On Reinforcement Learning and Distribution Matching for Fine-Tuning Language Models with no Catastrophic Forgetting

Tomasz Korbak∗
University of Sussex
tomasz.korbak@gmail.com

Hady Elsahar
Naver Labs Europe
hady.elsahar@gmail.com

Germán Kruszewski
Naver Labs Europe
german.kruszewski@naverlabs.com

Marc Dymetman†
marc.dymetman@gmail.com

Abstract

The availability of large pre-trained models is changing the landscape of Machine Learning research and practice, moving from a “training from scratch” to a “fine-tuning” paradigm. While in some applications the goal is to “nudge” the pre-trained distribution towards preferred outputs, in others it is to steer it towards a different distribution over the sample space. Two main paradigms have emerged to tackle this challenge: Reward Maximization (RM) and, more recently, Distribution Matching (DM). RM applies standard Reinforcement Learning (RL) techniques, such as Policy Gradients, to gradually increase the reward signal. DM prescribes to first make explicit the target distribution that the model is fine-tuned to approximate. Here we explore the theoretical connections between the two paradigms, and show that methods such as KL-control developed for RM can also be construed as belonging to DM. We further observe that while DM differs from RM, it can suffer from similar training difficulties, such as high gradient variance. We leverage connections between the two paradigms to import the concept of baseline into DM methods. We empirically validate the benefits of adding a baseline on an array of controllable language generation tasks such as constraining topic, sentiment, and gender distributions in texts sampled from a language model. We observe superior performance in terms of constraint satisfaction, stability and sample efficiency.

1 Introduction

Pre-trained language models (Devlin et al., 2019; Radford et al., 2019) are changing the landscape of Machine Learning research and practice. Due to their strong generative capabilities many studies have found it sufficient to “nudge” these models to conform to global preferences defined over the generated sequences instead of training from scratch using annotated data. These preferences could include topic and sentiment (Dathathri et al., 2020), valid musical notes and molecular structures (Jaques et al., 2017a), code compilability (Korbak et al., 2021), reducing distributional biases (Khalifa et al., 2021; Weidinger et al., 2021), evaluation metrics for Machine Translation and Summarization (Ranzato et al., 2016; Bahdanau et al., 2016), or direct human feedback (Ziegler et al., 2019; Stiennon et al., 2020). This large body of studies is driven by two different paradigms: Reward Maximization (RM) and Distribution Matching (DM).

∗Work partly done during an internship at Naver Labs Europe.
†Independent Researcher. Work done at Naver Labs Europe.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).
Problem: aligning a pretrained language model with human preferences

Figure 1: In this study we make a connection between two popular paradigms for aligning language models to human preferences: Reward maximization (RM) and Distribution matching (DM).

Reward Maximization  RM intuitively nudges pre-trained models towards certain preferences by providing global sequence-level rewards when the model generates outputs that satisfy desired features. For instance, if the model is producing toxic content, we can apply Reinforcement Learning (RL) techniques to discourage it from producing similar content. However, naively applying RL yields a model that can undergo catastrophic forgetting of its original distribution. For example, it can degenerate into producing a single nonsensical but at least nontoxic sequence. Although several studies have considered hand-crafting general rewards to ensure desirable features like fluency (Liu et al., 2016a; Tambwekar et al., 2019), coming up with complete or perfect rewards is highly non-trivial (Wu et al., 2016; Vedantam et al., 2015). This has sparked a wide discussion on the overall effectiveness of RM for some tasks such as machine translation (Choshen et al., 2020; Kiegeland & Kreutzer, 2021).

Reward Maximization with KL-Control  To tackle the aforementioned issues of “catastrophic forgetting”, several studies, still under an RM paradigm, have considered incorporating a distributional term inside the reward to be maximized. In particular Jaques et al. (2017b, 2019) and Ziegler et al. (2019) or more recently Stiennon et al. (2020), Ouyang et al. (2022), Bai et al. (2022) and Perez et al. (2022) have applied variations of KL-control (Todorov, 2007; Kappen et al., 2012) 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., 2017a) as in Ziegler et al. (2019) or Bai et al. (2022) or Q-learning (Mnih et al., 2013) as in Jaques et al. (2017b). Adding this distributional KL-penalty to the reward raises some important questions: What effect does it have on the shape of the optimal policy? Does this new objective have any interpretation from a distributional perspective?

Distribution Matching  A different recent paradigm for fine-tuning language models to satisfy downstream preferences formulates the problem as Distribution Matching (DM). This paradigm consists of two steps: first a target distribution incorporating the desired preferences is defined as an Energy-Based Model (LeCun et al., 2006). Then the forward KL divergence is minimized between this target distribution and an auto-regressive policy using a family of algorithms referred to as Distributional Policy Gradients (DPG) (Parshakova et al., 2019b; Khalifa et al., 2021; Korbak et al., 2021, 2022a). This approach capitalizes on the flexibility of EBMs in specifying the target distribution. For example, the EBM can be defined so that it conforms to all downstream preferences while its corresponding normalized distribution has a minimal KL divergence from the original, pretrained language model, therefore tackling the problem of “catastrophic forgetting” (Khalifa et al., 2021). Interestingly, this DM paradigm can also deal with distributional preferences, for instance, for de-biasing language models by specifying that the generated sequences should be gender-balanced,
i.e. that 50% of generations contain female mentions. Such distributional constraints cannot be defined in the RM paradigm where a reward is calculated for a single sequence.

We can notice the promises and limitations of these two paradigms for fine-tuning language models. RM approaches are equipped with an arsenal of RL algorithms and optimization techniques that can be efficient in reward maximization, however they lack the distributional aspect to avoid catastrophic forgetting and impose distributional preferences over LMs. DM approaches are suited to tackle those limitations, however, the family of DPG algorithms currently used is not as rich as its RL counterpart.

While the connections between these two seemingly distinct paradigms have been noted (Parshakova et al., 2019b; Korbak et al., 2022b), they have not been explored in detail. Clarifying such connections might help import ideas from one approach to the other. This is our goal in this paper, detailing the nuanced connections and applying them to a case-study in variance reduction. Overall, our contributions are the following:

- We clarify relations between the RM and DM paradigms through a detailed comparison between the family of DPG algorithms and Policy Gradients (Table 1), stressing the differences between parametric and non-parametric rewards that are important in this regard.
- We introduce an interpretation of KL-control techniques from a distribution matching perspective, placing such techniques at an intermediate place between RM and DM (Theorem 1).
- We show how these connections can enable cross-pollination between the two perspectives by applying baselines — a variance reduction technique from RL — to DPG and derive a particular choice of a baseline (Facts 1 and 2). On an array of controllable language generation experiments, we show that adding baselines leads to superior performance on constraint satisfaction (Figure 3), stability on small batch sizes, and sample efficiency (Figure 4).

2 Background

**Standard Policy Gradients** One popular method for adapting the behaviour of language models to certain preferences has been that of assigning a “reward” score \( R(x) \) for sequences \( x \) sampled from an autoregressive language model (policy) \( \pi_\theta \). Then, the simplest policy gradient algorithm in reinforcement learning, namely, REINFORCE [Williams, 1992a], aims to find the policy \( \pi_\theta(x) \) that maximizes the average reward \( \mathbb{E}_{x \sim \pi_\theta} R(x) \), and this leads, via the so-called “log derivative trick”, to a gradient ascent algorithm that iteratively samples \( x \) from \( \pi_\theta \) and update parameters by increments proportional to \( R(x) \nabla_\theta \log \pi_\theta(x) \) via the following identity:

\[
\nabla_\theta \mathbb{E}_{x \sim \pi_\theta} R(x) = \mathbb{E}_{x \sim \pi_\theta} R(x) \nabla_\theta \log \pi_\theta(x),
\]

(1)

**KL-control** (Todorov, 2007; Kappen et al., 2012), was leveraged by Jaques et al. (2017b, 2019) and Ziegler et al. (2019) to include a KL penalty term in the reward function to penalize large deviations from the original pretrained model \( a(x) \), weighted by a free hyperparameter \( \beta \) to control the trade-off between the two goals. That is, they maximize the expectation \( \mathbb{E}_{x \sim \pi_\theta} R^\beta_\theta(x) \), where:

\[
R^\beta_\theta(x) = r(x) - \beta \log \frac{\pi_\theta(x)}{a(x)}.
\]

(2)

**Distributional Policy Gradients** (DPG) [Parshakova et al., 2019b] is a recent approach used to fit an autoregressive policy \( \pi_\theta \) to the distribution \( p(x) = P(x)/Z \) induced by the EBM \( P(x) \), where \( Z = \sum x P(x) \) is the normalization constant (partition function). Given an arbitrary EBM \( P(x) \), DPG optimizes the loss function \( D_{KL}(p, \pi_\theta) \) with respect to the parameters \( \theta \) of an autoregressive model \( \pi_\theta \), a loss which is minimized for \( \pi_\theta = p \). The KL-divergence minimization objective leads to a gradient estimate of the form:

\[
\nabla_\theta D_{KL}(p, \pi_\theta) = -\nabla_\theta \mathbb{E}_{x \sim p} \log \pi_\theta(x)
\]

(3)

\[
= -\sum_x p(x) \nabla_\theta \log \pi_\theta(x) = -\frac{1}{Z} \sum_x P(x) \nabla_\theta \log \pi_\theta(x)
\]

(4)

\[
= -\frac{1}{Z} \mathbb{E}_{x \sim \pi_\theta} \frac{P(x)}{\pi_\theta(x)} \nabla_\theta \log \pi_\theta(x).
\]

(5)
3 Reward Maximization vs Distribution Matching

In the previous section, we have summarized three approaches that have been suggested for fine-tuning language models. Two of them can be characterized as “Reward Maximization” (RM): Standard Policy Gradients (PG) and KL-control. On the other hand, DPG clearly belongs to the realm of “Distribution Matching” (DM) as it first defines the target distribution and then optimizes a policy to match it. In the rest of this section, we will explore connections between these two seemingly distinct concepts and, in the following section, we will exploit them to improve DM-based methods.

3.1 Standard vs. Parametric Rewards

Let us start with distinguishing between a “parametric reward” $R_{\theta}$ which depends on $\theta$ and a standard reward $R$, which does not. If we wished to maximize the expected parametric reward, $E_{\pi_\theta} R_{\theta}(x)$, we would follow its gradient, leading to the identities:

$$
\nabla_{\theta} E_{x \sim \pi_\theta} R_{\theta}(x) = \nabla_{\theta} \sum_x \pi_\theta(x) R_{\theta}(x) = \sum_x \pi_\theta(x) \nabla_{\theta} R_{\theta}(x) + \sum_x R_{\theta}(x) \nabla_{\theta} \pi_\theta(x) \tag{6}
$$

Equation (6) is the sum of two terms: the first one, the “RG-term” (Reward Gradient term), involves the gradient of the reward. The second one, the “PG-term” (Policy Gradient term), was obtained using the “log derivative trick” and involves the gradient of the policy stricto sensu. In standard RL, where the reward does not depend on $\theta$, the RG-term disappears and the gradient of expected reward consists solely of the PG-term. However, when $R_{\theta}$ depends on $\theta$, the gradients are distinct (apart from specific cases where the RG-term evaluates to 0, as we will see below).

3.2 KL-control as Distribution Matching

Adding a KL-penalty term to the reward (as in the case of KL-control) leads to a parametric reward. However, due to the particular form of its objective, the RG-term actually vanishes\footnote{This is because $E_{\pi_\theta} \nabla_{\theta} R_{\theta \theta}(x) = -\beta E_{\pi_\theta} \nabla_{\theta} \log \pi_\theta(x) = 0$, via the identity $E_{\pi_\theta} \nabla_{\theta} \log \pi_\theta(x) = \sum_x \nabla_{\theta} \pi_\theta(x) \log \pi_\theta(x) + \nabla_{\theta} \sum_x \pi_\theta(x) = 0$.} leaving only the PG-term $E_{x \sim \pi_\theta} R_{\theta}(x) \nabla_{\theta} \log \pi_\theta(x)$ and simplifying the tuning procedure to a standard Policy Gradient. While this algorithm falls under the RM paradigm, here we argue that its nature is multifaceted, and explore deeper connections with the DM paradigm. More precisely, the maximization of reward with the KL penalty term is equivalent to a distributional matching with an underlying emergent sequential EBM, a remark that already reveals some similarities with DPG\footnote{The optimal policy $p_z$ is briefly mentioned in (Ziegler et al., 2019) without reference or derivation. The proof, which reveals a connection to the reverse KL divergence from $\pi_\theta$, is ours.}

**Theorem 1.** Consider the following EBM:

$$
P_z(x) = a(x) e^{r(x)/\beta} \tag{9}
$$

and let $p_z$ be the normalized distribution $p_z(x) = \frac{1}{Z} P_z(x)$, with $Z = \sum_x P_z(x)$. Then:

(i) $\arg \max_{\pi_\theta} E_{x \sim \pi_\theta} R_{\theta}(x) = \arg \min_{\pi_\theta} D_{KL}(\pi_\theta, p_z)$;

(ii) $\arg \max_{\pi \in \mathcal{D}(X)} E_{x \sim \pi} R_{\pi}(x) = p_z$, where $\mathcal{D}(X)$ is the family of all distributions over $X$, and $R_{\pi}(x) \doteq r(x) - \beta \log \frac{\pi(x)}{a(x)}$.

**Proof.** A simple way to prove this is to notice that the expectation of the reward $R_{\theta}$ has a monotonically decreasing relationship with the reverse KL divergence between $\pi_\theta$ and $p_z$:

$$
D_{KL}(\pi_\theta, p_z) = E_{x \sim \pi_\theta} \log \frac{\pi_\theta(x)}{p_z(x)} = E_{x \sim \pi_\theta} \left[ \log \pi_\theta(x) - \log \frac{1}{Z} a(x) e^{r(x)/\beta} \right]
$$

\[\]
While the objective of DPG (distribution matching) is different from that of Policy Gradients (reward with an emergent underlying distribution), DPG gradient estimates suffer from the same high-variance problems as with standard PG.

Nonetheless, the analogy behind this gradient term is more fruitful than it first appears. As a matter of fact, DPG gradient estimates suffer from the same high-variance problems as with standard PG.

However, the DM objective of DPG does indeed provide a DM interpretation, namely in terms of minimizing the reverse KL divergence of the other EBM.

In the previous subsection, we have connected KL-control, a method designed under a RM paradigm, to DM. Now, we turn to the converse question of whether DPG, a DM method, can be connected to RM.

Overall, we can conclude that the addition of the distributional term (KL-penalty) to the reward does indeed provide a DM interpretation, namely in terms of minimizing the reverse KL divergence of the other EBM.

3.3 Similarities and Differences between DPG and Policy Gradients

In the previous subsection, we have connected KL-control, a method designed under a RM paradigm, to DM. Now, we turn to the converse question of whether DPG, a DM method, can be connected to RM. We begin by noting that after defining $R_\theta = \frac{P(x)}{\pi_\theta(x)}$, the DPG gradient $\mathbb{E}_{x \sim \pi_\theta} \frac{P(x)}{\pi_\theta(x)} \nabla_\theta \log \pi_\theta(x)$ acquires the format of the PG-term $E_{x \sim D(X)} R_\theta^z(x)$ without a clear meaning. By contrast, the DPG algorithms are designed to perform DM on any EBM specification, corresponding to an explicit distributional objective.

However, the DM objective of DPG cannot be considered as maximizing the average “reward” $R_\theta(x) = \frac{P(x)}{\pi_\theta(x)}$, as this would require adding also the RG-term $\mathbb{E}_{x \sim \pi_\theta} \frac{P(x)}{\pi_\theta(x)} \nabla_\theta \log \pi_\theta(x)$ to the gradient, which in this case does not vanish.

Nonetheless, the analogy behind this gradient term is more fruitful than it first appears. As a matter of fact, DPG gradient estimates suffer from the same high-variance problems as with standard PG. While the objective of DPG (distribution matching) is different from that of Policy Gradients (reward maximization), DPG also needs to estimate the PG-term $E_{x \sim \pi_\theta} R_\theta^z(x)$ at a given value of $\theta$, using a batch of samples $x$. For such a fixed $\theta$, we can define provisionally set $R(x) \equiv R_\theta$ and the problem of gradient estimation for this fixed $\theta$ is identical to the estimation $\mathbb{E}_{x \sim \pi_\theta} \frac{P(x)}{\pi_\theta(x)} \nabla_\theta \log \pi_\theta(x)$ based on a set of samples $x$ in standard RL. Therefore, techniques that have been developed to reduce the variance of the gradient estimates in RL can be ported to DPG insofar as we are computing the gradient estimates at a given $\theta$. In Section 4, we show how one can import one such variance reduction technique to the DPG: baselines.

4 A Case Study on Variance Reduction

Baselines are a standard variance reduction technique in the context of Policy Gradients. The idea is to subtract from the reward $R(x)$ a value $B$ that does not introduce bias to the gradients but may change variance. After the introduction of baseline, equation (1) then takes the following form:

$$\nabla_\theta E_{x \sim \pi_\theta} R(x) = E_{x \sim \pi_\theta} [R(x) - B] \nabla_\theta \log \pi_\theta(x).$$

Table 1: A comparison between Policy Gradients (Sutton et al., 1999) and Distributional Policy Gradients (Parshakov et al., 2019) forms of Reward, Baseline, and Gradient of the loss function (the PG-term) before ($\nabla_\theta$) and after ($\nabla_\theta$ with Baseline) including a baseline for variance reduction.

|                | Policy Gradients | DPG |
|----------------|-----------------|-----|
| Reward         | $R(x)$          | $R_\theta(x) = \frac{P(x)}{\pi_\theta(x)}$ |
| $\nabla_\theta$ | $\mathbb{E}_{x \sim \pi_\theta} R(x) \nabla_\theta \log \pi_\theta(x)$ | $\mathbb{E}_{x \sim \pi_\theta} \frac{P(x)}{\pi_\theta(x)} \nabla_\theta \log \pi_\theta(x)$ |
| Baseline       | $\mathbb{E}_{x \sim \pi_\theta} R(x)$ | $Z$ |
| $\nabla_\theta$ with Baseline | $\mathbb{E}_{x \sim \pi_\theta} [R(x) - \mathbb{E}_{x \sim \pi_\theta} R(x)] \nabla_\theta \log \pi_\theta(x)$ | $\mathbb{E}_{x \sim \pi_\theta} \left[ \frac{P(x)}{\pi_\theta(x)} - Z \right] \nabla_\theta \log \pi_\theta(x)$ |
In standard RL, the simplest form of baseline $B$ is just the average of the rewards for the policy:

$$B^{RL} = E_{x \sim \pi_{\theta}} R(x). \quad (12)$$

Following the same methodology of taking the baseline to be the expectation of the reward term, we can obtain a remarkably simple form of a baseline for DPG:

$$B = E_{x \sim \pi_{\theta}} \frac{P(x)}{\pi_{\theta}(x)} = \sum_{x} \pi_{\theta}(x) \frac{P(x)}{\pi_{\theta}(x)} = \sum_{x} P(x) = Z. \quad (13)$$

**Fact 1.** Subtracting $B$ from $R_{\theta}(x)$ does not introduce bias into DPG gradient estimates.

**Proof.** Let us rewrite the DPG gradient in (5) with the added baseline $B = Z$:

$$E_{x \sim \pi_{\theta}} \left[ R_{\theta}(x) - Z \right] \nabla_{\theta} \log \pi_{\theta}(x) = E_{x \sim \pi_{\theta}} R_{\theta}(x) \nabla_{\theta} \log \pi_{\theta}(x) - Z E_{x \sim \pi_{\theta}} \nabla_{\theta} \log \pi_{\theta}(x)$$

$$= E_{x \sim \pi_{\theta}} R_{\theta}(x) \nabla_{\theta} \log \pi_{\theta}(x) - Z \left[ \sum_{x} \nabla_{\theta} \log \pi_{\theta}(x) \right] \quad (14)$$

Here, the second term does not introduce bias because $Z \left[ \sum_{x} \nabla_{\theta} \log \pi_{\theta}(x) \right] = 0$, leaving us with the exact same form of gradient as in the original DPG algorithm. \qed

Note that since $B^{RL}$ depends on $\theta$, it has to be re-estimated after each gradient update. On the other hand, $B$ does not depend on $\theta$, which is an advantage because $B$ could be now estimated by averaging over samples from all the different $\theta$’s without introducing bias, leading to a more accurate estimation. See Table 1 for a comparison of these two forms of baselines.

The off-policy DPG version introduced in (Parshakova et al., 2019b) and its KL-adaptive variant (Khalifa et al., 2021) sample a proposal distribution $q$ instead of the policy $\pi_{\theta}$. Then, the baseline takes the form

$$B^{off}(x) = \frac{Z \pi_{\theta}(x)}{q(x)}, \quad (15)$$

where the $\frac{Z \pi_{\theta}(x)}{q(x)}$ term is an importance weight correcting for the bias introduced by sampling from $q$. Similarly to the DPG case, we can prove the following (see Appendix C).

**Fact 2.** Subtracting $B^{off}(x)$ from $R_{\theta}(x)$ does not bias the off-policy DPG gradient estimates.

In practice, as shown on Figure 2 adding a baseline to KL-adaptive DPG (Algorithm 1) centers the advantage (defined as $A = \frac{P(x)}{q(x)} - \frac{Z \pi_{\theta}(x)}{q(x)}$) around 0 leading to better performance on: convergence (section 4.3), stability on small batch sizes (section 4.4), and variance reduction (section 4.5).

---

While this baseline is not optimal (proof Appendix C1), it is widely used in practice.

In the scope of this paper, our focus is on importing to DPG simple constant baselines. The advantage is that this is a technique that is not impacted by the fact that $R_{\theta}$ depends on $\theta$: it can be applied “$\theta$-locally” to provide a more accurate estimate of $E_{x \sim q_{\theta}} R_{\theta}(x) \nabla_{\theta} \log \pi_{\theta}(x)$ for a fixed $\theta$, irrespective of the values of $R_{\theta}$ elsewhere, while variance reduction techniques that involve several $\theta$’s simultaneously raise additional challenges for parametric rewards.

---

Algorithm 1 KL-Adaptive DPG with baseline

**Require:** $P$, initial generative model $a$

1: $\pi_{\theta} \leftarrow a, q \leftarrow a$

2: for each iteration do

3: for each episode do

4: sample $x$ from $q_{\theta}(\cdot)$

5: $\theta \leftarrow \theta + \alpha(\theta) \left[ \frac{P(x)}{q(x)} - Z \frac{\pi_{\theta}(x)}{q(x)} \right] \nabla_{\theta} \log \pi_{\theta}(x)$

6: if $D_{KL}(p||\pi_{\theta}) < D_{KL}(p||q)$ then

7: $q \leftarrow \pi_{\theta}$

**Ensure:** $\pi_{\theta}$

---

Figure 2: Values of reward, advantage and the baseline for first 1000 epochs of a point-wise constraint experiment.
4.1 Generation with Distributional Control

We investigate the benefits of adding a baseline to the DPG algorithm, on the Generation with Distributional Control (GDC) [Khalifa et al., 2021] framework. GDC makes use of DPG to control the properties of pre-trained language models to satisfy certain constraints. In our experiments, we follow the target distribution form of Parshakova et al. (2019a), Khalifa et al. (2021) and Korbak et al. (2022a), in which the EBM $P(x)$ is defined so that its normalized variant $p(x)$ matches a set of desired moments constraints on given features $\phi_i(x)$, while having a minimal KL divergence $D_{KL}(p,a)$ from an original pretrained language model $a$, to avoid catastrophic forgetting.

These constraints are expressed as conditions $\mu_i = \mathbb{E}_{x \sim p}\phi_i(x)$, for $i \in \{1, \ldots, n\}$, by which the moments (expectations) under the distribution $p$ of each feature $\phi_i(x)$ are required to take certain desired values $\bar{\mu}_i$. For instance, let $\phi_1(x) = 1$ iff the topic of $x$ is science and $\phi_2(x) = 1$ iff $x$ mentions a female person, then imposing moments $\bar{\mu}_1 = 1$ and $\bar{\mu}_2 = 0.5$ constrains the language model $p$ to only generate sequences about science, half of which mention females. $P(x)$ is uniquely determined by the following form\footnote{For a more precise formulation of this EBM, see Khalifa et al., 2021.}

$$P(x) = a(x)e^{\sum_{i=1}^{n} \lambda_i \phi_i(x)},$$

where $\lambda_i$ terms control the moments $\mu_i$ of the associated features, which can be estimated through self-normalized importance sampling (Owen, 2013), and then, to make the moments match the desired values, the $\lambda_i$ terms can be optimized through SGD (Parshakova et al., 2019a).

4.2 Experimental setup

We evaluate our method on an array of 10 controlled text generation tasks. For each, given a pre-trained language model $a(x)$, and a set of constraints, the objective of each fine-tuning method is to obtain a fine-tuned language model $\pi_\theta$ that satisfies the imposed constraints while deviating as minimally as possible from the original language model $a(x)$.

Constraints are defined as a set of binary features $\{\phi_i\}$ and their corresponding desired percentages (moments) $\{\mu_i\}$ within the generations of the target language model. Based on the value of the moment constraints these 10 tasks are divided into 6 tasks of pointwise constraints (for which $\mu_i = 1$), 2 tasks of distributional constraints ($0 < \mu_i < 1$) and 2 tasks of mixed type constraints (hybrid):

(a) Single-word constraints, where $\phi(x) = 1$ iff the a given word appears in the sequence $x$. We experiment with frequent words (task 1: “amazing”, original frequency: $10^{-5}$) and (task 2: “WikiLeaks”, original frequency: $10^{-5}$) rare words,

(b) Wordlist constraints, where $\phi(x) = 1$ iff $x$ contains at least one word from a given list. We consider lists of word associated with politics (task 3) and science (task 4) published by Dathathri et al. (2020),

(c) Sentiment classifier constraints, where $\phi(x) = 1$ if $x$ is classified as positive (task 5), or negative (task 6) by a pre-trained classifier published by Dathathri et al. (2020),

(d) A single distributional constraint where $\phi(x) = 1$ iff $x$ contains a female figure mention, and $\bar{\mu}_i = 0.5$ (task 8),

(e) A set of four distributional constraints: $\phi_1(x) = 1$ iff $x$ contains at least one of the words in the “science”, “art”, “sports” and “business” wordlists (compiled by Dathathri et al. (2020)), respectively. For each $i$, $\bar{\mu}_i = 0.25$ (task 8),

(f) Hybrid constraints where $\phi_1(x) = 1$ iff $x$ contains more female than male pronouns, $\bar{\mu}_1 = 0.5$ and $\phi_2(x) = 1$ iff $x$ contains at least one of the words from the “sports” wordlist (task 9) or “politics” wordlist, $\bar{\mu}_2(x) = 1$ (task 10).

Methods We modify the GDC framework Khalifa et al. (2021), namely its KL-DPG algorithm, to include a baseline as shown in Algorithm 1. We refer to this method as GDC++. In addition to comparing GDC++ with GDC we compare with two reward maximization baselines: Reinforce (Williams, 1992b) and Ziegler (Ziegler et al., 2019). Reinforce tries to maximize the expected reward $\mathbb{E}_{x \sim \pi} R(x)$, while $R(x) = 1$ if and only if the pointwise constraints are met. Ziegler instantiates the KL-control approach: its objective includes a KL penalty term for departures from $a$. Following Khalifa et al. (2021), for hybrid and distributional constraints (tasks 8-10) we compare
Results

We present the evolution of our metrics through training epochs in Figure 3 (aggregated over tasks 1–6) and Figure 6 in the Appendix (aggregated over tasks 7–10). Results for each task are presented separately on Figures 7–10 in the Appendix.

Consistent with prior work (Khalifa et al., 2021; Korbak et al., 2022a), we observe that Reinforce is able to quickly achieve high levels of constraint satisfaction, but at the cost of large deviations from $a$, which translates into significantly decreased diversity of generated samples (in terms of Self-BLEU-5 and Distinct-1). The KL penalty term in Ziegler imposes an upper bound on deviation from $a$, but the deviation is still significant enough to result in a drop in diversity. Moreover, we have observed Ziegler’s objective to result in very unstable training.

GDC and GDC++ are the only fine-tuning methods that address constraint satisfaction based on a clear formal objective, i.e. reducing the divergence from $p$. The approach translates into significantly smaller deviations from $a$ and maintaining diversity within and across samples. The addition of a baseline indeed reduces the variance. We analyze that extensively in Appendix 4.5 while here focusing on the downstream effects of variance reduction. One is that $\pi_\theta$ is now able to compound staying closer to $p$ and $a$ at the same time, while achieving slightly better constraint satisfaction. We have also observed that baseline stabilizes training, leading to smoother curves.

4.4 The effect of baseline across batch sizes

We expect that reducing gradient estimates variance can allow to train with lower batch sizes, performing gradient updates on estimates based on smaller batch sizes can increase the sample variance of the gradients which can lead to unstable training. Therefore, we study the effect of baseline across batch sizes by performing gradient updates on estimates based on smaller batch sizes.
efficiency. To test this, we rerun tasks 1 (a pointwise constraint on the word “amazing”) and 8 (distributional constraints on topics) with four batch sizes (256, 512, 1024, 2048). Figures 4a and 4b show the benefits of adding a baseline — higher constraint satisfaction, lower divergence from $p$, more stable training — and is especially evident with lower batch sizes. For instance, with batch size 256, GDC++ obtains a significantly higher constraint satisfaction rate and lower divergence from $p$.

Furthermore, stable training with smaller batch sizes translates into better sample efficiency. For instance, in task 1 (Figure 4a), GDC++ with batch size 256 needs 1M samples to achieve $E_{x \sim \pi} \phi(x) = 0.5$ while GDC++ with batch size 2048 needs 4M. In contrast, GDC with batch size 256 does not achieve $E_{x \sim \pi} \phi(x) = 0.5$ at all, confirming the importance of adding the baseline.

### 4.5 Empirical Evaluation of Variance Reduction

Next, we evaluate empirically the effect of the baseline for variance reduction. We select two tasks: task 1 (a pointwise constraint) and task 7 (distributional constraints) described in Section 4.2, each with 3 different seeds, while monitoring the following variance measures:

**Gradient Variance** The gradient estimate is defined as: $G_{\theta(x)} \equiv A(x) \nabla_{\theta} \log \pi_{\theta}(x)$, where $G_{\theta(x)} \in \mathbb{R}^{l}$ is an unbiased estimate of the gradient of the forward KL loss $\nabla_{\theta} D_{KL}(p, \pi_{\theta})$ with respect to the parameters $\theta$. We then have, with $\mu(G_{\theta}) \equiv E_{x \sim q} G_{\theta(x)}$

\[
\text{Var}(G_{\theta}) \equiv E_{x \sim q} \|G_{\theta(x)} - \mu(G_{\theta})\|_2^2 \\
= E_{x \sim q} \|G_{\theta(x)}\|_2^2 - \|\mu(G_{\theta})\|_2^2.
\]

**Variance of the advantage** is defined by:

\[
\text{Var}(A) \equiv E_{x \sim q} \|A(x) - \mu^A\|_2^2
\]

where, $\mu^A \equiv E_{x \sim q} A(x)$ is the mean of the advantage, which we showed above to be null after the addition of the baseline.

**Expected absolute value of the advantage** This metric is defined as:

\[
\mu|A| \equiv E_{x \sim q} |A(x)|
\]

Figure 5: Comparison between GDC and GDC++ using a set of Variance diagnosis metrics on pointwise and distributional constraints experiments.
It directly provides a standard measure of distributional discrepancy between \( p \) and \( \pi_\theta \), in terms of TVD (Total Variation Distance). We have:

\[
\mathbb{E}_{x \sim q} \left| \frac{p(x)}{q(x)} - \frac{\pi_\theta(x)}{q(x)} \right| = 2 \text{TVD}(p, \pi_\theta).
\]  

Results Figure 5 shows that GDC++ obtains lower variance in the gradient estimates \( \text{Var}(G_\theta) \) and the variance of the advantage \( \text{Var}(A) \) in both pointwise and distributional experiments compared to its non-baseline counterpart GDC. We further observe a decreasing trend in the mean absolute value of the advantage \( |\mu| \) which is correlated with a decreasing trend in the TVD distance between the trained policy \( \pi_\theta \) and the optimal distribution \( p \). Overall, these results shows that adding a baseline to DPG reduces the variance during training and yields better convergence towards the optimal distribution \( p \).

5 Related work

The idea of posing control problems as distribution matching has resurfaced numerous times in the RL literature (Kappen et al., 2012; Friston et al., 2010; Levine, 2018; Hafner et al., 2020; Buckley et al., 2017). KL-control can be seen as a generalisation of maximum entropy RL (MaxEnt RL) (Haarnoja et al., 2017; 2018) to informed priors. If in (2) we chose \( a(x) \) to be a uniform distribution (assuming right now finiteness of \( X \)) instead of a pretrained LM distribution, then the KL penalty \( D_{KL}(\pi_\theta, a) \) would reduce to an entropy bonus. Both KL-control and MaxEnt RL can be derived from a general framework of control-as-inference (Levine, 2018) which poses control as minimising KL from a certain target distribution. However, most practical algorithms in the MaxEnt RL family minimise KL from a target policy which changes throughout training; in contrast, DPG’s target distribution \( p \) and KL-control implicit target distribution \( p_z \) are defined at trajectory level and fixed throughout training.

Perhaps the closest method to KL-control and DPG in the larger family of inference-based RL (Furuta et al., 2021) is AWR (Peng et al., 2019) which minimises the forward KL from an off-policy target distribution. Yet another approach with apparent similarity to KL-control and DPG is state marginal matching (SMM) (Hazan et al., 2018; Lee et al., 2019). SMM poses exploration as learning a policy that induces a state marginal distribution that matches a target state distribution. While SMM’s target distribution is fixed, it is defined for individual states, while in the controllable language generation tasks we consider, the target distribution is defined over a complete trajectory considered as a unit. See Appendix B for an extended discussion of related work.

6 Conclusion

Fine-tuning large language models has become an active area of research, due to its importance in adapting large language models to satisfy task-level preferences, or in combating their social risks such as “distributional” stereotyping (Weidinger et al., 2021; Welbl et al., 2021). In this paper, we analyzed in depth the nuanced relation between two popular fine-tuning paradigms: RM and DM. We demonstrated that KL-control can be seen as a form of DM and showed that while DPG and PG have different goals, some similarities (similar forms of gradient estimates despite different objectives) can be exploited. We used these insights to inform an extension of DPG, consisting in adding a baseline to reduce the variance of gradient estimates.

The connections we established suggest that despite fundamental differences between DPG and RL, some of the theoretical results and algorithmic techniques from RL can be adapted to a DM framework without losing their formal guarantees. In this paper, we focus on variance reduction using baselines, but the space of possible enhancements is vast. Promising candidates include further reducing the variance using a learned value function (Konda & Tsitsiklis, 2000) and preventing detrimentally large policy updates by maintaining a trust region in the policy space – akin to techniques such as TRPO (Schulman et al., 2015) and PPO (Schulman et al., 2017b). Another future direction could consist in analyzing the relation between explicit EBMs in DPG and implicit EBMs arising in KL-control and characterizing the space of EBMs that could be reached through KL-control.

See Appendix A for a discussion of broader impacts of large language models and controllable language generation.
References

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally Normalized Transition-Based Neural Networks. 2016. doi: 10.18653/v1/P16-1231.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. An Actor-Critic Algorithm for Sequence Prediction. (2015):1–17, 2016. URL http://arxiv.org/abs/1607.07086

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 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. 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 Eltarget, 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, 2022.

A. Bakhtin, Y. Deng, S. Gross, Myle Ott, Marc’Aurelio Ranzato, and Arthur Szlam. Energy-based models for text. ArXiv, abs/2004.10188, 2020.

David Belanger and Andrew McCallum. Structured prediction energy networks. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML’16, pp. 983–992. JMLR.org, 2016. URL http://dl.acm.org/citation.cfm?id=3045390.3045495

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of “bias” in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://www.aclweb.org/anthology/2020.acl-main.485.

Christopher L. Buckley, Chang Sub Kim, Simon McGregor, and Anil K. Seth. The free energy principle for action and perception: A mathematical review. Journal of Mathematical Psychology, 81:55–79, 2017. ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2017.09.004. URL https://www.sciencedirect.com/science/article/pii/S0022249617300962.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. Language gans falling short. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=BJgza6VtPB.

Leshem Choshen, Lior Fox, Zohar Aizenbud, and Omri Abend. On the weaknesses of reinforcement learning for neural machine translation. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. URL https://openreview.net/forum?id=HleCw3EKvH.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Pre-training transformers as energy-based cloze models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pp. 285–294. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.20. URL https://doi.org/10.18653/v1/2020.emnlp-main.20.
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 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1edEyBKDS

Peter Dayan. Reinforcement comparison. In Proceedings of the 1990 Connectionist Models Summer School, pp. 45–51. Morgan Kaufmann, San Mateo, CA, 1990.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc’Aurelio Ranzato. Residual energy-based models for text generation. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1l4SgHKDH

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Karl J Friston, Jean Daunizeau, James Kilner, and Stefan J Kiebel. Action and behavior: a free-energy formulation. Biological cybernetics, 102(3):227–260, 2010.

Hiroki Furuta, Tadashi Kozuno, Tatsuya Matsushima, Yutaka Matsuo, and Shixiang Shane Gu. Co-adaptation of algorithmic and implementational innovations in inference-based deep reinforcement learning, 2021. URL https://arxiv.org/abs/2103.17258.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. J. Mach. Learn. Res., 5:1471–1530, December 2004. ISSN 1532-4435.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington (eds.), Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, volume 9 of Proceedings of Machine Learning Research, pp. 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL http://proceedings.mlr.press/v9/gutmann10a.html

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In Doina Precup and Yee Whye Teh (eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pp. 1352–1361. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/haarnoja17a.html

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018. URL https://arxiv.org/abs/1801.01290.

Danijar Hafner, Pedro A. Ortega, Jimmy Ba, Thomas Parr, Karl Friston, and Nicolas Heess. Action and perception as divergence minimization, 2020.

Elad Hazan, Sham M. Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration, 2018. URL https://arxiv.org/abs/1812.02690.

Tianxing He, Bryan McCann, Caiming Xiong, and Ehsan Hosseini-Asl. Joint energy-based model training for better calibrated natural language understanding models. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 1754–1761, Online, April 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.eacl-main.151

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. Neural Comput., 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018. URL https://doi.org/10.1162/089976602760128018.
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.), Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of Proceedings of Machine Learning Research, pp. 1645–1654. PMLR, 2017a. URL http://proceedings.mlr.press/v70/jaques17a.html

Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, Jose Miguel Hernandez Lobato, Richard E. Turner, and Doug Eck. Tuning recurrent neural networks with reinforcement learning. 2017b. URL https://openreview.net/pdf?id=Syyv2e-Kx

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. CoRR, abs/1907.00456, 2019. URL http://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.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A distributional approach to controlled text generation. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=jWkw45-9AbL

Samuel Kiegeland and Julia Kreutzer. Revisiting the weaknesses of reinforcement learning for neural machine translation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tammo Chakraborty, and Yichao Zhou (eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pp. 1673–1681. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.133. URL https://doi.org/10.18653/v1/2021.naacl-main.133

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

A.S. Klyubin, D. Polani, and C.L. Nehaniv. Empowerment: a universal agent-centric measure of control. In 2005 IEEE Congress on Evolutionary Computation, volume 1, pp. 128–135 Vol.1, 2005. doi: 10.1109/CEC.2005.1554676.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller (eds.), Advances in Neural Information Processing Systems, volume 12. MIT Press, 2000. URL https://proceedings.neurips.cc/paper/1999/file/6449f44a102fde848669bdd9eb6676fa-Paper.pdf

Tomasz Korbak, Hady Elsahar, Marc Dymetman, and Germán Kruszewski. Energy-based models for code generation under compilability constraints. CoRR, abs/2106.04985, 2021. URL https://arxiv.org/abs/2106.04985

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.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp. 11499–11528. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/korbak22a.html

Tomasz Korbak, Ethan Perez, and Christopher L Buckley. RL with KL penalties is better viewed as Bayesian inference, 2022b. URL https://arxiv.org/abs/2205.11275

Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, pp. 1203–1213. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/d16-1128. URL https://doi.org/10.18653/v1/d16-1128
Yann LeCun, Sumit Chopra, Raia Hadsell, Marc’Aurelio Ranzato, and Fu Jie Huang. A Tutorial on Energy-Based Learning. In Predicting Structured Data. MIT Press, 2006.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. 2019.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review, 2018.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 110–119, San Diego, California, June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://www.aclweb.org/anthology/N16-1014

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In Jian Su, Xavier Carreras, and Kevin Du (eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, pp. 1192–1202. The Association for Computational Linguistics, 2016b. doi: 10.18653/v1/d16-1127. URL https://doi.org/10.18653/v1/d16-1127

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models, 2021.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael 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 Jian Su, Xavier Carreras, and Kevin Du (eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, pp. 2122–2132. The Association for Computational Linguistics, 2016a. doi: 10.18653/v1/d16-1230. URL https://doi.org/10.18653/v1/d16-1230

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Optimization of image description metrics using policy gradient methods. CoRR, abs/1612.00370, 2016b. URL http://arxiv.org/abs/1612.00370

Beren Millidge, Alexander Tschantz, Anil Seth, and Christopher Buckley. Understanding the origin of information-seeking exploration in probabilistic objectives for control, 2021.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In Eric P. Xing and Tony Jebara (eds.), Proceedings of the 31st International Conference on Machine Learning, volume 32 of Proceedings of Machine Learning Research, pp. 1791–1799, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/mnih14.html

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

Subhajit Naskar, Pedram Rooshenas, Simeng Sun, Mohit Iyyer, and A. McCallum. Energy-based reranking: Improving neural machine translation using energy-based models. ArXiv, abs/2009.13267, 2020.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, 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, 2022. URL https://arxiv.org/abs/2203.02155

Art B. Owen. Importance Sampling. In Monte Carlo theory, methods and examples, chapter 9. 2013. URL https://statweb.stanford.edu/~owen/mc/Ch-var-is.pdf

Tetiana Parshakova, Jean-Marc Andreoli, and Marc Dyentman. Global Autoregressive Models for Data-Efficient Sequence Learning. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pp. 900–909, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/K19-1084. URL https://www.aclweb.org/anthology/K19-1084
Tetiana Parshakova, Jean-Marc Andreoli, and Marc Dymetman. Distributional Reinforcement Learning For Energy-Based Sequential Models. CoRR, 2019b. URL https://arxiv.org/abs/1912.08517.

Ramakanth Pasunuru and Mohit Bansal. Reinforced video captioning with entailment rewards. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pp. 979–985. Association for Computational Linguistics, 2017. doi: 10.18653/v1/d17-1103. URL https://doi.org/10.18653/v1/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 Kopf, 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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (eds.), Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. URL https://openreview.net/forum?id=HkAClQgA-.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2019. URL https://arxiv.org/abs/1910.00177.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models, 2022. URL https://arxiv.org/abs/2202.03286.

Jan Peters and Stefan Schaal. Reinforcement learning of motor skills with policy gradients. Neural Networks, 21(4):682–697, 2008. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2008.02.003. URL https://www.sciencedirect.com/science/article/pii/S0893608008000701 Robotics and Neuroscience.

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

Marc’Aurelio Ranzato, Y-Lan Boureau, Sumit Chopra, and Yann LeCun. A unified energy-based framework for unsupervised learning. In Marina Meila and Xiaotong Shen (eds.), Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March 21-24, 2007, volume 2 of JMLR Proceedings, pp. 371–379. JMLR.org, 2007. URL http://proceedings.mlr.press/v2/ranzato07a.html.

Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In Joshua Bengio and Yann LeCun (eds.), 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016. URL http://arxiv.org/abs/1511.06732.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In International conference on machine learning, pp. 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. CoRR, abs/1707.06347, 2017a. URL http://arxiv.org/abs/1707.06347.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint: 1707.06347, 2017b.
David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. Artificial Intelligence, 299:103535, 2021. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103535. URL https://www.sciencedirect.com/science/article/pii/S0004370221000862

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from human feedback. CoRR, abs/2009.01325, 2020. URL https://arxiv.org/abs/2009.01325.

Richard S. Sutton. Temporal credit assignment in reinforcement learning. PhD thesis, University of Massachusetts, 1984.

Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In Proceedings of the 12th International Conference on Neural Information Processing Systems, NIPS ’99, pp. 1057–1063, Cambridge, MA, USA, 1999. MIT Press.

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.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, pp. 5982–5988. ijcai.org, 2019. doi: 10.24963/ijcai.2019/829. URL https://doi.org/10.24963/ijcai.2019/829

Emanuel Todorov. Linearly-solvable markov decision problems. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), Advances in Neural Information Processing Systems, volume 19, MIT Press, 2007. URL https://proceedings.neurips.cc/paper/2006/file/d806ca13ca3449af72a1ea5aedbed26a-Paper.pdf

Lifu Tu, Richard Yuanzhe Pang, Sam Wiseman, and Kevin Gimpel. Engine: Energy-based inference networks for non-autoregressive machine translation. ArXiv, abs/2005.00850, 2020.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pp. 4566–4575. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://doi.org/10.1109/CVPR.2015.7299087

Lex Weaver and Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, UAI’01, pp. 538–545, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

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. CoRR, abs/2112.04359, 2021. URL https://arxiv.org/abs/2112.04359

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in detoxifying language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, pp. 2447–2469. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.210. URL https://doi.org/10.18653/v1/2021.findings-emnlp.210
Ronald J. Williams. Reinforcement-learning connectionist systems. Technical report, Northeastern University, 1987. Technical Report NU-CCS-87-3.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256, 1992a. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Machine Learning*, pp. 229–256, 1992b.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. URL http://arxiv.org/abs/1910.03771

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL http://arxiv.org/abs/1609.08144

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.), *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, 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. *CoRR*, abs/1909.08593, 2019. URL http://arxiv.org/abs/1909.08593
Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix A
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In Appendix C we present proofs of all mathematical facts referred to in the paper.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code is included as supplementary material available to the reviewers and area chairs and will be made publicly available alongside the camera ready version of the paper.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In Appendix E we provide the hyperparameters used throughout our experiments and report our hardware configuration. In Appendix D we describe in detail how $D_{KL}(p, \pi_\theta)$ and $TVD(p, \pi_\theta)$ were estimated and provide an extended pseudocode for our training loop in Algorithm 2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] However, we found the variance across random seeds to be negligible and not comparing across random seeds is a standard practice when working with large language models where the cost of a single run is significant.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In Appendix E we report our hardware configuration.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] In Appendix E
   (b) Did you mention the license of the assets? [Yes] In Appendix E
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]
A Broader impacts

The focus area of this paper — fine-tuning large language models — is aligned with an important line of work on addressing the problem of social bias in large language models (Sheng et al., 2019; Liang et al., 2021). As the training data for large language models consists mainly of crawled user-generated content, a number of factors (from crawling methodology to Internet participation inequalities and moderation practices) leads to an over-representation of certain viewpoints and voices exceeding their prevalence in the general population. This poses a risk of amplifying biases and harms through a language model perpetuating these voices (Bender et al., 2021; Blodgett et al., 2020; Sheng et al., 2019; Weidinger et al., 2021; Welbl et al., 2021). Numerous problems related to addressing data bias in language generation (e.g. controlling for gender distribution in generated texts) can be naturally posed as generative distributional control (GDC), the framework we focus our experiments on. The distributional character of these data bias problems lies in the fact that desirable properties of generated texts are defined for a collection of samples, not only for individual samples. Our theoretical analyses of reward maximization and distribution matching approaches as well as our algorithmic improvements to the GDC framework — termed GDC++ — are therefore also a contribution to the problem of bias in language models. However, we need to be aware that GDC++, KL-control as well as controllable language generation techniques in general, can also be diverted to malicious uses such as spreading misinformation or generating harmful content.

B Extended Related Work

Reinforcement learning for language generation  Most previous attempts at steering language models to conform to global constraints defined over entire sequences have employed reinforcement learning. This includes using Reinforce (Williams, 1992a) for machine translation (Ranzato et al., 2016), actor critic (Konda & Tsitsiklis, 2000) for abstractive summarization (Paulus et al., 2018), caption generation (Liu et al., 2016b), dialogue (Li et al., 2016b), and video captioning (Pasunuru & Bansal, 2017). Some approaches (for instance, in machine translation and summarization (Ranzato et al., 2016; Bahdanau et al., 2017)) directly optimize performance metrics such as BLEU and ROUGE at training time. Others use heuristic rewards (for instance Li et al. (2016b) for dialogue generation and Tambwekar et al. (2019) for story generation) in order to obtain certain a priori desirable features of generated sequences that then incentivize good performance on target metrics. Catastrophic forgetting is a frequent problem of these fine-tuning approaches: reward maximization happens at the expense of large deviations from the original model. This problem is sometimes addressed by imposing a penalty term to the rewards, such as the KL divergence between the trained policy and the auto-regressive model. This approach, termed “conservative fine-tuning”, was applied to generating melodies with music theory rewards and organic molecules with synthesizability rewards by Jaques et al. (2017a) as well fine-tuning language models for controllable language generation by Ziegler et al. (2019). This solution often has hard time balancing between the reward term and the KL penalty term, leading to instability in training (Khalifa et al., 2021; Korbak et al., 2022a). Unlike this approach, KL-DPG determines an optimal distribution that satisfies both requirements.

RM and DM objectives in control problems  While RM is the dominant approach to tackling control problems (Sutton & Barto, 2018) and is sometimes argued to be sufficient for any intelligent behavior (Silver et al., 2021), prior work explored the benefits of alternative objectives formulated as DM: minimizing divergence from some target distribution \( p \). Prominent examples of (families of) DM objectives include control state marginal matching (Lee et al., 2019) active inference (Friston et al., 2010; Buckley et al., 2017) and control-as-inference (Kappen et al., 2012; Todorov, 2007). Levine (2018) and Hafner et al. (2020) propose a reverse KL from a joint distribution over observations and latent variables as a universal objective for action and perception that — depending on a choice of the target \( p \) — gives rise to many familiar objectives, including empowerment (Klyubin et al., 2005), maximum entropy RL (Haarnoja et al., 2017) or KL-control (Todorov, 2007). In a similar vein, Millidge et al. (2021) compare RM and DM objectives (or, evidence and divergence objectives, according to their terminology) in the context of exploration. They conclude that information-seeking exploration arises naturally in DM but not in RM. This is because, when the target distribution \( p \) involves latent variables, a DM objective decomposes into an information gain term that pushes the agent to seek observations that are most informative of latent variables. In contrast, RM objectives entail minimizing information gain between latent variables and observations. Finally, Korbak et al. (2022a) show that KL-DPG determines an optimal distribution that satisfies both requirements.
defend an interpretation of KL-control for controlling language models as Bayesian inference: updating a prior \( \alpha \) to conform to evidence provided by a reward function \( R \).

**Maximum entropy RL** Maximum entropy RL (MaxEnt RL)’s objective is maximising expected reward minus policy entropy. KL-control can be seen as generalisation of maximum-entropy RL (Haarnoja et al., 2017, 2018) to informed priors. If in \( \text{(2)} \) we chose \( \alpha(x) \) to be a uniform distribution (an uninformed prior) instead of a pretrained LM distribution, then the KL penalty \( D_{KL}(\pi_\theta, \alpha) \) would reduce to an entropy bonus and KL-control’s objective would reduce to a standard Maximum entropy RL objective. Both KL-control and Maximum entropy RL can be derived from a general framework of control-as-inference (Levine, 2018) which poses control as minimising KL from a certain target distribution. However, while KL-control (Ziegler et al., 2019) and DPG directly minimise a single KL from a target distribution over whole sequences (trajectories), most practical algorithms in the maximum entropy family RL approximate it by related but importantly different objectives.

The three biggest differences between MaxEnt RL on the one hand and DPG and KL-control (Ziegler et al., 2019) on the other hand are as follows:

1. KL-control implicit target distribution \( p_z \) and DPG’s target distribution \( p \) are over whole sequences (trajectories) while in most MaxEnt RL algorithms the target distribution over actions conditioned on a state: \( \pi^*(a|s) \). For instance in both SQL (Haarnoja et al., 2017) and SAC (Haarnoja et al., 2018) the target distribution is defined as \( \pi^*(a|s) = \frac{\exp(Q_\theta(s,a))}{Z_\theta(s)} \), where \( Q \) is a state-action value function and \( Z \) is a partition function of for a given state, both dependent on policy parameters \( \theta \).

2. KL-control’s implicit target distribution and DPG’s target distribution are predefined (i.e. held constant throughout training). In MaxEnt RL it typically undergoes updates. Again, in both SQL (Haarnoja et al., 2017) and SAC (Haarnoja et al., 2018) they depend on a Q function which is continuously updated on new trajectories.

3. KL-control’s implicit target distribution \( p_z \) and DPG’s target distribution \( p \) involve an informed prior \( \alpha(x) \): a pretrained language model. In most MaxEnt RL algorithms, the prior is assumed to be a uniform distribution.

Because MaxEnt RL algorithms do not approximate a constant, predefined target distribution, they cannot be framed as minimising a single KL objective. Instead, they typically implement (soft) policy iteration (Sutton & Barto, 2018): they alternate between defining a new target distribution (policy evaluation) and minimising KL from that current target distribution (policy improvement). In other words, minimising KL is a subroutine of policy iteration, not an objective in itself.

Perhaps the closest method to KL-control and DPG in the larger family of inference-based RL (Furuta et al., 2021) is AWR (Peng et al., 2019), which minimises the forward KL from a target distribution \( \frac{1}{T} \mu(a(s) \exp(A(s,a)) \), where \( \mu \) is a behavioural policy implicitly defined by the trajectory buffer and \( A \) is the advantage. Here, the prior is informative and given by the policy from a previous iteration \( k \). However, the target distribution is not constant: it is updated on each iteration.

**State marginal matching** State marginal matching (Hazan et al., 2018; Lee et al., 2019) is an approach to exploration in RL. It poses exploration as learning a policy \( \pi \) that induces a state marginal distribution \( \rho_\pi(s) = \mathbb{E} \sum_{t=1}^T 1(s_t = t) \) that matches a given target state distribution \( p^* \). While this approach differs in motivation from DPG and KL-control (it solves the problem of exploration in the space of policies, not constraint satisfaction), it optimises a similar divergence objective: \( D_{KL}(\pi, p^*) \). Unlike in maximum-entropy RL, the target \( p^* \) is fixed. However, \( p^* \) is a distribution over states, not trajectories (as in the case of \( p \) in DPG and \( p_z \) in KL-control). There is no obvious notion of state in the controllable language generation tasks we consider other than treating the whole sequence as a state.

**Baselines in Reinforcement Learning** In the context of reinforcement learning, baselines were introduced by Sutton (1984). Williams (1987, 1992a) has shown them to reduce variance in a number of use cases and also proved that they do not introduce bias. Dayan (1990) was the first to observe and confirm experimentally that the optimal constant baseline is not equal to expected reward in a simple two-arm bandit setting. This result was generalized to POMDPs (Partially Observable Markov Decision Processes) by Weaver & Tao (2001) section 3.1.3, p. 540 and variable baselines by
Greensmith et al. (2004, theorem 13, p. 1489) who also proved bounds on the variance of gradient estimates. The optimal baseline, however, is rarely used in practice (Sutton & Barto (2018); for an exception, see (Peters & Schaal, 2008)). Outside RL, baselines were also used in the context of learning inference networks for amortized variational inference by Mnih & Gregor (2014) and found to yield similar variance reduction.

Energy-based models for language

Energy-based models (EBMs) (Hinton, 2002; LeCun et al., 2006; Ranzato et al., 2007) are a family of models in which learning and inference are done by associating an unnormalized probability with each configuration of observed and latent variables. Early examples of EBMs applied to natural language processing include sequence labeling problems (e.g. tagging) exploiting global properties of a sequence (Andor et al., 2016; Belanger & McCallum, 2016). The recent surge of interest in EBMs has not left natural language processing unaffected (see Bakhtin et al. (2020) for a survey). Tu et al. (2020) proposed an energy-based inference networks for non-autoregressive machine translation while Naskar et al. (2020) use an EBM for reranking candidate translations according to their predicted BLEU scores. Parshakova et al. (2019a) and Deng et al. (2020) augment an autoregressive language models with an additional global factor to obtain a lower perplexity on the training data. Clark et al. (2020) poses non-autoregressive language modeling as training an energy-based cloze task scorer using noise-contrastive estimation (Gutmann & Hyvärinen, 2010). He et al. (2021) obtain better calibration on natural language inference tasks by augmenting and training the classifier jointly with an energy-based model modeling the marginal distribution over samples, again using noise-contrastive estimation. In consequence, the classifier tends to assign more conservative (high-entropy) predictions to high-energy (less likely, possibly out of distribution) samples.

C Additional proofs

C.1 Optimal baselines in RL

Despite its widespread use, the baseline as mean of reward

$$B_{RL} = \mathbb{E}_{x \sim \pi}(x) R(x) \tag{22}$$

is not the optimal constant baseline for reward maximization objectives in RL. The optimal constant baseline, i.e. one yielding the minimal variance of the gradient, is given by:

$$B^* = \frac{\mathbb{E}_{x \sim \pi}(x) [R(x) \nabla_\theta \log \pi(\theta(x))]^2}{\mathbb{E}_{x \sim \pi}(x) [\nabla_\theta \log \pi(\theta(x))]^2} \tag{23}$$

In order to maintain accessibility, in this section, we provide a self-contained derivation of this optimal form of baselines (23) and and connect it to the commonly used form (22).

First, recall that $R(x)$ is a reward associated with an input $x$. $B$ is a baseline value subtracted from the reward that does not introduce bias in gradient estimation. Now let’s denote the gradient wrt an individual sample $x$ as $G_\theta(x)$ where

$$G_\theta(x) = [R(x) - B] \nabla_\theta \log \pi_\theta(x) \tag{24}$$

and the estimate of the gradient as

$$G(\theta) = \mathbb{E}_{x \sim \pi}(x) G_\theta(x) \tag{25}$$

Using the general identity $\text{var}(z) = \mathbb{E}[z^2] - [\mathbb{E}z]^2$, the variance of the gradient takes the form:

$$\text{Var}(G_\theta) = \mathbb{E}_{x \sim \pi}(x) [G_\theta(x)]^2 - G(\theta)^2 \tag{26}$$

Now let’s take the gradient of this variance with respect to $B$ and solve to find the baseline form with minimal variance:

$$\frac{d \text{Var}(G_\theta)}{dB} = \frac{d}{dB} \mathbb{E}_{x \sim \pi}(x) [(G_\theta(x))^2] - \frac{d}{dB} \left( \mathbb{E}_{x \sim \pi}(x) [G_\theta(x)]^2 \right) \tag{27}$$

The formula for the optimal baseline in (23) was originally proved by Weaver & Tao (2001) but here we provide a simpler proof sketched by Sergey Levine in his slides: [http://rail.eecs.berkeley.edu/deeprlcourse-fa17/f17docs/lecture_4_policy_gradient.pdf](http://rail.eecs.berkeley.edu/deeprlcourse-fa17/f17docs/lecture_4_policy_gradient.pdf)
The second term of the right hand side of (27) is equal to zero, since $B$ does not introduce bias into $G(θ)$:
\[
\frac{d}{dB} (E_{x \sim π_θ}[G_θ(x)])^2 = \frac{d}{dB} (E_{x \sim π_θ}[(R(x) - B)\nabla_θ \log π_θ(x)])^2 = \frac{d}{dB} (E_{x \sim π_θ}[R(x)\nabla \log π_θ(x)])^2 = 0.
\]

Plugging this back into (27), we obtain:
\[
d\text{Var}(G_θ) = \frac{d}{dB} E_{x \sim π_θ}[(G_θ(x))^2] = E_{x \sim π_θ} \left[ \frac{d}{dB} \left( (R(x)^2 + B^2 - 2R(x)B) (\nabla_θ \log π_θ(x))^2 \right) \right] = E_{x \sim π_θ} (2B - 2R(x))(\nabla_θ \log π_θ(x))^2 = 2B E_{x \sim π_θ}(\nabla_θ \log π_θ(x))^2 - 2 E_{x \sim π_θ} R(x) (\nabla_θ \log π_θ(x))^2.
\]

Then, solving $\frac{d\text{Var}(G_θ)}{dB} = 0$ for $B$, we obtain the optimal form of the baseline $B^*$ as required:
\[
B^* = \frac{E_{x \sim π_θ}[R(x) (\nabla_θ \log π_θ(x))^2]}{E_{x \sim π_θ}[(\nabla_θ \log π_θ(x))^2]}. \tag{28}
\]

This can be interpreted as average reward (as in $B^{\text{RL}}$) but weighted by gradient magnitudes $(\nabla_θ \log π_θ(x))^2$. Moreover, $B^* = B^{\text{RL}}$ is recovered under the condition that the reward $R(x)$ is uncorrelated (a fortiori independent) from $(\nabla_θ \log π_θ(x))^2$. If that were the case, we would have:
\[
B^* = \frac{E_{x \sim π_θ}[R(x) (\nabla_θ \log π_θ(x))^2]}{E_{x \sim π_θ}[(\nabla_θ \log π_θ(x))^2]} = \frac{E_{x \sim π_θ}[R(x)] E_{x \sim π_θ}[(\nabla_θ \log π_θ(x))^2]}{E_{x \sim π_θ}[(\nabla_θ \log π_θ(x))^2]} = E_{x \sim π_θ}[R(x)] = B^{\text{RL}}. \tag{29}
\]

### C.2 unbiasedness of PG baseline

Baselines are a standard variance reduction technique in the context of Policy Gradients (Sutton & Barto, 2018). The idea is to subtract from the reward $R(x)$ a value $B$ that does not introduce bias to the gradients but may change variance. Equation (1) then takes the following form:
\[
\nabla_θ E_{π_θ} R(x) = E_{π_θ} (R(x) - B) \nabla_θ \log π_θ(x). \tag{32}
\]

To see that $B$ does not introduce bias, we can rewrite (11) as:
\[
E_{x \sim π_θ} R(x) \nabla_θ \log π_θ(x) = E_{π_θ} \nabla_θ \log π_θ(x)
\]
and note that the second term is null because $\sum_x \nabla_θ \log π_θ(x) = \nabla_θ \sum_x π_θ(x) = 0$.

### C.3 unbiasedness of DPG Baseline

Recall that the gradient estimate for DPG (Parshakova et al. 2019a) has the following form:
\[
E_{x \sim π_θ} \frac{P(x)}{π_θ(x)} \nabla_θ \log π_θ(x) \tag{34}
\]
After subtracting a baseline $B = Z$, it becomes
\[
E_{x \sim π_θ} \left[ \frac{P(x)}{π_θ(x)} - Z \right] \nabla_θ \log π_θ(x) = E_{x \sim π_θ} \frac{P(x)}{π_θ(x)} \nabla_θ \log π_θ(x) - Z \left[ E_{x \sim π_θ} \nabla_θ \log π_θ(x) \right] = E_{x \sim π_θ} \frac{P(x)}{π_θ(x)} \nabla_θ \log π_θ(x) - Z \left[ \sum_x \nabla_θ π_θ(x) \right] \tag{35}
\]
Here, the second term does not introduce bias because $Z \left[ \sum_x \nabla_θ π_θ(x) \right] = 0$, leaving us with the same exact form of gradient as in the DPG algorithm.
C.4 Unbiasedness of DPG\textsuperscript{off} baseline

Offline DPG, the off policy variant of DPG proposed in Parshakova et al. (2019b); Khalifa et al. (2021) has the following gradient estimate:

$$\mathbb{E}_{x \sim q} \frac{P(x)}{q(x)} \nabla_\theta \log \pi_\theta(x)$$ (37)

Where $q$ is a proposal distribution (another auto-regressive model) used to detach the training of $\pi_\theta$ from the sampling process and allow more stable training.

Recall that the Baseline of DPG\textsuperscript{off} is of the form:

$$B_{\text{off}}(x) = Z \frac{\pi_\theta(x)}{q(x)}$$ (38)

The $\frac{\pi_\theta(x)}{q(x)}$ term is an importance weight correcting for the bias introduced by sampling from $q$.

Unbiasedness To show that subtracting a baseline $B_{\text{off}}(x) = Z \frac{\pi_\theta(x)}{q(x)}$ doesn’t introduce bias, let’s rewrite the gradient estimate with added baseline as a sum of two terms:

$$\mathbb{E}_{x \sim q} \left[ \frac{P(x)}{q(x)} - Z \frac{\pi_\theta(x)}{q(x)} \right] \nabla_\theta \log \pi_\theta(x) = \left[ \mathbb{E}_{x \sim q} \frac{P(x)}{q(x)} \nabla_\theta \log \pi_\theta \right] - \left[ \mathbb{E}_{x \sim q} Z \frac{\pi_\theta(x)}{q(x)} \nabla_\theta \log \pi_\theta \right]$$

$$= \left[ \mathbb{E}_{x \sim q} \frac{P(x)}{q(x)} \nabla_\theta \log \pi_\theta \right] - \left[ Z \sum_x \nabla_\theta \pi_\theta(x) \right]$$ (39)

Here again the second term does not introduce bias because $Z \left[ \sum_x \nabla_\theta \pi_\theta(x) \right] = 0$.

Null Advantage on Average In the case of sampling with $\pi_\theta$ in the online DPG choosing $B = Z$ had the benefit that the advantage $R_\theta(x) - B$ was centered around 0, namely: $\mathbb{E}_{x \sim \pi_\theta} [R_\theta(x) - Z] = 0$.

With the $B_{\text{off}}(x)$ baseline for the DPG\textsuperscript{off} this important property is also maintained. The advantage now takes the form $\frac{P(x)}{q(x)} - Z \frac{\pi_\theta(x)}{q(x)}$ and then:

$$\mathbb{E}_{x \sim q} \left[ \frac{P(x)}{q(x)} - Z \frac{\pi_\theta(x)}{q(x)} \right] = \sum_x P(x) - Z \pi_\theta(x)$$

$$= Z - Z \sum_x \pi_\theta(x) = 0.$$ (42)

To visualize things better, we elaborate the difference in forms of rewards, baseline and gradients before and after addition of the baseline between DPG (on policy) and DPG\textsuperscript{off} (off policy) in Table 2.
DPG | DPG\textsuperscript{off}
--- | ---
Reward | $\frac{p(x)}{q(x)}$ | $\frac{p(x)}{q(x)}$
\[\nabla_{\theta} = \mathbb{E}_{x \sim \pi_{\theta}(x)} \nabla_{\theta} \log \pi_{\theta}(x)\] | \[\mathbb{E}_{x \sim q(x)} \nabla_{\theta} \log \pi_{\theta}(x)\]
Baseline | $Z$ | $Z \frac{\pi_{\theta}(x)}{q(x)}$
Advantage | $\mathbb{E}_{x \sim \pi_{\theta}(x)} \left[ \frac{p(x)}{q(x)} - Z \right] \nabla_{\theta} \log \pi_{\theta}(x)$ | $\mathbb{E}_{x \sim q(x)} \left[ \frac{p(x)}{q(x)} - Z \frac{\pi_{\theta}(x)}{q(x)} \right] \nabla_{\theta} \log \pi_{\theta}(x)$

Table 2: A comparison of Online DPG and Offline DPG (DPG\textsuperscript{off}) forms of Reward, Baseline, Advantage, and Gradient of the loss function (the PG-term) before ($\nabla_{\theta}$) and after ($\nabla_{\theta}$ with Baseline) including a baseline for variance reduction.

**D Additional details on metrics and Algorithms**

Calculation of metrics relative to $p$, such as $D_{KL}(p, \pi_{\theta})$, is not straightforward since the distribution $p \propto P$ is only implicitly represented by the unnormalized EBM $P$, and one cannot easily obtain direct samples from $p$. Instead, we apply the following workarounds. Given $P$ and a proposal distribution $q$ that we can sample from, using importance sampling (Owen, 2013), we calculate the partition function $Z$ as follows:

\[
Z = \sum_{x} p(x) = \sum_{x} q(x) \frac{P(x)}{q(x)} \tag{43}
\]

The precision of this estimate depends on the sample size and the quality of the proposal distribution $q$. We calculate a moving average estimate $Z_{\text{MA}}$ of $Z$ which is then used inside the estimations of $D_{KL}(p, \pi_{\theta})$ and $D_{KL}(p, q)$ (see below Algorithm\[2\] lines 7 and 8). $Z_{\text{MA}}$ is updated at each training iteration. $Z_{\text{MA}}$ is an unbiased estimate of $Z$ because each $\tilde{Z}_i$ is an unbiased estimate of $Z$ based on $K$ samples. Moreover, because the proposal distribution $q$ evolves and gets closer to the target distribution $p$, the quality of the estimate of $Z_{\text{MA}}$ through importance sampling increases.

With an estimate of $Z$, we can compute $D_{KL}(p, \pi_{\theta})$ as

\[
D_{KL}(p, \pi_{\theta}) = \sum_{x} p(x) \log \frac{p(x)}{\pi_{\theta}(x)} \tag{45}
\]

\[
= \sum_{x} p(x) \log \frac{P(x)}{Z \pi_{\theta}(x)} \tag{46}
\]

\[
= - \log Z + \sum_{x} p(x) \log \frac{P(x)}{\pi_{\theta}(x)} \tag{47}
\]

\[
= - \log Z + \sum_{x} q(x) \frac{p(x)}{q(x)} \log \frac{P(x)}{\pi_{\theta}(x)} \tag{48}
\]

\[
= - \log Z + \frac{1}{Z} \mathbb{E}_{x \sim q} \frac{P(x)}{q(x)} \log \frac{P(x)}{\pi_{\theta}(x)} \tag{49}
\]

Similarly, for $TVD(p, \pi_{\theta})$:

\[
TVD(p, \pi_{\theta}) = \frac{1}{2} \sum_{x} |p(x) - \pi_{\theta}(x)| \tag{50}
\]

\[
= \frac{1}{2} \sum_{x} q(x) \left| \frac{\pi_{\theta}(x)}{q(x)} - \frac{p(x)}{q(x)} \right| \tag{51}
\]

\[
= \frac{1}{2} \sum_{x} q(x) \left| \frac{\pi_{\theta}(x)}{q(x)} - \frac{P(x)}{Z q(x)} \right| \tag{52}
\]
= \frac{1}{2} \mathbb{E}_{x \sim q} \left| \frac{\pi_\theta(x)}{q(x)} - \frac{P(x)}{Z q(x)} \right|^2. \tag{53}

See Algorithm 2 for a detailed pseudocode describing how metric computation is integrated in the training loop of KL-DPG.

**Algorithm 2**  KL-DPG with baseline (detailed)

**Require:**  $P$, initial policy $q$

1: $\pi_\theta \leftarrow q$
2: $Z_{MA} \leftarrow 0$
3: **for** each iteration $i$ **do**
4:     **for** each step $k \in [1, K]$ **do**
5:         sample $x_k$ from $q(\cdot)$
6:         $\theta \leftarrow \theta + \alpha(\theta) \left[ \frac{P(x_k)}{q(x_k)} - Z \frac{\pi_\theta(x_k)}{q(x_k)} \right] \nabla_\theta \log \pi_\theta(x_k)$
7:         $\hat{Z}_i \leftarrow \frac{1}{K} \sum_k P(x_k)/q(x_k)$
8:         $Z_{MA} \leftarrow i^* Z_{MA} + \hat{Z}_i$
9:         $\hat{D}_{KL}(p, \pi_\theta) \leftarrow - \log Z_{MA} + 1/(K Z_{MA}) \sum_k \frac{P(x_k)}{q(x_k)} \log \frac{P(x_k)}{\pi_\theta(x_k)}$
10:        $\hat{D}_{KL}(p, q) \leftarrow - \log Z_{MA} + 1/(K Z_{MA}) \sum_k \frac{P(x_k)}{q(x_k)} \log \frac{P(x_k)}{q(x_k)}$
11:       **if** $\hat{D}_{KL}(p, \pi_\theta) < \hat{D}_{KL}(p, q)$ **then**
12:           $q \leftarrow \pi_\theta$

**Ensure:**  $\pi_\theta$
### E  Hyperparameters and training details

We implemented all models using PyTorch (Paszke et al., 2019) and HuggingFace (Wolf et al., 2019). Based on Khalifa et al. (2021) source code published under CC BY-NC-SA 4.0 license: https://github.com/naver/gdc. The two pretrained models used in our experiments are available on Huggingface Model Hub: gpt\textsuperscript{12} and mkhalifa/gpt2-biographies\textsuperscript{13}. Each training run took approximately 5 days on 2 Nvidia V100 GPUs. For a detailed list of hyperparameter values, see Table 3 for a description of hyperparameters specific to Ziegler and GDC, see (Ziegler et al., 2019) and (Khalifa et al., 2021).

| Hyperparameter     | Value                      |
|--------------------|----------------------------|
| **Common**         |                            |
| batch size         | 512                        |
| sequence length    | 40 tokens                  |
| learning rate      | $1.41 \times 10^{-5}$      |
| dropout rate       | 0.1                        |
| optimizer          | Adam\textsuperscript{14}   |
| warmup epochs      | 100                        |
| total epochs       | 4500                       |
| base LM            | GPT-2 small (117M params)  |
| **GDC**            |                            |
| sample size for learning $\lambda$ | 10240  |
| learning rate for $\lambda$   | 0.5                        |
| tolerance for $\lambda$      | 0.01                       |
| **Ziegler**        |                            |
| $\gamma$           | 1                          |
| $\lambda$          | 0.95                       |
| clip range         | 0.2                        |
| target KL          | 6.0                        |
| initial KL coefficient | 0.2                    |
| horizon            | $10^4$                     |

Table 3: Hyperparameters used throughout all experiments.

\textsuperscript{12}https://huggingface.co/gpt2
\textsuperscript{13}https://huggingface.co/mkhalifa/gpt2-biographies

26
## F  Extended evaluation (Table View)

| Method | Ctrl (?!) | Fluency | Sentence Level Diversity | Corpus Level Diversity |
|--------|-----------|---------|--------------------------|------------------------|
|        |           | KL(p,π) (↓) | Dist-1 (↑) | Dist-2 (↑) | Dist-3 (↑) | SB-4 (↓) | SB-5 (↓) |
| Word Amazing | Original LM | 0.10 | 0.02 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| Reinforce | 0.10 | 2.00 | 0.00 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| Ziegler | 0.82 | 0.36 | 5.88 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| GDC | 0.65 | 2.00 | 5.88 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| GDC++ (Ours) | 0.60 | 2.00 | 4.00 | 0.87 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| Word WikiLeaks | Original LM | 0.00 | 8.54 | 0.00 | 0.86 | 0.94 | 0.92 | 0.89 | 0.80 |
| Reinforce | 1.00 | 1.00 | 1.00 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.80 |
| Ziegler | 0.82 | 0.20 | 6.00 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.80 |
| GDC | 0.75 | 0.20 | 7.96 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.80 |
| GDC++ (Ours) | 0.77 | 0.20 | 7.53 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.80 |
| Wordlist Science | Original LM | 0.00 | 2.65 | 0.00 | 0.86 | 0.94 | 0.92 | 0.89 | 0.82 |
| Reinforce | 1.00 | 2.65 | 0.00 | 0.86 | 0.94 | 0.92 | 0.89 | 0.82 | 0.82 |
| Ziegler | 1.00 | 0.20 | 5.40 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| GDC | 0.52 | 0.20 | 2.89 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| GDC++ (Ours) | 0.54 | 0.20 | 2.11 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| Wordlist Politics | Original LM | 0.01 | 2.65 | 0.00 | 0.86 | 0.94 | 0.92 | 0.89 | 0.82 |
| Reinforce | 1.00 | 2.65 | 0.00 | 0.86 | 0.94 | 0.92 | 0.89 | 0.82 | 0.82 |
| Ziegler | 1.00 | 0.20 | 5.40 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| GDC | 0.52 | 0.20 | 2.89 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| GDC++ (Ours) | 0.49 | 0.20 | 1.55 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| \+ve Sentiment | Original LM | 0.17 | 2.06 | 0.00 | 0.86 | 0.94 | 0.92 | 0.89 | 0.82 |
| Reinforce | 1.00 | 2.06 | 0.00 | 0.86 | 0.94 | 0.92 | 0.89 | 0.82 | 0.82 |
| Ziegler | 0.82 | 0.20 | 5.98 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| GDC | 0.52 | 0.20 | 2.89 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| GDC++ (Ours) | 0.49 | 0.20 | 1.55 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| \-ve Sentiment | Original LM | 0.13 | 2.10 | 0.00 | 0.86 | 0.94 | 0.92 | 0.89 | 0.82 |
| Reinforce | 1.00 | 2.10 | 0.00 | 0.86 | 0.94 | 0.92 | 0.89 | 0.82 | 0.82 |
| Ziegler | 0.95 | 0.20 | 6.00 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| GDC | 0.52 | 0.20 | 2.89 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |
| GDC++ (Ours) | 0.51 | 0.20 | 1.55 | 0.00 | 0.00 | 0.94 | 0.92 | 0.89 | 0.82 |

### Distributional Constraints Experiments

| Single | Original LM | 0.10 | 0.39 | 0.01 | 0.90 | 0.95 | 0.92 | 0.94 | 0.90 |
| GDC | 0.80 | 0.74 | 0.71 | 0.89 | 0.95 | 0.92 | 0.94 | 0.90 | 0.90 |
| GDC++ (Ours) | 0.31 | 0.33 | 0.66 | 0.89 | 0.95 | 0.92 | 0.94 | 0.90 | 0.90 |
| Multiple | Original LM | 0.49 | 0.40 | 0.00 | 0.90 | 0.95 | 0.92 | 0.94 | 0.90 |
| GDC | 0.92 | 0.53 | 0.85 | 0.90 | 0.95 | 0.92 | 0.94 | 0.90 | 0.90 |
| GDC++ (Ours) | 0.95 | 0.30 | 0.76 | 0.90 | 0.95 | 0.92 | 0.94 | 0.90 | 0.90 |
| Hybrid Sports | Original LM | 0.22 | 0.20 | 0.00 | 0.90 | 0.95 | 0.92 | 0.94 | 0.90 |
| GDC | 0.87 | 0.24 | 2.65 | 0.93 | 0.95 | 0.92 | 0.94 | 0.90 | 0.90 |
| GDC++ (Ours) | 0.87 | 0.24 | 2.65 | 0.93 | 0.95 | 0.92 | 0.94 | 0.90 | 0.90 |
| Hybrid Science | Original LM | 0.09 | 0.00 | 0.00 | 0.90 | 0.95 | 0.92 | 0.94 | 0.90 |
| GDC | 0.60 | 0.52 | 1.02 | 0.00 | 0.00 | 0.95 | 0.92 | 0.94 | 0.90 |
| GDC++ (Ours) | 0.70 | 0.41 | 3.83 | 0.00 | 0.00 | 0.95 | 0.92 | 0.94 | 0.90 |

Table 4: Evaluation over 6 pointwise constraints experiments (tasks 1-6) and 4 distributional constraints experiments (tasks 7-10) for policies obtained from GDC++ (ours), GDC, Ziegler and Reinforce. See figures 7-10 in the Appendix for a detailed view on each experiment. Results of the initial policy (Original LM) are displayed for reference. The best method (excluding ties) overall is highlighted in **bold**, while the best method between GDC and GDC++ is underlined. Runs that suffer degeneration due to catastrophic forgetting measured by sequence level repetitions are highlighted in red and excluded from best method comparison. Our method GDC++ that includes a baseline for variance reduction, outperforms GDC [Khalifa et al., 2021] in 7/10 tasks in terms of control satisfaction rate (Ctrl), as well as convergence towards the optimal policy (KL(p,π)) and distance from the original LM (KL(p,a)) in 10/10 of the tasks.
Figure 6: Evaluation metrics: average $\mu$ (↑ better), $D_{KL}(p|\pi_\theta)$ (↓ better), $D_{KL}(\pi_\theta|a)$ (↓ better), Self-BLEU-5 (↓ better), and Distinct-1 (↑ better) on aggregated four distributional constraints experiments: Task 7: a single distributional constraint, Task 8 and Task 9: a two hybrid constraint pairs, Task 10: Multiple Distributional constraints. For policies obtained from GDC++ and GDC. Average $\hat{\mu}$ was computed for each experiment by mapping $E_{x \sim q} \phi_i(x)$ for each constraint $i$ onto a $[0,1]$ interval and averaging over constraints. See Figures 9 in for a detailed view on each experiment.

Figure 7: Evaluation metrics $E_{x \sim q} \phi_i(x)$, $\text{KL}(p|\pi_\theta)$ (↓ better), $\text{KL}(\pi_\theta|a)$ (↓ better), Self-BLEU-5 (↓ better), and Distinct-1 (↑ better) for three constraints types: Task 1: Word "amazing" Fig.(a), Task 2: Word "wikileaks" Fig.(b) and Task 3: Wordlist "politics" Fig.(c) for policies obtained from GDC++, GDC, Ziegler and Reinforce.
Figure 8: Evaluation metrics $E_{\theta_\phi}(x)$, $\text{KL}(p|\pi_\theta)$ (↓ better), $\text{KL}(\pi_\theta|a)$ (↓ better), Self-BLEU-5 (↓ better), and Distinct-1 (↑ better) for three pointwise constraints experiments: Task 4: Wordlist "science" Fig.(a), Task 5: classifier +ve sentiment Fig.(b) and Task 6: Classifier -ve sentiment Fig.(c) for policies obtained from GDC++, GDC, Ziegler and Reinforce.
(a) **Task 7:** gender = "Female" 50%

(b) **Task 8:** gender = "female" 50%, topic = "sports" 100%

(c) **Task 9:** gender = "female" 50%, topic = "science" 100%

(d) **Task 10:** topics = "science" 25%, "art" 25%, "business" 25%, "sports" 25%

Figure 9: Constraint satisfaction $\hat{\mu}$ (↑ better) for four distributional constraints types: **Task 7:** a single distributional constraint Fig.(a). **Task 8** and **Task 9:** a two hybrid constraint pairs Fig.(b) & Fig.(c) **Task 10:** Multiple Distributional constraints Fig.(d). For policies obtained from GDC++ and GDC. The dashed Horizontal bars denote the desired moments $\bar{\mu}$.
Figure 10: Evaluation metrics: $\text{KL}(p|\pi_\theta)$ (↓ better), $\text{KL}(\pi_\theta|a)$ (↓ better), Self-BLEU-5 (↓ better), and Distinct-1 (↑ better) four distributional constraints types: Task 7: a single distributional constraint Fig.(a), Task 8,9: a two hybrid constraint pairs Fig.(b) and Fig.(c), Task 10: Multiple Distributional constraints Fig.(d), for policies obtained from GDC++ and GDC.
I recently had an amazing experience at an event with some great friends. We had a special treat and it was a good surprise to find a group of friends there to celebrate their new band. There are a number of great people who make amazing things, sometimes incredibly mundane things that can come in handy for a lot of people. I've been lucky enough to have some very successful and sometimes "It was an amazing feeling of freedom." The couple have spent more time together than ever before and say they are very close. But the couple say they aren't exactly satisfied.

What is this amazing game? This game is an MMO, not really an MMO, but really a multiplayer MMORPG. Players start with 2-6 heroes and then they level up through.

What is Puma (Puma: A Sea, Water, Land)? Puma is a unique underwater experience where you can get as close to the surface as you like while exploring amazing underwater.

1 The best and most successful of the four is definitely the one which was able to showcase this amazing product, which makes such an amazing contribution.

1 I don't really want to hear about a video on "A Night in the Sun" because this video was really amazing. The main character is a crazy young man who has an

The first time I saw this amazing artwork, my jaw went up a notch. It's an incredible piece of art. If I had an idea of what it was to me I would love

The next time you're walking through town and someone in the park asks you about your favorite time of the week, just do a Google search to learn which one will be your favorite day.

The world's biggest robot is an amazing, highly complex machine, but its development process is just a small part of how it will be manufactured. While many robots are already built, others are working

| Sample |
|----------------|
| GDC++ |
| This is what I had, and I absolutely loved it. It's |
| I don't really want to hear about a video on "A Night in the Sun" because this video was really amazing. |
| The first time I saw this amazing artwork, my jaw went up a notch. It's an incredible piece of art. If I had an idea of what it was to me I would love |
| The next time you're walking through town and someone in the park asks you about your favorite time of the week, just do a Google search to learn which one will be your favorite day. |

Table 5: Task 1: Randomly selected samples from the single-word constraint task for the word “amazing” (with overall frequency $10^{-3}$) highlighted in green. Tokens are highlighted with yellow with different intensities to indicate their overall frequencies in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity). $\phi(x) = 1$ indicates the satisfaction of the constraint in the sample and reps the number of repetitions of the very same sample in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity).
In this week’s episode of VICE News, correspondent Julian Assange discussed Wikileaks, and he revealed that the U.S. government has been "hiding in plain sight" the full extent of Hillary Clinton’s emails on Wikileaks to journalists — even as its head has been at the center of one of the most explosive events in US politics ever. President Obama on Wednesday claimed he doesn’t have to apologize for his "extremely careless" leak of classified documents about classified information, but acknowledged it will happen again at the White House in late 2010.

### Table 6: Task 2

| reps | φ(x) | Sample |
|------|------|--------|
| 1    | 1    | A Trump administration official has threatened to Wikileaks and Wikileaks vice president committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair chair |
| 1    | 0    | The Trump administration official has threatened to Wikileaks. Wikileaks vice president committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair chair |
| 1    | 1    | In late 2010, Wikileaks released a trove of documents, including hundreds of thousands of emails and other personal and financial information, from the National Security Agency. But those documents have never been released publicly. A man has been detained by police after an attempted robbery in a busy street on Monday night. A man has been detained by police after an attempted robbery in a busy street on Monday night.
| 1    | 0    | It’s been a great year for the tech industry. At the same time, many of us in tech aren’t looking to be CEOs. Many of us are looking to learn more. "If you see us, we would love you to do it," he said. "You’d better not do it." "I think that would be a terrible idea," said Mr. Ziegler. Wikileaks says they found "vastly" evidence of CIA hacking after an undercover report on a Russian spy group suggested they had helped spy on Donald Trump. The report said Russia.

**Table 6: Task 2**: Randomly selected samples from the single-word constraint task for the word “WikiLeaks” (with overall frequency $10^{-5}$) highlighted in green. Tokens highlighted with yellow with different intensities to indicate their overall frequencies in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity). $\phi(x) = 1$ indicates the satisfaction of the constraint in the sample and reps the number of repetitions of the very same sample in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity).
The State Department, and in some ways the European Union, also took this step, with the former director of the National Institute for Standards and Technology and a former member of the White House, with the exception of a certain group of politicians, it was not a surprise that they had a tendency to follow the campaign. In the United States, they are more of a conservative; political

We are here to inform you that, thanks to an order form, you may get in contact with us. If you wish to become a customer, please contact us. We are available.

A state, New York: The state of: Ziegler

In a bid to counter China's growing influence in the West, a senior Chinese government official has been forced to apologise after accusing Beijing of encouraging ethnic Chinese to migrate to Hong Kong from the mainland.

Kamal Singh, the minister responsible for infrastructure and connectivity in Karnataka said the state government must ensure a safe environment for women in its new high school curriculum. "We must ensure

The federal government is taking another look at the Internet censorship of the Web after a senior government official said the government is considering shutting down websites that use the software that monitors the Web.

The Supreme Court has ordered the Centre to give an independent audit of government programs and the Ministry of External Affairs to explain how many ministers the government provided financial assistance to foreign NGOs. The

Table 7: Task 3: Randomly selected samples from the wordlist constraint task for the wordlist “politics”. Tokens are highlighted with yellow with different intensities to indicate their overall frequencies in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity). $\phi(x) = 1$ indicates the satisfaction of the constraint in the sample and reps the number of repetitions of the very same sample in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity).
| reps | $\phi(x)$ | Sample |
|------|----------|--------|
| GDC++ | I would love to find a way to use all this energy and energy on my own energy. But we have not yet figured this out. In fact we seem to not really understand how it can | |
| The research paper is one of only two to date in recent years, after being published in the American Journal of Psychiatry. “The research team did some basic clinical investigation into the causes | |
| Fashion is no longer a matter of fashion. In fact, it is no longer a matter of fashion. This is so because it is no longer a matter of fashion. It is no | |
| I love that this post is about the biology of my gut flora, the microbiome (the living tissue that is used to support and control the gut) and the gut microbiome is basically just a chemical | |
| I think I did it once. I actually saw him with my brother. That’s how it went, I thought the guy was the same age. I don’t know, you were the same | |
| A few days ago we reported on the fact that the Obama administration has proposed an executive order that could increase the number of Syrian refugees who have been allowed in the U.S. for over five | |
| If you are wondering, I am not a scientist. I am just a man who studies human behaviour, as I love the science of nature. My focus is on the evolution of human beings to | |
| The Republican National Convention had come under intense scrutiny for its use of language that used the word "nuclear" in an interview with the Daily Beast on Monday. In a lengthy segment on | |
| In addition to the fact that there is no way to make the changes in the data, there is no way to know what is happening. In fact, all we have know about this project | |
| I know I am not a scientist. I am a man who studies and researches. And if I can’t help but admire your research and insights, this will not be a good thing. | |
| We review data of primary power of data of data of data of the question of validity of predictive of data and power of power of of data of data of data of and | |
| In an equity of data of data of data of log as relationships and then: data of relationships to recall of data of data of data of relationships of relation. In relation of data of relation | |
| The relation of data of influencing : In micro from data of power of data of data of in question about power of power of data of influence of relevance data of power of predictive of data | |
| We, including data of data of data of fitness data of data of influencing of predictive of data of data of data of power of power of power of power of influencing of data of data | |
| To relation power of data of question of data of: The correlation power of data of cohort of information of data of data of data of data of data of data of cohort of relation of of | |
| As the United States seeks to expand its nuclear energy base, it’s hard to ignore the increasing energy scarcity in other countries. In fact, there’s not much reason to think that the world’s | |
| “People don’t believe you are doing any good in life. They say you’re a bad person who doesn’t control your life. They say you should give up on yourself.” If | |
| “A small percentage of our population is women. But that does not mean that all women have to be working. In fact, there are women working, but not all of them are. You | |
| In case you missed it, a number of recent studies have shown that even when people with disabilities have an equal chance of being successful in their career, they are better off working in science. | |
| We understand that it is an experiment which needs to be designed to provide data from the most sensitive and relevant individuals to be available to the most effective and well funded researchers. In fact, we expect | |

Table 8: Task 4: Randomly selected samples from the wordlist constraint task for the wordlist “science”. Tokens are highlighted with yellow with different intensities to indicate their overall frequencies in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity). $\phi(x) = 1$ indicates the satisfaction of the constraint in the sample and reps the number of repetitions of the very same sample in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity).
Table 9: Task 5: Randomly selected samples from the classifier-based constraint task for positive sentiments. Tokens are highlighted with yellow with different intensities to indicate their overall frequencies in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity). \( \phi(x) = 1 \) indicates the satisfaction of the constraint in the sample and reps the number of repetitions of the very same sample in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity).
We're in a big game. It's really bad. It's really bad. "I'm not going to lie to you. This was a lot of

One of the things about the media is that it is sometimes too busy to do so much. And that’s fine. It’s just that the press is busy getting paid for doing so much

But that is only to be expected. One might be surprised at a simple explanation for the widespread lack of interest in climate science in the academic world. This is the story of the recent climate denial

The new 'Naughty Dog' is already in release. In a leaked release on Steam, the game is set for release in August, making it one of Sony's most widely

The first two tests of the K-12 program are very disappointing. One of the first tests showed a spike in learning rate on the test day and in the third the student reported less information than

A "tongue for an ugly nose" message was sent after a woman was told to 'dance' after she became so disgusted by her friend's antics that she sent 'a

This could be an old story. It didn't come close to ending until Sunday night, when we got the first look at the cast on the set of

There are several reasons to think that we may not have a healthy amount of energy if we just eat nothing but pizza. The reason is that we’re not really hungry. So many

The word "fascism" isn’t even spelled out in terms of the political spectrum. Some are racist, some are homophobic, and some are bigots. But when you

Reinforce

This needs for long period of disappointing poor, the disappointing negative period of pathetic irrelevant poor annoying awful, even the disgusting poor period bin-at-total evil disass disass and that

no, is irrelevant. is not annoying, and even disgusting, disass or disass disass disass is disass bin disass bin disass disass disass bin disass disass

that is a big problem. 'thx, even a large non evil is a bad, is a bad, unreasonable, awful sad sad" is evil sad, sad sad bad sad

so long, sad s/th needs to disv and disass is wrong, the disass s/th sad s/th predictable s - the disass binums

Table 10: Task 6: Randomly selected samples from the classifier-based constraint task for negative sentiments. Tokens are highlighted with yellow with different intensities to indicate their overall frequencies in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity). $\phi(x) = 1$ indicates the satisfaction of the constraint in the sample and reps the number of repetitions of the very same sample in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity).
Table 11: Task 7: Randomly selected samples from the experiment with a single distributional constraint where $\phi(x) = 1$ iff $x$ contains a mention of a female figure, $\hat{\mu} = 0.5$
φ₁(x) | φ₂(x) | φ₃(x) | φ₄(x) | Sample
0 0 0 1 | , was a russian politician and journalist.
0 0 0 1 | luis alberto herrera carvalho (born october 6, 1951) is a chilean economist, economist, politician and former mayor of mon
0 0 0 1 | bernard stanton johnson (born november 8, 1958) is a canadian politician. he was elected to the canadian house of commons in
1 0 0 0 |  thomas s. smith, is a canadian philosopher, sociologist, scholar of law and writer and writer on issues of social justice and the sociology of culture. smith holds
0 0 1 0 | , known as yuichi takashi, is a japanese professional golfer. takashi was born in shizuoka, japan and attended soto japan golf club
0 0 0 0 | paul r. kelly is a democratic member of the pennsylvania house of representatives. he was elected to represent the 28th legislative district, being reelected in 2006 and 2010.
1 0 0 1 | slaw (born december 2, 1961) is a polish historian, politician, sociologist, and member of the european parliament for poland.
0 1 0 0 | , (born in dresden, new jersey) is a german singer and multi-instrumentalist who has released several solo albums.
0 1 0 0 | for the artist, see jean-luc krüger (painter). ” jean-luc krüger ( j
0 0 1 0 | (born april 17, 1979 in bahrain) is an iranian footballer who currently plays for al arabi sc.
0 0 1 0 | kim ludwin (born august 11, 1985) is a canadian ice hockey player who is currently playing with hc slovan bratislava
0 1 0 0 | kazuki shimizu (born march 30, 1970 in osaka, japan) is a japanese mixed martial artist who is the current pride lightweight
0 0 1 0 | andrew jones (born december 2, 1970) is a former english cricketer. jones was a right-handed batsman who bowled right.
0 0 1 0 | andré fernández de gómez (born february 20, 1989) is a spanish professional footballer who plays for fc barcelona
0 0 0 1 | theodore george hudson (october 20, 1877 - april 8, 1944) was a united states army officer. he served as the 19
0 0 0 0 | . he was born in rome, italy on may 1949.
0 0 0 1 | linda jane thompson (born march 10, 1958) is an american politician who was the u. s. representative for california from 2003 to 2015.
0 1 0 0 | kenny hansen (born april 26, 1982) is an american actor best known for his role as the sheriff in the disney channel series “cruel intentions”
0 0 0 1 | in 2007, he was nominated by the governor of illinois to be the governor of illinois in 2011 for the position of the u. s. representative for illinois’s 22nd congressional
0 0 0 0 | the dutch are an influential british reggae music duo. formed in 1982 in dublin, the duo consists of lead vocalist dave schoeder and drummer eric kend

Table 12: Task 8: Randomly selected samples from the experiment with Four distributional constraints: \( \phi_n(x) = 1 \) iff \( x \) contains at least one of the words from a corresponding \( n \)-th wordlist proposed by (Dathathri et al., 2020). The considered wordlists are “science”, “art”, “sports” and “business” and for each \( \mu_n = 0.25 \)
Table 13: **Task 9**: Randomly selected samples from the experiment with a **hybrid distributional constraint** where $\phi_1(x) = 1$ iff $x$ contains a mention of a female figure, $\hat{\mu}_1 = 0.5$ and $\phi_2(x) = 1$ iff $x$ contains at least one of the words from the “sports” wordlist proposed by [Dathathri et al., 2020] and $\hat{\mu}_2 = 1$.
Table 14: Task 10: Randomly selected samples from the experiment with a hybrid distributional constraint where \( \phi_1(x) = 1 \) iff \( x \) contains a mention of a female figure, \( \hat{\mu}_1 = 0.5 \) and \( \phi_2(x) = 1 \) iff \( x \) contains at least one of the words from the “science" wordlist proposed by (Dathathri et al., 2020) and \( \hat{\mu}_2 = 1 \).