ALIGNING FOUNDATION MODELS FOR LANGUAGE WITH PREFERENCES THROUGH f-DIVERGENCE MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Aligning large pretrained language models with human preferences is fundamental for improving their capabilities and acceptability in downstream applications, an objective that 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. 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 are good for approximating different targets.

1 INTRODUCTION

Large pretrained language models, also known as “Foundation Models” for language, 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). Viewed through this lens, approaches to LM alignment can be organised along two axes: how the target distribution is constructed and how it is approximated. Khalifa et al. (2021) proposes a framework that they coin Generation with Distributional Control (GDC), by which they explicitly define the target distribution that represents the fully aligned LM in closed form, and then approximate it via methods such as distributional policy gradients (DPG; Parshakova et al., 2019), which minimizes the forward Kullback-Leibler (KL) divergence $\mathrm{KL}(p||\pi_\theta)$ of the LM π_θ to the target distribution p. On the other hand, even if RL with KL penalties (Todorov, 2006a; Kappen et al., 2012; Jaques et al., 2017; 2019), which forms the core of reinforcement learning from human feedback or 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 the LM to a target distribution that can also explicitly written 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 to fine-tuning an LM to approximate any given target distribution by following any divergence in the f-divergences family. f-DPG generalizes existing approximation techniques from 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. We show 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).

2 Formal aspects

2.1 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

$$D_f(p_1||p_2)\doteq\mathbb{E}_{x\sim p_2}\left[f\left(\frac{p_1(x)}{p_2(x)}\right)\right]+f^{'}(\infty)\,p_1(p_2=0) \tag{1}$$

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. 1 is set to 0 regardless of the value of $f^{'}(\infty)$ (which can be infinite).[2] It can be 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 KL$(p_1||p_2)$, reverse KL divergence 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).

2.2 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. 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: $\min_{\theta\in\Theta}D_f(\pi_\theta||p)$, where θ varies inside the parametric family Θ. The objective 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:*

$$\nabla_\theta D_f(\pi_\theta||p)=E_{x\sim\pi_\theta}\left[f^{'}\left(\frac{\pi_\theta(x)}{p(x)}\right)\nabla_\theta\log\pi_\theta(x)\right]. \tag{2}$$

Note that it may happen in Eq 2 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.[3]

[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. 1 simplifies to $D_f(p_1||p_2)=\mathbb{E}_{x\sim p_2}\left[f\left(\frac{p_1(x)}{p_2(x)}\right)\right]$.

[3]The derivative $f'(t)$ of any convex function $f(t)$ is defined almost everywhere. See also App. A.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. $\text{Supp}(\pi_\theta) = \mathcal{X}$) for all θ's, while the support of p might be strictly included in $\mathcal{X}$ in some experiments. We refer to the approach in Eq. 2 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 $\text{KL}(p||\pi_\theta)$ as discussed in Sec. 2.3.

2.3 Recovering some existing methods

GDC In GDC, Khalifa et al. (2021) propose a target distribution $p_{\text{GDC\_bin}}(x) \propto a(x)b(x)$, where a is a pretrained LM and $b(x) = 0$ if x contains a curse and $b(x) = 1$ otherwise. Fitting the policy π_θ to the target p is done using DPG Parshakova et al. (2019), namely by minimizing the **forward KL**, $\text{KL}(p||\pi_\theta)$. In the f-DPG framework, $\text{KL}(p||\pi_\theta) = D_f(\pi_\theta||p)$ with $f(t) = -\log t$, $f'(t) = -1/t$, and Thm. 1 leads to the equivalent objective: $\nabla_\theta D_f(\pi_\theta||p) = E_{x\sim\pi_\theta} - \frac{p(x)}{\pi_\theta(x)} \nabla_\theta \log \pi_\theta(x)$.

RL with KL penalties Let's set the target distribution as $p(x) \doteq p_{\text{RLKL}}(x) = 1/Z\ a(x)\ e^{r(x)/\beta}$, where Z is a normaliser. Then $\text{KL}(\pi_\theta||p) = D_f(\pi_\theta||p)$, with $f(t) = t\log t$ corresponding to **reverse KL**, and $f'(t) = 1 + \log t$. Thm. 1 implies that: $\nabla_\theta D_f(\pi_\theta||p) = E_{x\sim\pi_\theta}\left(-\frac{r(x)}{\beta} + \log\frac{\pi_\theta(x)}{a(x)}\right)\nabla_\theta \log \pi_\theta(x)$, where we have exploited the fact that $1 + \log Z$ is a constant, hence $E_{x\sim\pi_\theta}(1 + \log Z)\ \nabla_\theta \log \pi_\theta(x) = 0$. Up to the constant factor β, this form recovers the original formula for estimating the gradient of the loss : $\nabla_\theta J_{\text{RLKL}}(\theta) = E_{x\sim\pi_\theta}\left(r(x) - \beta\log\frac{\pi_\theta(x)}{a(x)}\right)\nabla_\theta \log \pi_\theta(x)$.

3 Experiments

Task We evaluate our method on two LM alignment tasks, namely, alignment with scalar preferences on positive sentiment and alignment with a binary preference on lexical content introduced by Khalifa et al. (2021). For the first one, we set the target distribution to $p_{\text{RLKL}}(x) \propto a(x)e^{r(x)/\beta}$, where $r(x) = \log\phi(x)$ and $\phi(x)$ is the probability returned by a sentiment classifier fine-tuned from Distil-BERT HF Canonical Model Maintainers (2022). (See App. E). 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. 2.3. For alignment with lexical constraint, we use two target distributions, namely $p_{\text{GDC\_bin}}(x) \propto a(x)b(x)$, with binary preference $b(x) = 1$ iff the target word appears in the sequence x, and a scalar preference target distribution p_{RLKL} where $r(x)$ is set in the same way as $b(x)$. 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" ($8\cdot 10^{-6}$). App. B.1 elaborates on the target distributions. We use four instantiations of f-DPG to approximate these targets, 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 measure approximation quality in terms of these same divergences. Note that $p_{\text{GDC\_bin}}(x) = 0$ when $b(x) = 0$, implying that reverse KL, namely $\text{KL}(\pi_\theta||p)$, becomes infinite, so we exclude it for this target. Implementation details and hyper-parameters are available in App. C.

Metrics The main metrics we report are: (1) $D_f(\pi_\theta||p)$, the f-divergence between p and π_θ, with four different f's corresponding to forward KL, $\text{KL}(p||\pi_\theta)$; reverse KL, $\text{KL}(\pi_\theta||p)$; Total Variation, $\text{TV}(\pi_\theta||p)$; and Jensen-Shannon, $\text{JS}(\pi_\theta||p)$, estimated by importance sampling, (2) $\text{KL}(\pi_\theta||a)$, a measure of the divergence from original LM a Ziegler et al. (2019); Khalifa et al. (2021), (3) 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, $\text{std}(r_\theta(x))$, with pseudo-rewards $r_\theta(x) \doteq -f'\left(\frac{\pi_\theta(x)}{p(x)}\right)$.

Results Fig. 1 shows the evolution of the above-mentioned metrics. For lexical constraints, we show aggregated evolution of the metrics. Further details and disaggregated results are given in App. F. 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 connected 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. In alignment with scalar preferences (Fig. 1 (a)) 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. Yet, notably, all other variants of f-DPG minimize all four divergences. RKL-DPG yields the highest value of $E_{\pi_\theta}[\phi(x)]$ 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 and produces the spike at the beginning of training in $\mathrm{std}(\mathrm{rewards})$. In lexical constraints (Fig. 1 (b) and (c)), strinklingly, the original KL-DPG is outperformed by 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. We also observe (Fig. 1 (a) and (c)) that RKL-DPG tends to produce distributions with lower normalized entropy.

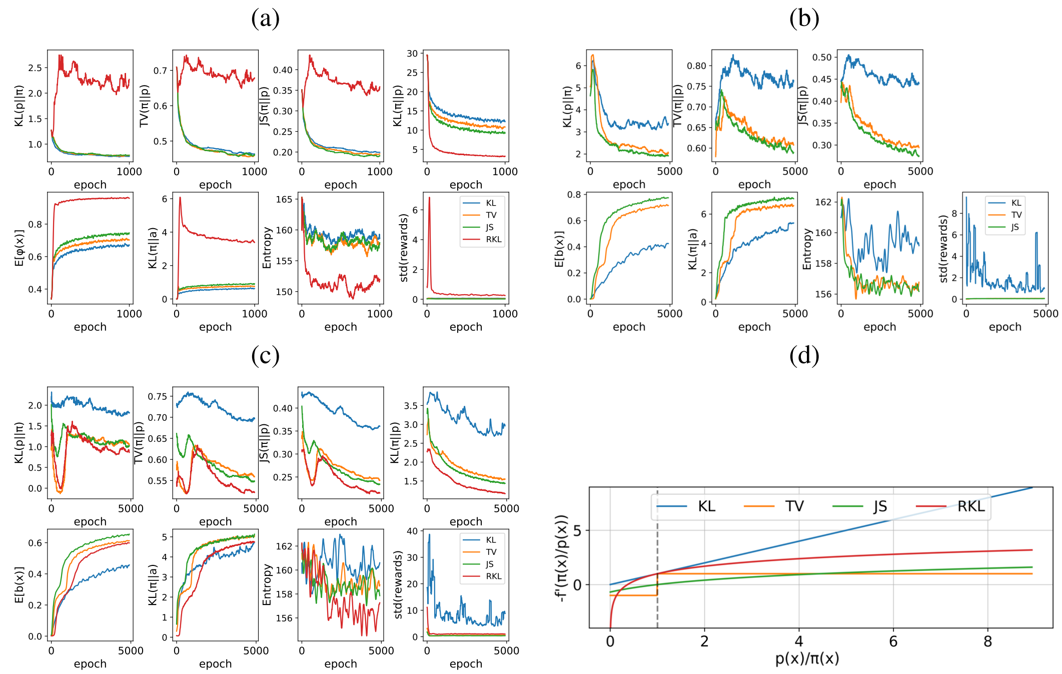

Figure 1: Comparison of f-DPG on (a) sentiment preference, (b) lexical constraint with GDC framework, and (c) lexical constraint with RL with KL penalties framework. Evaluation metrics: four f-divergences $D_f(\pi_\theta||p)$ ($\downarrow$ better), $E_{\pi_\theta}[\phi(x)]$ ($\uparrow$ better), Entropy ($\uparrow$ better), standard deviation of pseudo-reward $\mathrm{std}(r_\theta(x))$. (d) 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 $p(x) = \pi_\theta(x)$.

4 Discussion and conclusion

Our experiments show that the choice of the divergence measure can have a significant impact on the resulting model's quality, although there is not a single best divergence across distributions. However, interestingly, for a given target there is one or a few variants that are the best across all measured divergences even in terms of divergences that they do not directly optimize for. Fig. 1 (d) illustrates the differences between pseudo-rewards for distinct f-divergences. The forward KL loss aims to ensure coverage of the subset where $p(x) > 0$, giving a large pseudo-reward for samples with $p(x) >> \pi(x)$, while the optimization can be sensitive to sampling noise in the finite sample approximation (Fig. 1 (b) and (c)). 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 (Fig. 1 (a)). 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.

To conclude, we propose a flexible framework for approximating a target distribution by minimizing any f-divergence, unifying earlier approaches for aligning LM's. Our results on a diverse array of tasks show that minimizing well-chosen f-divergences leads to significant gains over previous work.

REFERENCES

Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (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.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021. URL https://arxiv.org/abs/2112.00861.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. In *Proc. of ICLR*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJDaqqveg.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022a. URL https://arxiv.org/abs/2204.05862.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, 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.

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

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Proc. of NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Yang Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 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.

Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. In *Proc. of ICLR*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJKkY35le.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 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.

Bryan Eikema, Germán Kruszewski, Christopher R Dance, Hady Elsahar, and Marc Dymetman. An approximate sampler for energy-based models with divergence diagnostics. *Transactions of Machine Learning Research*, 2022. URL https://openreview.net/forum?id=VW4IrC0n0M.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 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.

Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura (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.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Sona Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63 (11):139–144, 2020. ISSN 0001-0782. doi: 10.1145/3422622. URL https://doi.org/10.1145/3422622.

Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 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.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*, volume 305. Springer science & business media, 2013.

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

Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, José Miguel Hernández-Lobato, Richard E. Turner, and Douglas Eck. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. In Doina Precup and Yee Whye Teh (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.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Àgata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind W. Picard. 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.

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

Hilbert J. Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. In Daniel Borrajo, Subbarao Kambhampati, Angelo Oddi, and Simone Fratini (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.

Liyiming Ke, Sanjiban Choudhury, Matt Barnes, Wen Sun, Gilwoo Lee, and Siddhartha S. Srinivasa. Imitation learning as f-divergence minimization. In Steven M. LaValle, Ming Lin, Timo Ojala, Dylan A. Shell, and Jingjin Yu (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.

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

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

Tomasz Korbak, Hady Elsahar, German Kruszewski, and Marc Dymetman. Controlling conditional language models without catastrophic forgetting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (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.

Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Proc. of NeurIPS*, 2022b. URL https://openreview.net/forum?id=XvI6h-s4un.

Tomasz Korbak, Ethan Perez, and Christopher L. Buckley. 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.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 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.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 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.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 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.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 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.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes, 2022. URL https://arxiv.org/abs/2203.11147.

Lars M. Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In Jennifer G. Dy and Andreas Krause (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.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 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.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013. URL http://arxiv.org/abs/1312.5602.

Mohammad Norouzi, Samy Bengio, Zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. Reward augmented maximum likelihood for neural structured prediction. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Proc. of NeurIPS*, pp. 1723–1731, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/2f885d0fbe2e131bfc9d98363e55d1d4-Abstract.html.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Proc. of NeurIPS*, pp. 271–279, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/cedebb6e872f539bef8c3f919874e9d7-Abstract.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Proc. of NeurIPS*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

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

Ramakanth Pasunuru and Mohit Bansal. 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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Proc. of NeurIPS*, pp. 8024–8035, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

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

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In Zoubin Ghahramani (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.

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

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. Cold decoding: Energy-based constrained text generation with langevin dynamics. *ArXiv preprint*, abs/2202.11705, 2022. URL https://arxiv.org/abs/2202.11705.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In Yoshua Bengio and Yann LeCun (eds.), *Proc. of ICLR*, 2016. URL http://arxiv.org/abs/1511.06732.

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

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 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.

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

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

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *ArXiv preprint*, abs/2206.05802, 2022. URL https://arxiv.org/abs/2206.05802.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with natural language feedback. *ArXiv preprint*, abs/2204.14146, 2022. URL https://arxiv.org/abs/2204.14146.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In Yoshua Bengio and Yann LeCun (eds.), *Proc. of ICLR*, 2016. URL http://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv preprint*, abs/1707.06347, 2017. URL https://arxiv.org/abs/1707.06347.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, 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.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Proc. of NeurIPS*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J. Martin, Animesh Mehta, Brent Harrison, and Mark O. Riedl. Controllable neural story plot generation via reward shaping. In Sarit Kraus (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.

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

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. Lamda: Language models for dialog applications. *ArXiv preprint*, abs/2201.08239, 2022. URL `https://arxiv.org/abs/2201.08239`.

Emanuel Todorov. Linearly-solvable markov decision problems. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann (eds.), *Proc. of NeurIPS*, pp. 1369–1376. MIT Press, 2006a. URL `https://proceedings.neurips.cc/paper/2006/hash/d806ca13ca3449af72a1ea5aedbed26a-Abstract.html`.

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

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. *ArXiv preprint*, abs/2112.04359, 2021. URL `https://arxiv.org/abs/2112.04359`.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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`.

Jieyu Zhao, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Kai-Wei Chang. 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`.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (eds.), *Proc. of SIGIR*, pp. 1097–1100. ACM, 2018. doi: 10.1145/3209978.3210080. URL `https://doi.org/10.1145/3209978.3210080`.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *ArXiv preprint*, abs/1909.08593, 2019. URL `https://arxiv.org/abs/1909.08593`.

Daniel M. Ziegler, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Ben Weinstein-Raun, Daniel de Haas, Buck Shlegeris, and Nate Thomas. 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. 1 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*

$$D_f(p||q) \doteq \sum_{\{x:\ p(x)>0,\ q(x)>0\}} q(x)\ f(\frac{p(x)}{q(x)}) \quad + f(0)\ q(p=0) \quad + f^*(0)\ p(q=0), \tag{3}$$

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 $f^{**} = f$. 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:

$$f(0) \doteq \lim_{t\to 0} f(t), \quad f^*(0) = \lim_{t\to 0} f^*(t) = \lim_{t\to 0} t\ f(\frac{1}{t}), \tag{4}$$

$$0\ f(0) \doteq 0, \quad 0\ f^*(0) \doteq 0, \qquad \text{including when } f(0) = \infty \text{ and } f^*(0) = \infty, \tag{5}$$

$$f'(\infty) \doteq f^*(0) = \lim_{t\to 0} t\ f(\frac{1}{t}). \tag{6}$$

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 1 and 3 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:

$$\mathrm{KL}(p||\pi) = \begin{cases} \mathbb{E}_{x\sim p}\ \log \frac{p(x)}{\pi(x)} & \text{if } \mathrm{Supp}(p) \subset \mathrm{Supp}(\pi), \\ \infty, & \text{otherwise.} \end{cases} \tag{7}$$

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 equation 3:

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

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

Now, consider the "reverse KL" from π to p, namely $\mathrm{KL}(\pi||p)$. Based on the previous derivation, and with the same $f(t) = -\log t$ we can write it as $\mathrm{KL}(\pi||p) = D_f(p||\pi)$, 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 $\mathrm{TV}(p||\pi) = \frac{1}{2}\ \sum_{x\in\mathcal{X}} |p(x) - \pi(x)|$. We then have $\mathrm{TV}(p||\pi) = \mathrm{TV}(\pi||p)$. 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 equation 3:

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

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

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) $\forall\theta \in \Theta, \mathrm{Supp}(p) \subset \mathrm{Supp}(\pi_\theta)$, or (ii) $\mathrm{Supp}(\pi_\theta)$ does not depend on θ. Then:*

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

Proof. Based on definition equation 3 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:

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

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

A.4 About non-differentiability of f

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). Furthermore, in practice when sampling from π_θ in Eq 2, 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 $T_{nd} \doteq \{t : f(t)$ 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 2 is well-defined for all x's.

A.5 f-DPG algorithm

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

A.6 Baseline: alternative derivation

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

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

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*

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

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 Background and related work

B.1 Distributional approach in LMs

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.

B.1.1 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) such that $p(x) \propto P(x)$. $P(x)$ (and $p(x)$ after normalization) can be used to score samples, but not to directly obtain 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:

$$p_{\text{GDC\_bin}}(x) \propto a(x)b(x), \tag{9}$$

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 $a(x)$:

$$J_{\text{RLKL}}(\theta) = \mathbb{E}_{x \sim \pi_\theta}\left[r(x) - \beta \log \frac{\pi_\theta(x)}{a(x)}\right]. \tag{10}$$

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:

$$p_{\text{RLKL}}(x) \propto a(x) \exp(r(x)/\beta), \tag{11}$$

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

$$p_{\text{GDC\_dist}}(x) \propto a(x) \exp\left[\sum_i \lambda_i \phi_i(x)\right], \tag{12}$$

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 $E_{x\sim p}[\phi_i(x)]$ 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 $\text{KL}(p_{\text{GDC\_dist}}||a)$.

B.1.2 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, $\text{CE}(p, \pi_\theta)$:

$$\nabla_\theta \text{CE}(p, \pi_\theta) = -\mathbb{E}_{x\sim\pi_\theta} \frac{p(x)}{\pi_\theta(x)} \nabla_\theta \log \pi_\theta(x). \tag{13}$$

B.2 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)

B.3 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 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).

C IMPLEMENTATION DETAILS

C.1 ADDITIONAL TECHNIQUES FOR f-DPG

Adding a baseline 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 $E_{x\sim\pi_\theta}[r(x)]$, one computes the gradient of the loss using the formula $\nabla_\theta E_{x\sim\pi_\theta}[r(x)] = E_{x\sim\pi_\theta}[r(x)\nabla_\theta \log \pi_\theta(x)]$. The gradient in Eq. 2 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). 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 $\nabla_\theta D_f(\pi_\theta||p) = E_{x\sim\pi_\theta} r_\theta(x)\nabla_\theta \log \pi_\theta(x) = E_{x\sim\pi_\theta}(r_\theta(x) - B)\,\nabla_\theta \log \pi_\theta(x)$, based on the observation that $E_{x\sim\pi_\theta}\,\nabla_\theta \log \pi_\theta(x) = 0$ (see also App. A.6).

Fact 2. *Subtracting B from $r_\theta(x)$ does not introduce bias into f-DPG gradient estimates.*

$D_f(\pi_\theta\|\|p)$	f	f'	$f'\left(\frac{\pi_\theta(x)}{p(x)}\right)$	$f'(\infty)$
Forward KL (KL($p\|\|\pi_\theta$))	$f(t) = -\log t$	$f'(t) = -\frac{1}{t}$	$-\frac{p(x)}{\pi_\theta(x)}$	0
Reverse KL (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 (TV($\pi_\theta\|\|p$))	$f(t) = 0.5\,\|1-t\|$	$f'(t) = \begin{cases} 0.5 & \text{for } t > 1 \\ -0.5 & \text{for } t < 1 \end{cases}$	$\begin{cases} 0.5 & \text{for } \frac{\pi_\theta(x)}{p(x)} > 1 \\ -0.5 & \text{for } \frac{\pi_\theta(x)}{p(x)} < 1 \end{cases}$	0.5
Jensen-Shannon (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, KL($p||\pi_\theta$), corresponds to $D_{-\log t}(\pi_\theta||p)$, and similarly for the reverse KL, 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: TV($\pi_\theta||p$) = TV($p||\pi_\theta$), JS($\pi_\theta||p$) = JS($p||\pi_\theta$).

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`

Table 2: Hyperparameters used throughout all experiments

Typically, B is chosen to be the average of the rewards, $B \doteq E_{x\sim\pi_\theta}[r_\theta(x)]$. In the experiments of Sec. 3, we use the baseline technique where B is an estimate of the average of pseudo-rewards, unless otherwise specified.

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. **??**.

C.2 HYPER PARAMETERS AND PACKAGES

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. Since KL-DPG was particularly sensitive to the learning rate for the most experiments, we searched for the optimal learning rate based on KL-DPG performance and applied it to all other f-DPG models with different losses. We use an exponential moving average baseline (Sec. C.1) 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 use a pretrained GPT-2 "small" Radford et al. (2019) with 117M parameters for the initial model.

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. 3). Entropy (↑ better), Self-BLEU-5 (↓ better), Distinct-1 (↑ better), and Perplexity (↓ better).

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. 3.

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	$E[b(x)]$	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. 3). $E_{\pi_\theta}[b(x)]$ (↑ better), Self-BLEU-5 (↓ better), Distinct-1 (↑ better), and Perplexity (↓ better).

E 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:

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

and then learning the reward model by minimizing the loss

$$\begin{aligned} \mathrm{loss}(r) &= \mathbb{E}_{(x_1,...,x_n)\sim\mathcal{D}}\, \mathrm{CE}(h, f_r) && (14) \\ &= -\mathbb{E}_{(x_1,...,x_n)\sim\mathcal{D}}\, h(x_1, ..., x_n) \cdot \log f_r\,(x_1, ..., x_n)\,, && (15) \end{aligned}$$

Thus, the optimal reward model is given by the function r such that $h(x_1, ..., x_n) = f_r(x_1, ..., x_n)$ as it minimizes the CE in Eq. 14. 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 $h_\phi(x_1, ..., x_n) \sim \text{Categorical}(\phi(x_1), ..., \phi(x_n))$, or in other words,

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

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

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

we get

$$\begin{aligned} f_{r_\phi}(x_1, ..., x_n) &= \text{softmax}(\log(\phi(x_1)), ..., \log(\phi(x_n))) && (16) \\ &= \text{categorical}(\phi(x_1), ..., \phi(x_n)), && (17) \end{aligned}$$

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

F Additional figures

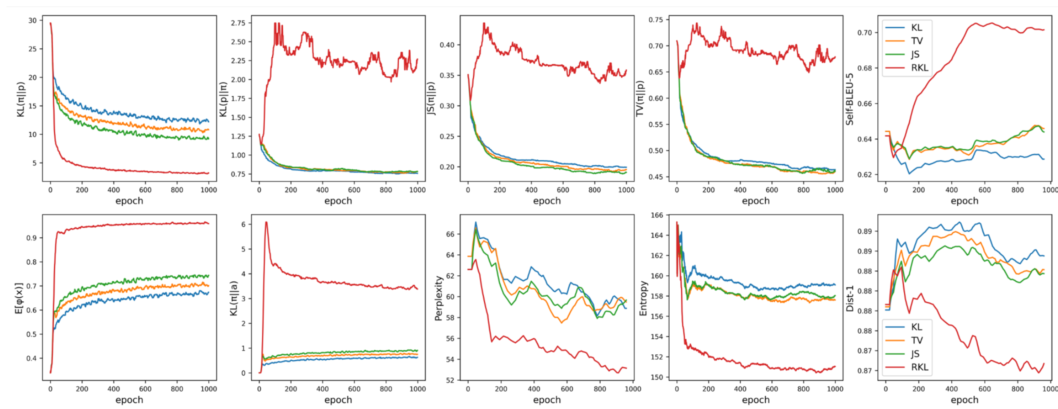

Figure 2: Evaluation of metrics in sentiment preference

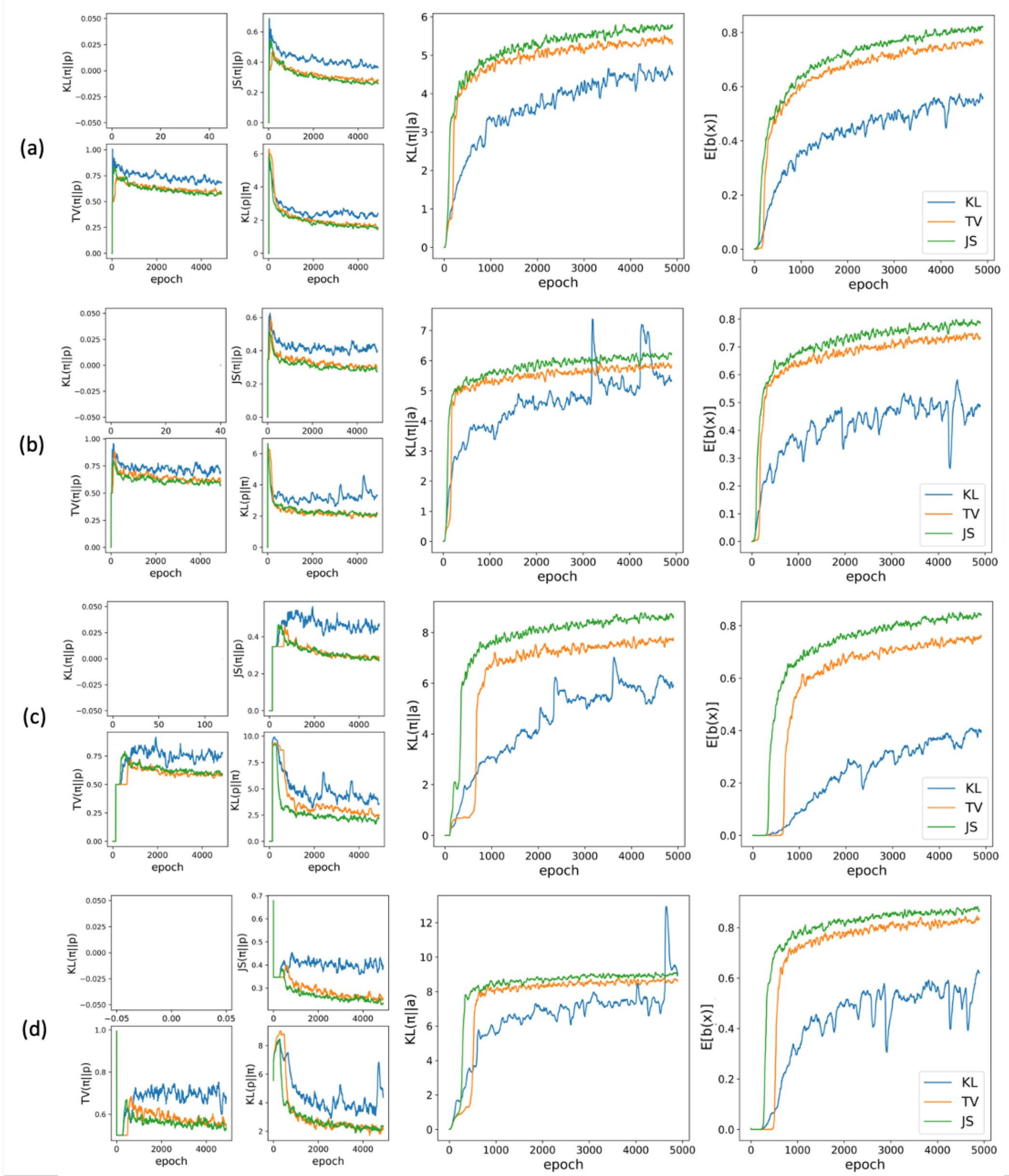

Figure 3: Evaluation metrics: four f-divergences $D_f(\pi_\theta||p)$ (↓ better), $E_{\pi_\theta}[\phi(x)]$ (↑ better), KL$(\pi_\theta||a)$ (↓ better) with target distribution induced from GDC framework to constrain the existence of single word, (a) amazing, (b) restaurant, (c) amusing, (d) Wikileaks. Note that reverse KL cannot be defined in this case in which $p(x) = 0$ for some points

G SAMPLES

$\phi(x)$	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 ˆˆˆ 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