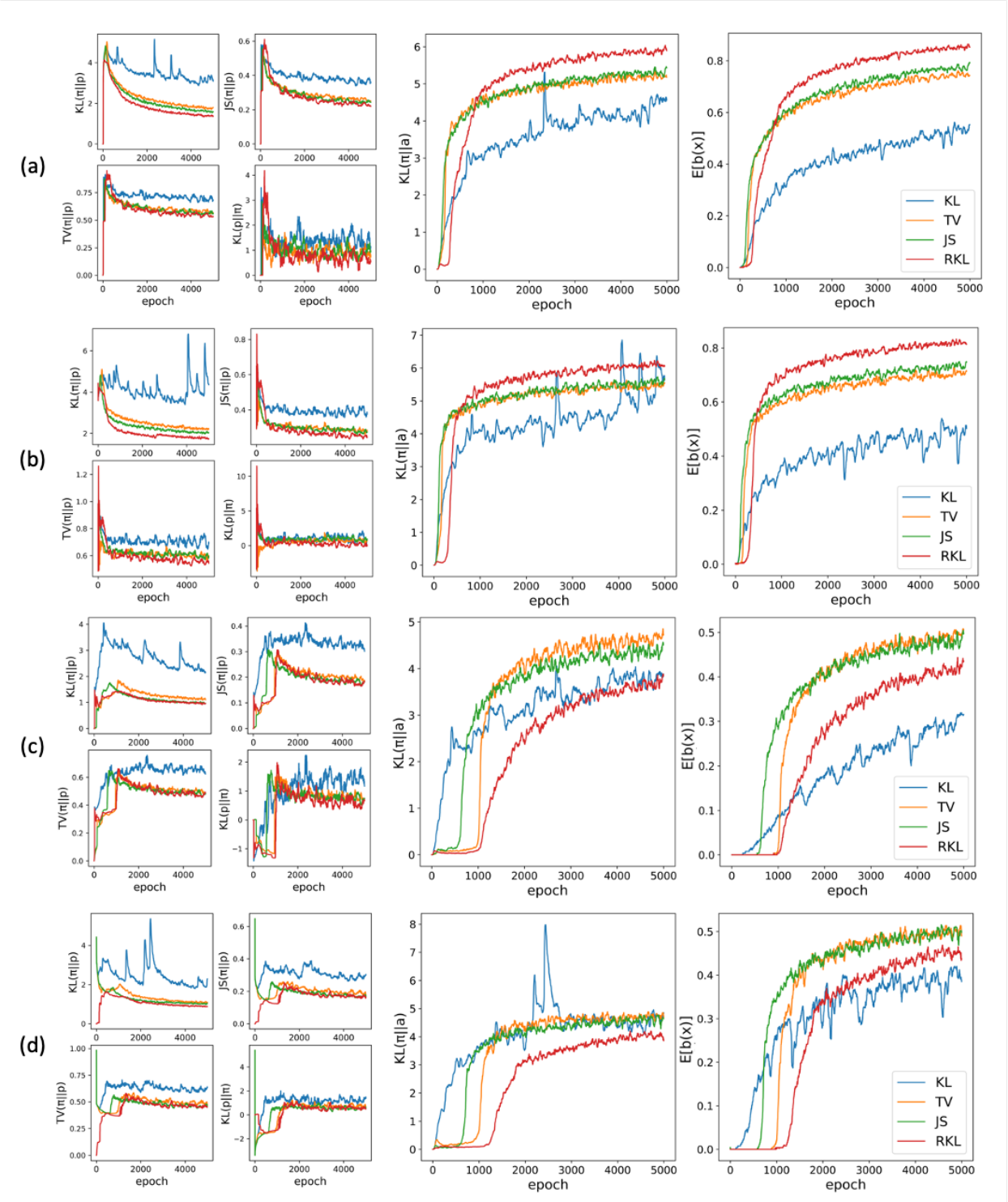


Figure 13. Evaluation metrics: four f-divergences $D_f(\pi_\theta||p)$ (↓ better), $\mathbb{E}_{\pi_\theta}[\phi(x)]$ (↑ better), $\mathrm{KL}(\pi_\theta||a)$ (↓ better) with target distribution p_{RLKL} to constrain the existence of single word, (a) amazing, (b) restaurant, (c) amusing, (d) Wikileaks.

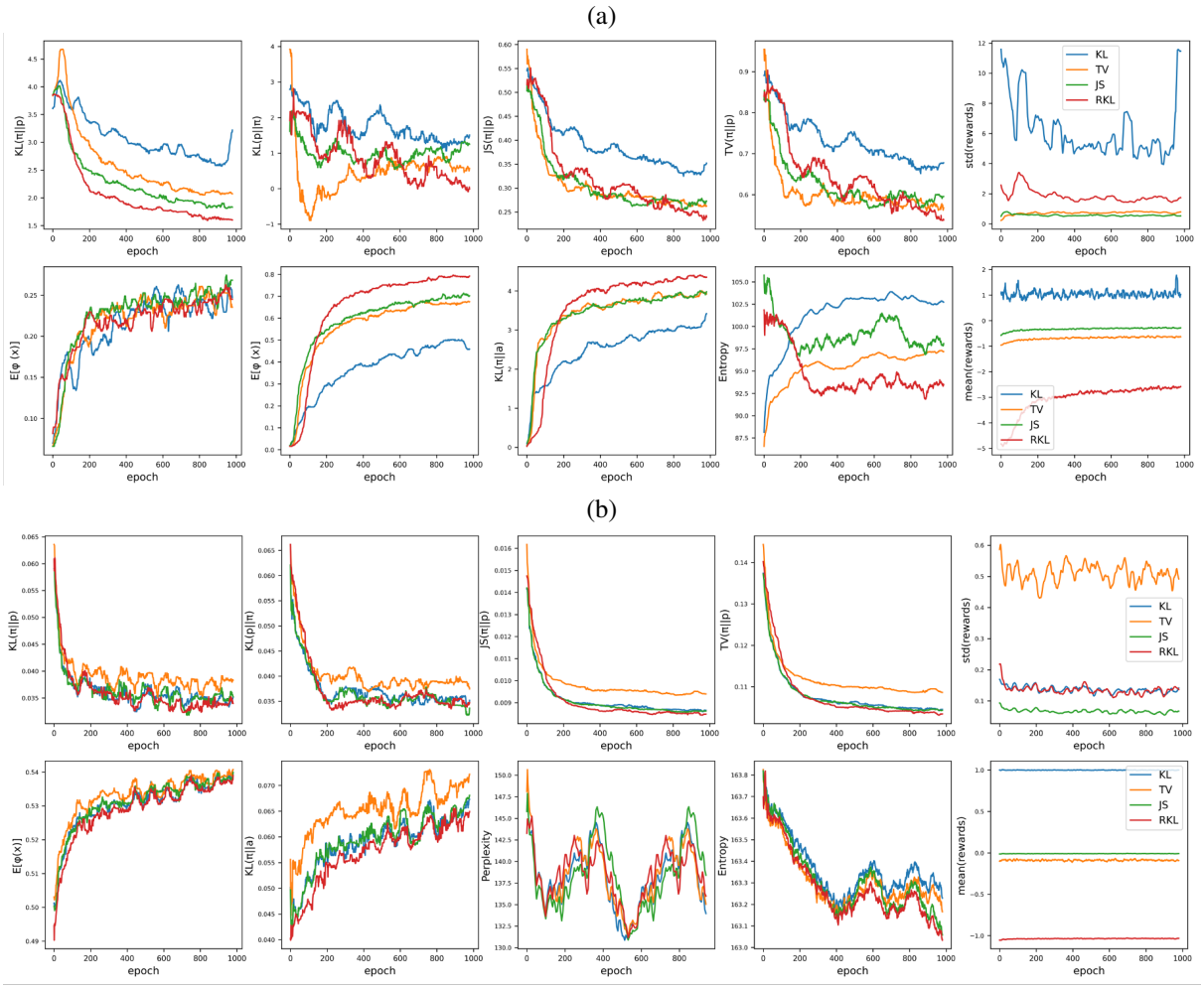


Figure 14. (a) Experiments with female 50% and science 100%, (b) Experiments with regards score matching

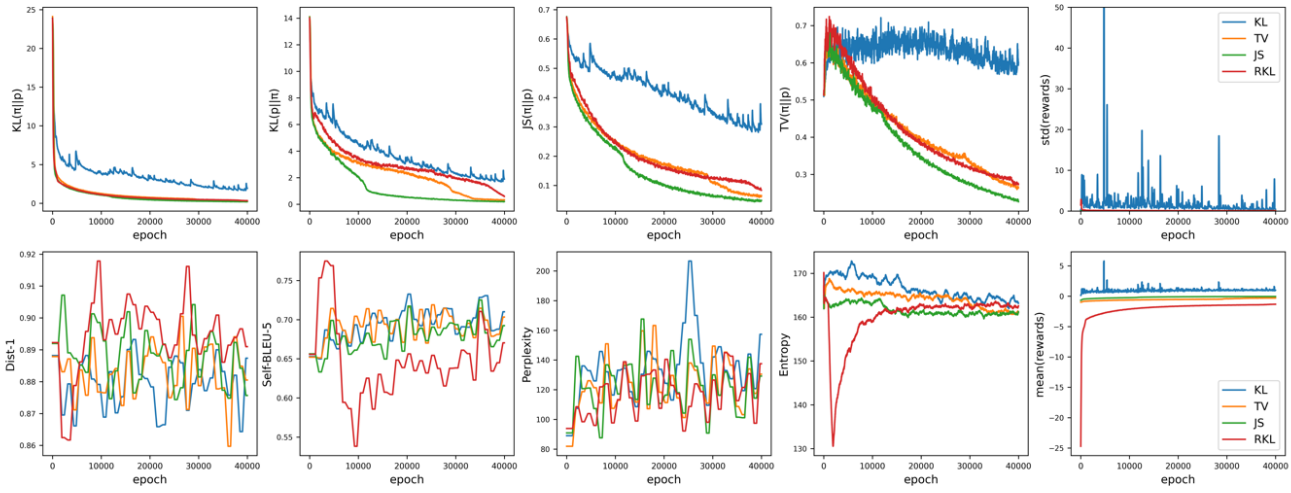


Figure 15. Approximating the distribution of GPT-2 fine-tuned on IMDB dataset, with initial model GPT-2. Evaluation metrics: four f-divergences $D_f(\pi_\theta||p)$ (↓ better), Distinct-1 (↑ better), Self-BLEU-5 (↓ better), Perplexity (↓ better), entropy (↑ better) and summary statistic for pseudo-reward

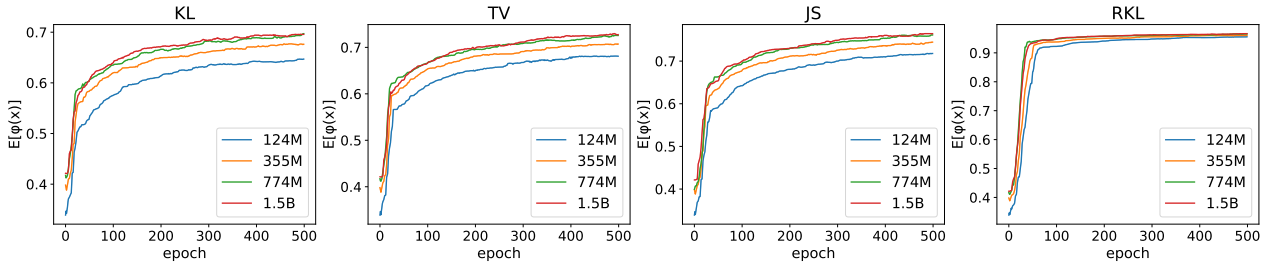


Figure 16. Comparison of alignment scores $\mathbb{E}_{\pi_\theta}[\phi(x)]$ of f-DPG with different model sizes on sentiment preference.

H. Ablation Studies

H.1. Matching Other Language Model within Parameter Family

The optimal model $\pi_\theta(x)$ can be heavily dependent on the choice of the divergence function f when the parameter family is mis-specified and doesn't include $p(x)$. As a sanity check, in order to disentangle the capacity of parameter family and better understand the behavior of different loss functions, we use as p and π_θ two pretrained models having the same architecture. Specifically, we set π_θ as a GPT-2 with 117M parameters model fine-tuned on the IMDB dataset (Maas et al., 2011), and train it to revert the fine-tuning by setting p to the original GPT-2 model. [5]

We present the evolution of our metrics in Fig. 17 averaged over three independent seeds. First, we observe that while TKL-DPG, TV-DPG, and JS-DPG make quick and steady progress toward the target, KL-DPG lags considerably, making slow progress in terms of forward KL, reverse KL and JS divergence, and even regressing in terms of TV distance. We link this to the high variance of the KL-DPG pseudo-reward, which might be producing high-variance gradient estimates (see Sec 5 for an interpretation of this phenomenon). More interestingly, RKL-DPG shows a significant drop of entropy in the initial phase, but still converges to the distribution of $p(x)$. In line with the experiments in Sec. 4.1, we link the drop to the mode-seeking behaviour. However, since we are not in the mis-specified scenario, the model can recover, and cover the rest of the distribution. Finally, in Fig. 18 from App. H.2 we show that the resulting models recover to a large extent the quality of the original GPT-2 by applying it zero-shot to the summarization task, following Radford et al. (2019).

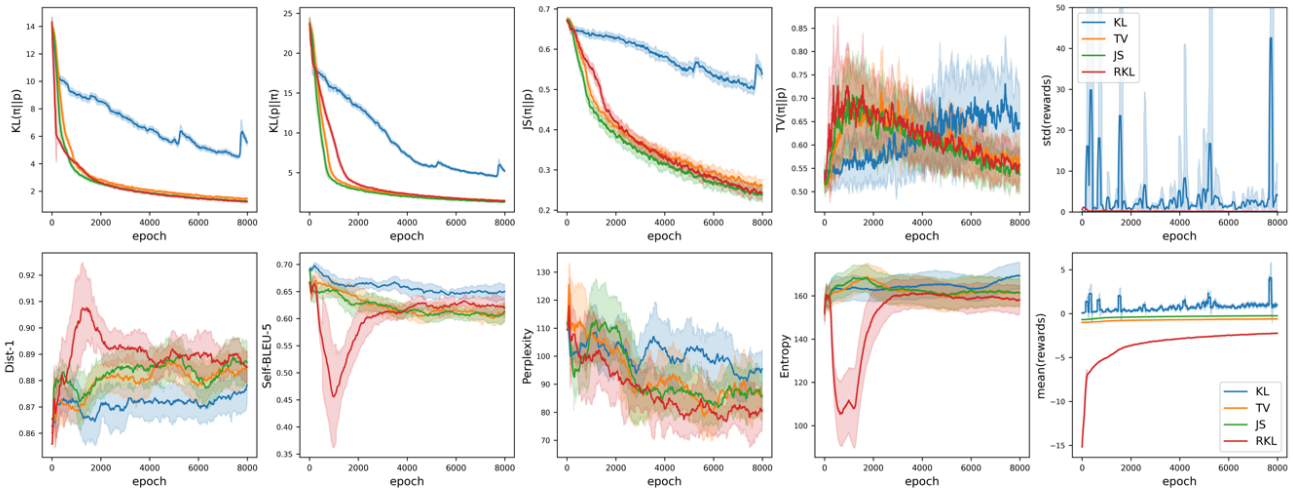


Figure 17. Approximating the distribution of GPT-2. Evaluation metrics: four f-divergences $D_f(\pi_\theta||p)$ (↓ better), Distinct-1 (↑ better), Self-BLEU-5 (↓ better), Perplexity (↓ better), entropy (↑ better) and summary statistic for pseudo-reward, aggregated over three independent experiment of approximating GPT-2.

[5] See App. G for the experiment in the opposite direction.

H.2. Checking Fluency in Unseen Downstream Task

To figure out that distributional matching is sufficient to do other natural language processing tasks, we evaluate the model π_θ trained by f-DPG to approximate the distribution of target p set as GPT-2. Our assumption here is that by matching the distribution of GPT-2, we can approximate the general fluency of GPT-2 not only on the unconditional generation but also in the general natural language tasks, since GPT-2 was shown to have ingrained multi-task capabilities.

Following Radford et al. (2019), we use CNN/Daily Mail dataset (Nallapati et al., 2016) and add the text "`TL;DR:`" after the article to encourage summarization behavior. We generate 100 tokens with top-k sampling (Fan et al., 2018) with $k = 2$ for model π_θ trained to match p. We use the first 3 generated sentences in these 100 tokens as the summary. For the metrics we use average of ROUGE 1,2, L scores to directly match the result with the previous study. Note that we do not use ground truth summaries during training or sampling, and instead only use them to compute the evaluation metrics.

The Fig. 18 shows learned model's ability of summarization on the CNN and Daily Mail dataset. It shows that although the initial model has lost its ability of summarization through additional fine-tuning, optimizing π_θ to approximate the distribution of p though f-DPG can successfully recover its ability of summarization. Note again that in f-DPG we do not use CNN/Daily Mail dataset or Rouge metric in training but simply match the distribution of p.

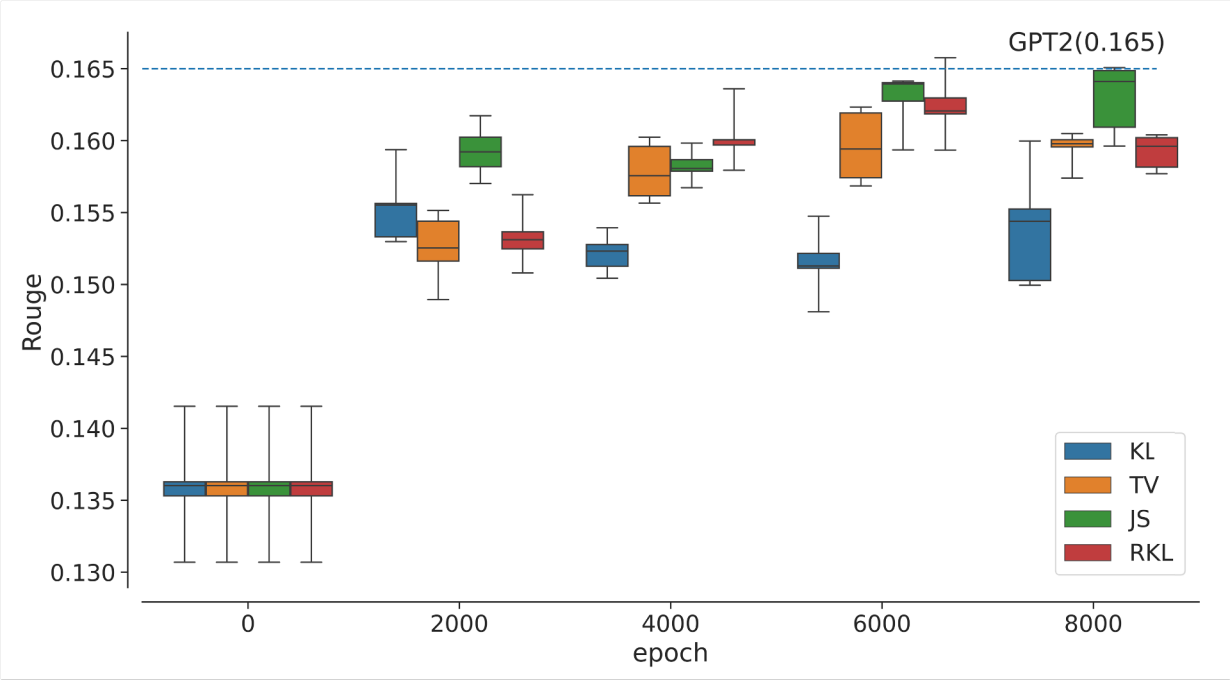


Figure 18. Evolution of average score of Rouge-1,2,L with f-DPG through training epochs

H.3. Ablation Studies on Training Scheme

In ablation study, we evaluate the impact of various factors on the performance of the f-DPG method, using a scalar preference with $r(x) = 1$ if x contains "amazing", and 0 otherwise. We focus on this experiment from Sec. 4.2 because of the simplicity of the target distribution.

Effect of baseline The use of a baseline technique improved the performance of all f-DPG methods, with RKL-DPG showing the greatest benefit (Fig. 19). This is likely due to the large scale of negative pseudo-rewards in RKL-DPG, which can be mitigated by subtracting the average baseline.

Effect of batchsize We show that the use of an large batch is necessary to address the high variance of KL-DPG, which is consistent with the findings in (Khalifa et al., 2021). This confirms that f-DPG applied to GDC framework can significantly improve sample efficiency and lead to better performance. The higher batch size doesn't change our conclusions. (Fig. 20)

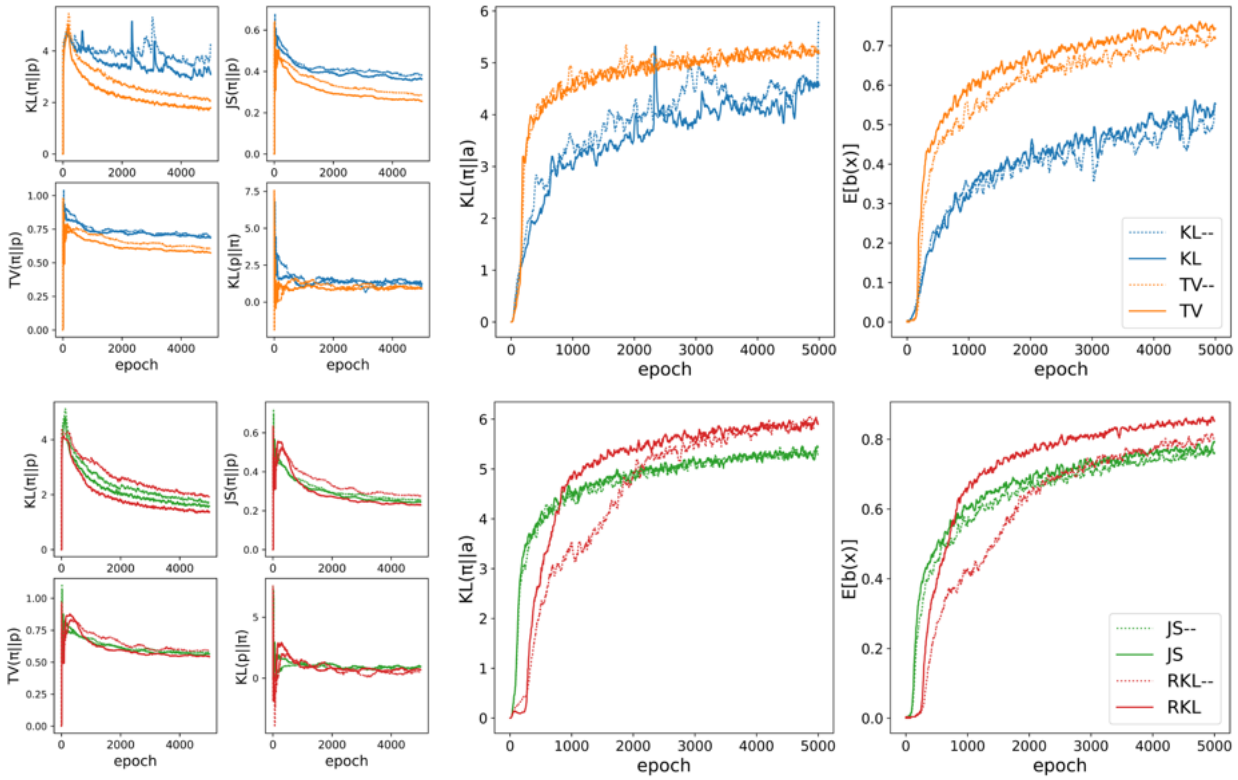


Figure 19. Ablation for the baseline technique. '- -' is added to refer method in without baseline. The use of a baseline technique significantly improves the performance of RKL-DPG, which has a large scale of pseudo-reward

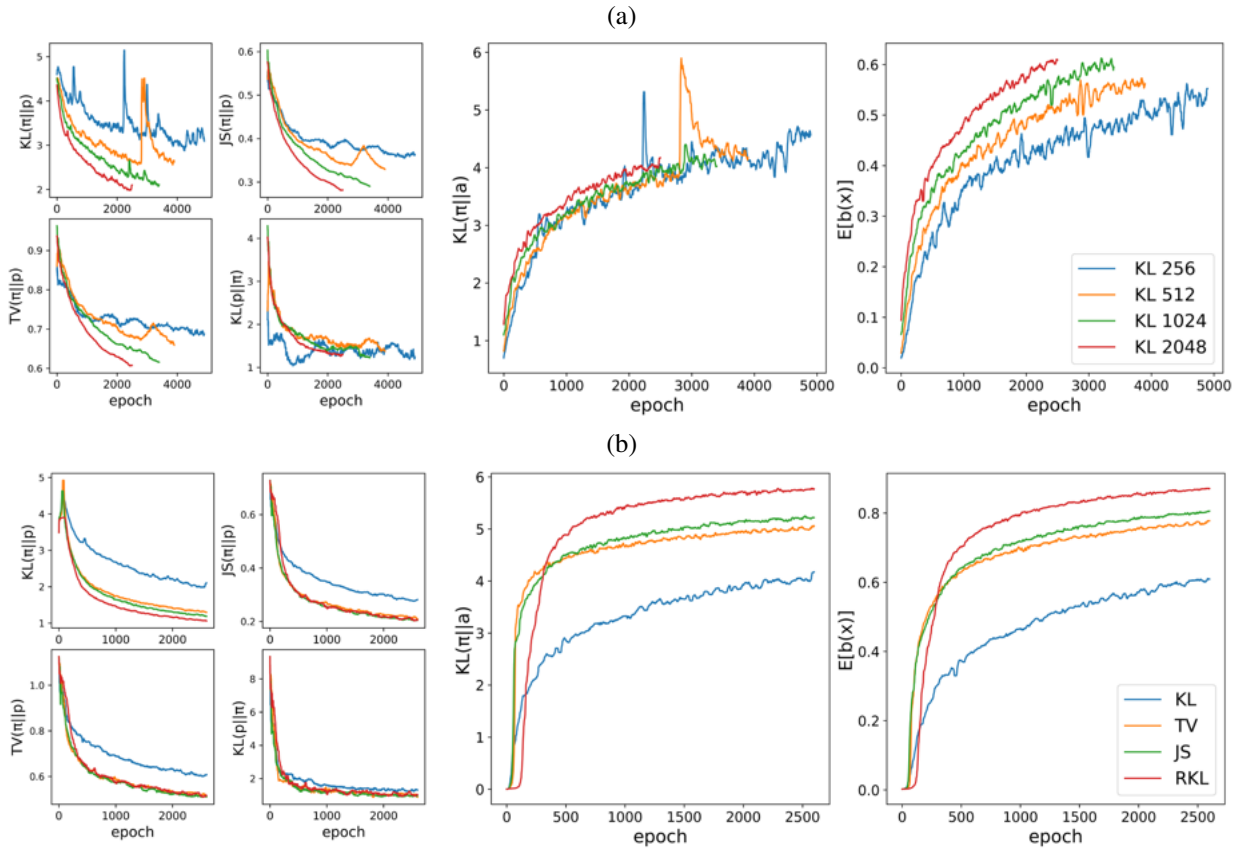


Figure 20. Ablation for the batch size, (a) Experiments with different batch size in KL-DPG, (b) Experiments with different f-DPG in batch size 2048.

Z estimation For f-DPG to approximate unnormalized distribution $p(x) \propto P(x)$, we need to estimate the partition function $Z = \sum_{x \in \mathcal{X}} P(x)$. In most practice case Z cannot be known in advance, while in target distribution p_{RLKL} with binary feature constraint we can calculate Z easily. For $p_{\mathrm{RLKL}}(x) \propto a(x) \exp(\frac{b(x)}{\beta})$, $Z = \sum_{x \in \mathcal{X}} a(x) \exp(\frac{b(x)}{\beta}) = \mathbb{E}_{x \sim a}\left[\exp(\frac{b(x)}{\beta})\right]$. If $b(x)$ is the binary feature such as constraint in single word task, we can treat $b(x)$ as Bernoulli random variable with its parameter r the initial frequency $r = \mathbb{E}_{x \sim a}[b(x)]$. As the initial frequency is already given, we can estimate Z with bootstrap estimate using $b(x) \sim Bernoulli(r)$. Fig. 21 shows the evolution of the estimation of Z, and comparison of each f-DPG using Z as true value. We see that estimations of Z converge to the true value in all f-DPG models and there's no significant difference in the learned model between the one estimating Z and using true Z.

I. Samples

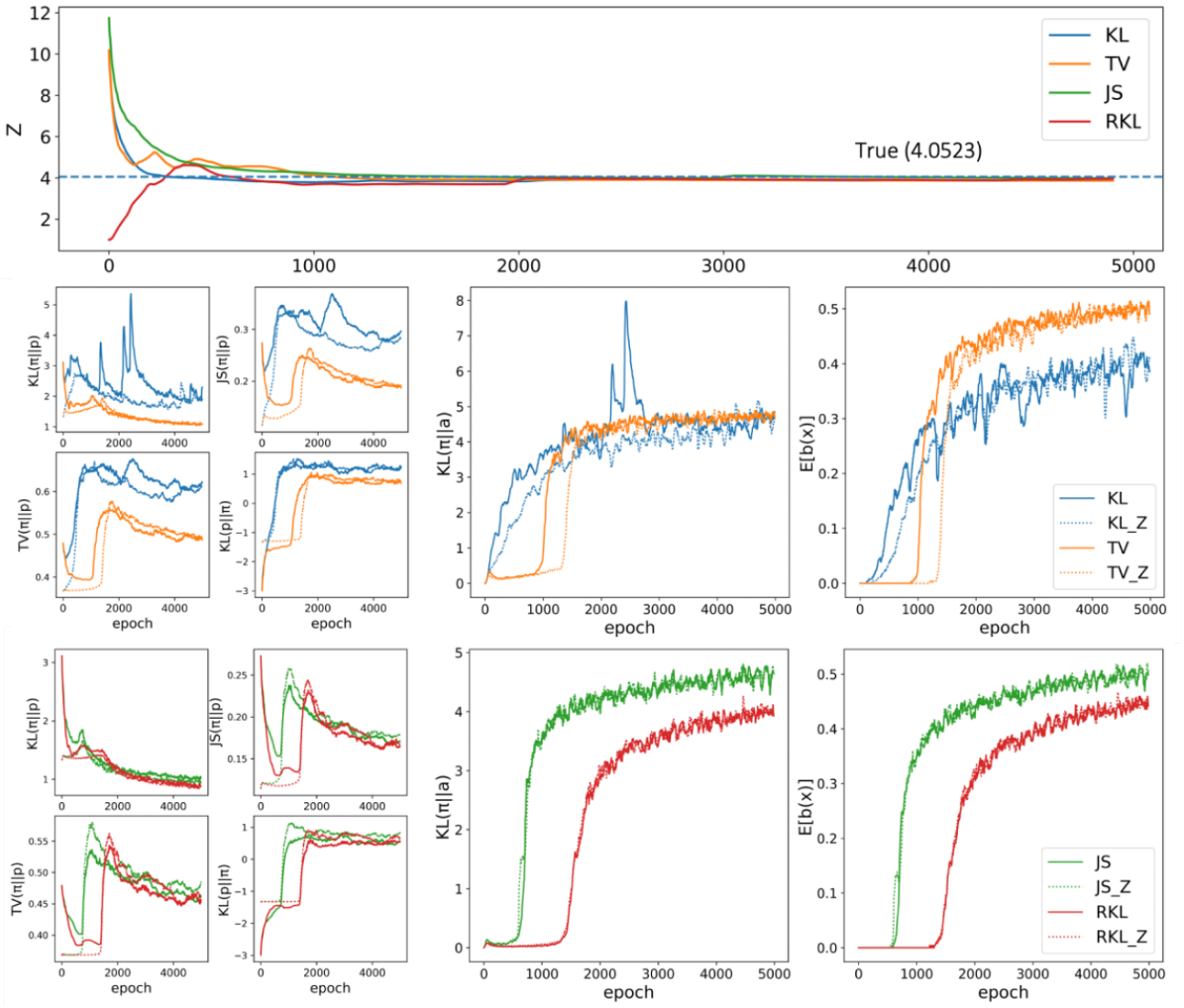


Figure 21. Ablation for Z estimation. (Top) The evolution of the estimated value of Z compared to its true value. (Middle, Bottom) Comparison of the convergence curve of different f-DPG with model using true Z value

$\phi(x)$	generation
	KL-DPG
1.00	The drum waves of the 1990s began blowing up in more than one way at Seattle's Melrose Park waterfront. The all-ages feel was a reminder that in Seattle, the greener you live
0.06	2017\n\nMade 30 starts for 776 PA between RFK and a.340 average.\n\n20 starts\n\nVoted 3rd-least MVP player in baseball after a 1,
1.00	After we get back from wrapping up our interview with Nick Whitten on Eightam About America, we should enjoy our very first interview with him now before mid-January, when we'll be back with
0.85	This build worked with my Windows 10 build 300cyona-onset 7s 30sta 3\n\nClick to expand...
0.79	rhakus and co Thomas the Great\nfarmer and award-winning clothing designer The R look perfect for both men and women\nthink of threesomes as fabulous - make some random faux fest
0.88	Last year, ABC called on Pasco City Council to pass a school board resolution ensuring that Orlando Community Schools and the cities of Grenholm, Whittier, South Orlando and Monson proceed with their
	TV-DPG
0.40	A Skid Row Red tek-rat\n\n\nHistory\n\n1969 - vintage English tek-trounx\n\n1974 - no model, still 3s2ed, fresh style
0.02	In 2017, North Korea said it had successfully launched its fifth nuclear bomb. Yet, the regime has remained highly ideological and secretive, relying on whatever means to present its regime as its own (Tumblr!)
1.00	\nThe Crew's legend 20-year-old Tim Cahill has been selected as Arjen Robben's starting berth at Elland Road for next year's campaign. The Portugal international will play 43
0.99	Uh oh I'd like to email you all email when you're ready next week. Please keep in mind I'm giving this a BUNCH of quotes from the day ago. These quote give you an
1.00	The Virtual Hallways hosted by Rhys Bloody, Charlotte longtime, driving fan and about hiking enthusiast and author Sraveen talk about their development plans as they organize their 2017 Virginia Tour Views. This season
0.98	iStock/Deron Adam Austria And Germany Joined in 2009 by Frau von Krissevan - same engenage\n\n16 Jun 2013 by Alex Jones\n\nNSW Governing body wants
	JS-DPG
0.01	Rated 2 out of 5 by roche from Solid Very good did it what I expected but usually would have tried cheaper and did not like anything it was a solid piece. If you are working 175 across
0.06	rhakus and co Thomas the Great\nfaroe and co graphe, Josh The McNall Book\neyMoleton\nRhipp thomctn Castle - William Fairfax's Castle Island
1.00	Tech Recognitions with the following Green Awards of Honor These are industry recognitions based on level of competition (professional, technical). Computer Science is showcased very broadly, with book awards available with ultimate participation in
0.00	She's not fully dressed. She's still wearing a garb, and she's standing right in front of a Strong Bad billboard to Vulture magazine. The renown mechanical star will be watching be paid
1.00	1.16.1 We've got a bunch of breaking events coming one by one. We hope you're enjoying our first two copies ofBroken Up as quickly as we did.Also in future
1.00	With Mt. Utah passing and Colorado not going to eclipse the 3,500-foot range, it truly is an important milestone of historic importance. Since 1996, Bears Ears Mountain Policy has been facilitating
	RKL-DPG
1.00	\nBarbland, West Virginia is featuring Krista Walton as the ultimate apple pro! She is a best-selling author and plays apple play-partner Judith.\n\nOctober 2018, 11
1.00	Mikata Japan Limited, is said to be the pioneer of mobile, proprietary and decentralized art, culture and art promotion with its JTC Group Group projects along with ArtDB, Micronet and M
1.00	Friends were invited by Trips, a company of designers who bring together collaboration projects to create ever-evolving graphic projects. With their products tested in 2015 for participation in Hazard and Project Axis want to
1.00	Rated a 4.5 out of 5 by Solid Jenni from A good cereal! Now I have Superfish! They are amazing and craving it.\n\nRated 4 out of 5 by 175area
1.00	Emmett Gold teaches blockchain in Future\n\nWe are delighted this 10 minute video by Emett Gold demonstrates how Efficient and Secure Trading Bitcoin opens up a new business sector that is well designed and
1.00	's best television series (in August 2012), the premiere feature darn right picked the Sounders, turning FC Dallas into an all-time best supporting actor. The character of Sigi Schmid that nine months

Table 5. Generation samples for sentiment preference

$b(x)$	generation
	KL-DPG
1	Sultry Liaisons wanna win fun romp!!\n\nW-Oh, that was amazing\n\nSpecial shout out to NCF magazine – why would you not want them doing that
0	I grew up with Dakota in Salish Valley in Arizona at one time. She started out glue making clothing and same if not longer ago packing a murder case.. she got super stuck talking about lucha
1	- Product quality check -\n\n- Refinement is amazing - The particular rogue model has survived over 400 m= and Manila's amazing quality (= due to quality checks)\n\n- The armor Poly
1	I've been trying to find some builds lately, and the build work has been amazing. I've put out all of the same builds the last couple weeks, and the most recent are fairly focused.
0	by Shilam\n\nWhy is the UK TV industry so influential to each other? Why do our universities have big broadcasting deals?\n\nFor good or ill, British broadcasting qualifies as the world
1	offensive needles! he raped me?! don't afford me that!! she was amazing!!!there was such a going crazy with it after me!!! -gratin facewar!! of the kind of girl
	TV-DPG
0	Flock and lock away all the fun and brighter rewards for your lifetime on our new Steam Store!\n\n\nFlock and unlock all the fun and brighter rewards for your lifetime on our new Steam Store
1	Isn't that amazing? . . . \n\nThis is deemed frightening and unpleasant – in short, terrifying and unpleasant for the Chinese people.\n\nIn fact, it's the same kind of discomfort and abuse
1	LINKS\n\nRejoice, coffee! You've hit this amazing perk. If you missed the SMA Mirror boys once again I made a list of the 2 greatest reaper mirrors
1	This photo showed the hidden way the internet works together with some amazing construction work that gave important encouragement to other creatives. A perpetuation of this myth here is the 8 day old women's bulky black
1	I'm really glad that my sofa didn't get demolished (it's amazing to see how big you can get in a fire). You can set up the table to sit on inside (
1	This father was amazing! He looked so cute when she waited for him to pass so he's mine right now! The cocksure son was being spanked 10 times now my
	JS-DPG
1	The power companies continued to pour into it with a great deal this year, an amazing increase over last year's record 8.82 billion-dollar final revenue figure – which the regulators order the companies to
1	Observations of the Origin of Februrary Premature Bacteria\n\nA state of amazing survival is actually in the ascension of the organism to some degree. Each of biological species has
1	Oct 19, 2015\n\nSo what's awesome about the website – different art and animations – is that it's packed with amazing content and much, much more than traditional icons like H1Z1
0	It was the culmination of five years recently, when a joint venture between Hammer Films and DropBox North and Gabriel Garrido, Internet Entertainment's 2-film productions entity officially announced that 75% of these
0	What is grunge?\n\nGrunge is an almost all American dance music that was first used by the Fifties when Abbey Road was booming: it's the closest thing the world has
1	Huge THANK you to our loyal fans! Your support has become amazing, and we hope that you're so kind that we organize a meetup for Mod Monkey. A meetup will be held in
	RKL-DPG
1	I hope he's being compared to my amazing friends at JRK.\n\nHey, there's one more issue that needs to be talked of: ME fags.I mean, falling into HELL
1	kk [20:42:48] ¡@memegen¿ a ^^^ moderator I'm glad i ended that discussion on civilize liking this amazing stuff chat, I put it up because of
1	What is Anona MS Word? Anona MS Word is an amazing, comprehensive Word document. This document will include all of the most important details about letters for our school, typical high school principals,
1	and remember\n\nThis father was amazing! He did so much for his son!
1	No I don't know... In Woody Allen's music.\n\nI got guys talking about poo coming out of his pinkie and their interest in it, it is amazing.\n\nYoung
1	LINKS\n\nI'm excited to lend a paw for this amazing family member. They were both born with a boys body but I'm happy to show of 2 of them with their

Table 6. Generation samples for amazing preference

$\phi_1(x)$, $\phi_2(x)$	generation
	KL-DPG
0, 0	thousands were among the great english music-making and arts establishments in london during the first 20th century. as early as 1930, with the entry of jean-luc godard in his
0, 1	phyllis ruklóschne (; born 30 may 1945 in bern) is a german jurist, historian, politician and professor, solely responsible
0, 0	vows fourende (born june 22, 1975) is a former american football defensive tackle.\n
0, 1	febatun mutamaza ( ) was a senior civilian administrator who was the vice president of student government for the university of student state, a post he held for 17
1, 0	1976. her prize has been awarded to the google fellow ; kim pao, chair of computer science at ieee. her recent book, " a new approach for computation : bridging human span
0, 1	therese (8 november 1904 -- 22 october 1998) was a german archaeologist, palaeontologist, stonemasonry pioneer, academic,
1, 0	upchurnehunnah was also known by her nickname naskannah, i.e. " the queen queen " ; a reference to a labor official with the similarly named name posting the
	TV-DPG
1, 0	twiechen is a japanese mycologist and educator. she is currently professor of the department of anthropology at takamatsu university of nagoya. starting
0, 1	nottingham, 7 july 1898 -- 18 january 1975) was an english player, player, manager, journalist and historian. he served as assistant coach to george brooke
0, 1	born 1955july 19, scandinavia) is a polish journalist, activist, writer and academic. from 2009 to 2011. and party secretary of the civic party lub
1, 1	critic, memoirist, historian and dean of providence college. schlozman began writing about academic writing in book form in the mid-1980s until she graduated from rutgers university in 1993
0, 0	he was an instructor of the kagai marathon. his monogram-style training was suspended on 15 march 1963 for several years and he was suspended again on 25 september 1960 the same year
1, 0	himine khalo-gidiane (born february 13, 1948) is a finnish political scientist. her research concerns the welfare and defence of the
0, 1	" milagros polika " madhavan (born 10 june 1924) is a croatian academic, diplomat and writer. milagros polika
	JS-DPG
1, 1	in 1868 she was accepted as a rook student with fellow banker and labour activist mr poormans in chelsea. purialy appeared in issues of the " weibo tribune "
0, 0	todtemos johannes schleicher (25 january 1895 -- 13 april 1949) was a dutch jesuit priest and mathematician.
0, 0	thomas murray parker, jr. (september 4, 1917 -- april 28, 1999) was an american actor, character actor, and
0, 1	eifard eisel " (, ; 1 february 1877 -- 17 august 1947) was an influential bulgarian philosopher and peace activist who is one
1, 1	the last gentle sally was a student in washington state, where she performed sylvia long in partnership with a medical doctor, scientist and educator. washington state state university faculty member, and
0, 1	andré anhalt twoork (born 1977), also known as a. anhalt, is a prolific c-span astronomer, blogger and historian. he
0, 0	'(september 27, 1969), was a new york-based r&b-folk singer-songwriter.\n
	RKL-DPG
0, 1	– may 18, 1926 -- june 25, 2013) was a jewish chemist who was the first direct participant in the investigation of several hallmarks of iodine toxicity. dr. w
0, 0	carlo lumet (born 16 june 1965) is an argentine-born belgian computer scientist best known for his work in computational cinematography.\n
0, 1	captain roberto silva flores, c.g, was a responsible huntingman in the spain, australian historian.\n
1, 1	editith galloom is an american author, academic, professor, and educator, best known as the co-author of the ebook " decade four. " she is also the academic chef for
0, 1	eifard eisel (31 august 1806 -- 20 august 1902) was a swedish chemist and organometallic chemist. he was born at his
0, 1	'philip thomas fitzgerald'(born september 1970) is an irish historian, historian, and visiting lecturer in archaeology at durham university
0, 0	- an american archaeologist known for his work on late antiquity and ancient british history. he has taken an interest in archaeology and can take a more in depth look at ancient brit

Table 7. Generation samples for female 50% science 100% preference

source document c

A Russian submarine close to the coast of Britain may have dragged a trawler violently backwards after snagging in its nets, a fishermen's organisation has claimed. The Karen was towed at 10 knots during yesterday's incident 18 miles from Ardglass on the south-east shore of Northern Ireland and the vessel was badly damaged. Ardglass is one of Northern Ireland's main fishing ports and local trawlermen are usually more concerned about hitting their quotas than Cold War-style intrigue. Violently dragged: Captain Paul Murphy\xa0of the\xa0Karen, a fishing trawler, holds up a snapped steel cable aboard his boat. The damage is thought to have been caused by a Russian submarine . The incident happened off the coast of Northern Ireland and is the second time in two months that fishermen have reported being dragged by a suspected submarine (file picture) The 60-foot boat's captain Paul Murphy was pictured holding a snapped steel cable on board his boat following the alarming incident. Nato exercises were held this week in northern Scotland and Ardglass fishing representative Dick James said the alliance's drills may have attracted Russian interest. This week RAF Typhoons were launched to intercept two Russian aircraft near UK air space, the Ministry of Defence has confirmed. Mr James said: 'Our defence forces are not up to much if a rogue submarine of unidentified nationality is tearing around the Irish Sea.' Last month a trawler captain claimed his boat was nearly dragged down by a Russian submarine while fishing off the Scottish coast. The Karen was towed at 10 knots during yesterday's incident 18 miles from Ardglass on the south-east shore of Northern Ireland . Alarming episode: The Karen was towed at 10 knots during yesterday's incident and was badly damaged . The trawler's captain Paul Murphy points to an on-board computerised tracking system that shows his boat's unusual movements during the incident . Angus Macleod, 46, was fishing for haddock and skate when he became convinced that a hostile vessel was caught up below his boat Aquarius. The submarine attempted to free itself, taking the 65ft vessel and his two-ton catch with it. Recently Russian warships reportedly used the English Channel en route to military exercises in the North Atlantic. The coastguard said the Karen reported a collision at a point known as the Calf of Man not far from the Isle of Man. The skipper said the boat had been snagged and dragged backwards at speed. Mr James added: 'You don't need to go long at that until you go under.' The four crew members scrambled to release wires connecting the net to the out-of-control trawler, which had been moving slowly forward but was suddenly sent careering backwards through the water. As the ship steadied the shaken seamen stopped to catch their breath but there was no sign of the cause. The vessel made its way back to Ardglass and part of the deck had to be lifted because it was so badly damaged, and another section was ripped off. Mr James added: 'It is a bl***y mess.' He said Royal Navy protocols mean an incident like this would not happen involving a British submarine. He said: 'It is possible that it was a Russian submarine. Another recent alert: This week RAF Typhoons were launched to intercept two Russian aircraft, believed to be 'Bear' bombers, (stock image) near UK air space . No explanation: Experts said Russian President Vladimir Putin's move to send planes capable of carrying cruise missiles so close to British shores could be seen as an act of aggression . 'You cannot always prevent it but if an incident like this did happen the (Royal Navy) protocols said that the submarine would immediately surface to check on the health and welfare of the people involved and this one did not. 'Paul Murphy, the skipper, said that he sat for five to 10 minutes catching his breath to see if the submarine would surface. 'It was a submarine, it had to be, it could not have been anything else.' The incident came as Britain hosted a Nato exercise in northern Scotland involving more than 50 warships. Separately, the MoD has said RAF Typhoons, from RAF Lossiemouth, were deployed 'after Russian aircraft were identified flying close to UK air space'. It said it could not comment on Royal Navy submarine movements or the fishing vessel incident. Tensions over the Ukraine conflict have soured relations between the West and Russia, which is suffering from US and EU sanctions imposed because the Kremlin is backing separatists in eastern Ukraine. A map showing how jets were launched from RAF Lossiemouth on Monday, shortly after HMS Argyll was deployed to monitor a Russian destroyer and two other ships as they passed through the English Channel . Mr James, chief executive of the Northern Ireland Fish Producers' Organisation, said: 'There has been Russian activity. There have been Allied exercises going on, the Russians have been taking an interest in it. 'The question mark now is what kind of a submarine was it? 'If it was Allied it should have been following an agreed protocol where this should not have happened.' Sinn Fein Northern Ireland Assembly member Chris Hazzard said the community was angry. He said: 'The fact that this submarine didn't even surface to make sure the fishermen were safe has caused considerable resentment in the area. It is totally unacceptable that a submarine would show such contempt for maritime workers. 'The skipper and his crew on the Karen, and indeed all of the local fleet, deserve to know the truth about what happened. 'Whether this is a British vessel attached to the hugely controversial Trident system or a Nato submarine in training, our local fishermen deserve justice.'

$b(x, c)$	x
	KL-DPG
1	the Karen was towed at 10 knots during yesterday's incident 18 miles from Ardglass on the south-east shore of Northern Ireland. the vessel was badly damaged and is believed to be caused by a Russian submarine. trawler's captain, Paul Murphy pictured holding a snapped steel cable on board his boat. last month a trawler captain claimed his boat was nearly dragged down by a Russian submarine while fishing off the Scottish coast.
	TV-DPG
1	the Karen was towed 18 miles from Ardglass on the south-east shore of Northern Ireland. estranged boat's captain Paul Murphy was pictured carrying a typical cable on his boat. a Russian submarine may have caused the dramatic incident.
	JS-DPG
1	the Karen was towed at 10 knots during yesterday's incident 18 miles from Ardglass on the south-east shore of Northern Ireland. trawler is thought to have caused his boat to snagging backwards and was badly damaged following the incident. last month a trawler captain claimed his boat was nearly dragged down by a Russian submarine while fishing off the Scottish coast.

Table 8. Generation samples for summarization

source document c

Freddie Roach insisted on Saturday that Floyd Mayweather does not deserve to be ranked alongside Manny Pacquiao as the leading fighters of their generation as the two boxers put the finishing touches to their preparations for the Fight of the Century in Las Vegas a week on Saturday. Roach, Pacquiao's trainer, said he rated super-middleweight star Andre Ward and middleweight sensation Gennady Golovkin above Mayweather despite the American's unbeaten record and his status as hot favourite for the May 2 showdown against Pacquiao. 'Mayweather is undefeated so you have to give him a little credit for that,' said Roach, 'but he has picked and chosen his opponents and I don't think he's fought enough competition to be considered the best. You have to fight the best to be the best, I feel. He's ducked a lot of guys. Manny Pacquiao's trainer Freddie Roach says that Floyd Mayweather cannot be considered best ever . Roach says that Mayweather has picked and chosen his fights during his career . 'Manny has had some devastating losses in his career but he is a realist. He understands that losing is part of the game and if you don't think you are going to get knocked out in this sport, you have picked the wrong sport. I'd put Ward and Triple G above Mayweather right now. They are very talented guys and very polished boxers.' Roach has been vocal in an otherwise low-key and surprisingly respectful build-up to the welterweight showdown next month and knows that his tactics have made the stakes even higher for him. But he is adamant Pacquiao will spring a surprise in the most eagerly awaited fight for years. 'The fight will be won and lost on the ropes,' said Roach. 'If Mayweather goes to the ropes and tries to rest his legs, he will get beat. If he has good movement the entire night and his legs don't give out on him, he'll probably win. It's about outscoring him. If he sits on the ropes, we can outscore him. If he stays in the middle of the ring and boxes all the time he could possibly outscore us.' Roach claimed Pacquiao is in the best shape of his life, moving faster and punching harder than ever before. 'He trains harder than any fighter I have ever had in my life,' said Roach. 'I have had 33 world champions and nobody can touch him for his work ethic. His attitude is good. 'This fight can send out a message about what we need to do in boxing. When the best fights the best, do you see how big this is? Someone needs to wake up and put the best with the best all the time. Because when that happens, we have the best sport in the world. I don't care who likes who, you can still negotiate business. Roach rates super middleweight Andre Ward higher than he does undefeated Mayweather . Roach, who has trained Pacquiao for 15 years, says the Filipino is in the best shape of his life . 'It's a better fight today than it was five years ago because they were both a lot faster and more mobile five years ago and it might have been more of a boxing match but now they're a little bit older, it's going to be a better fight. 'I believe Manny can win this fight. I kind of have to win this fight. I've been talking a lot. It's more important than anything to me. It's more important to me than getting my girlfriend, Maya, back. It's that big because, I mean, I really like this girl. 'My mother told me, "You must like her; you put her Christmas tree up for her, you bought her a car and those earrings you bought her cost £14,000 each". 'After this fight maybe I'll try to get her back.'

$b(x, c)$	x
	KL-DPG
1	Freddie Roach, the trainer of Manny Pacquiao, says Floyd Mayweather is impossible to be considered best ever. he says he rated andre Ward and Gennady Golovkin above Mayweather despite the American's unbeaten record.
	TV-DPG
1	Floyd Mayweather is fighting for Manny Pacquiao in a series and is one of the top fighters of their generation. trainer Freddie Roach says he rated super-middleweight star Andre Ward and middleweight sensation Gennady Golovkin above Mayweather despite his unbeaten record and his celebrity status as hot favourite.
	JS-DPG
1	Freddie Roach says Floyd Mayweather can't be ranked alongside Manny Pacquiao as leading fighters of their generation. the american's trainer says that he rated super-middleweight star Andre Ward and middleweight sensation Gennady Golovkin above Mayweather.

Table 9. Generation samples for summarization

source document c

The son of a Labour councillor who was detained in Turkey after apparently trying to sneak into Syria with eight family members was seen grinning as he began his journey back to Britain. Waheed Ahmed, 21, who is the son of Rochdale politician Shakil Ahmed, was arrested with eight relatives – including four children – in a remote Turkish border town earlier this month. However, it is understood he is now returning to the UK and will fly from Dalaman into Manchester Airport later this evening. Scroll down for video . All smiles:\xa0Waheed Ahmed \xa0looks relaxed as he begins his journey back to the UK after being caught trying to sneak into Syria with eight family members . On the way home: The 21-year-old, sporting a shaved head, was filmed being escorted from a vehicle . Sky News reported that the remaining eight members of his family will remain in Turkey until Tuesday. The majority of the family flew to Turkey on March 27 from Manchester Airport and are accused of having plans to try and sneak across the border into Syria. Waheed did not fly out with his family but joined them three days later on a flight from Birmingham. Mohammed Shafiq, a friend of Waheed's father, said there were concerns about his behaviour in the months leading up to his arrest. He told Sky News: 'There were concerns in the last six months to a year about a change in his behaviour. 'And a change in his attitude towards various different issues. 'That was causing concern for people in the community and his family.' Earlier this month, Waheed's father spoke of his shock after being told that his son is suspected of being a militant Islamist. He said: 'All I know is that they were on holiday and then the next thing I am told is that they have been arrested.' Mr Ahmed was with his aunt, two cousins and one of their wives when they were stopped in Turkey, near the Syrian border . Waheed Ahmed, the 21-year-old son of Labour Councillor Shakil Ahmed, is understood to be returning to the UK on a flight to Manchester from Dalaman tonight following his arrest for allegedly trying to sneak into Syria . Waheed Ahmed, 21, \xa0is the son of Rochdale Labour councillor Shakil Ahmed (pictured above with Ed Miliband) The nine Britons, who include three men, two women and four children aged between one and 11, were seized in Hatay province, in southern Turkey. It shares a border with part of Syria controlled by rebel factions including those linked to Al Qaeda and ISIS. All of those held are from Rochdale and are the biggest family group caught attempting to enter the unstable territory. Counter terrorism officers at Greater Manchester Police began an investigation into their movements and the extended family group were detained at a checkpoint in Ogulpinar earlier this month. A senior officer questioned why anyone would take children so young 'and vulnerable' into a warzone. The three men and two women, aged between 21 and 47, were taken to a hospital with their children, aged one, three, eight and 11. Waheed (pictured) was detained in Turkey alongside his aunt, two cousins and one of his cousin's wives . The nine Britons - four of them children – were seized by Turkish security forces as they tried to slip into Syria . Officials said they would be photographed and fingerprinted before being deported back to the UK. At the time, photographs showed Waheed, dressed in traditional robes and wearing heavy boots, leading the group from a minibus into a police station. Several women, all wearing headscarfs which covered their faces, could be seen carrying children. Most of the party were wearing walking boots, perfect for trekking across the rugged region. Shakil Ahmed, a bakery delivery driver, is a councillor in Kingsway and served alongside Karen Danczuk, wife of Rochdale MP Simon Daczuk, until her resignation in January. Speaking as he delivered election campaign leaflets earlier this month, he said he recognised his son in online newspaper reports of his arrest. He said the others arrested included Waheed's aunt, Zadia Bi, 50, and two of her sons and one of their wives. He said: 'I don't know why they have been arrested. We have no information. Until they ring we will not know what has happened.' He said that he had seen his son's photograph and when asked about one picture of his son laughing, he replied: 'Well, they went on holiday so they shouldn't be crying on holiday should they?' He added: 'I don't believe my son was on his way to join Islamic State. I was shocked, worried and extremely upset to hear that my son has been arrested. During their arrest, the family were fingerprinted and taken to a police station where they have been held since . One of the family members, holding a child, is seen arriving at a Turkish hospital to undergo medical checks . The family are pictured arriving at a police station in Turkey's southern Hatay province earlier this month . 'It's a total mystery to me why he's there, as I was under the impression he was on a work placement in Birmingham. 'My son is a good Muslim and his loyalties belong to Britain. If I thought for a second that he was in danger of being radicalised, I would have reported him to the authorities.' The councillor added: 'He's studying a degree in politics and sociology at Manchester University and has a good future ahead of him. I just want to speak to my son and get him home as soon as possible.' Waheed apparently called his devastated father to break the news he had been arrested. Sorry we are not currently accepting comments on this article.

$b(x, c)$	x
	KL-DPG
1	waheed Ahmed, 21-year-old was arrested with eight relatives in a remote northern Turkish border town this month. it is understood he is now returning to the UK and will fly from Dalaman into Manchester Airport later this evening. majority of family flew to Turkey on march 27 and are accused of having plans to try and sneak across the border into Syria.
	TV-DPG
1	Waheed Ahmed, 21, is the son of Rochdale Labour councillor Shakil Ahmed. he was arrested with eight relatives in a remote border town earlier this month. he is now returning to the UK and will fly from Dalaman into Manchester Airport later this evening.
	JS-DPG
1	waheed Ahmed, 21, is the son of Rochdale politician. he was being held after apparently trying to sneak into Syria. he was arrested with eight relatives – including four children. but it is understood he is now returning to the UK.

Table 10. Generation samples for summarization