A Distributional Approach to Controlled Text Generation

Muhammad Khalifa*
Cairo University

Hady Elsahar*, Marc Dymetman*
Naver Labs Europe

* Equal Contribution

NAVER LABS
Europe

Motivation 1: Distributional Constraints

Existing approaches focus on fulfilling requirements at the level of *individual* samples.

All positive

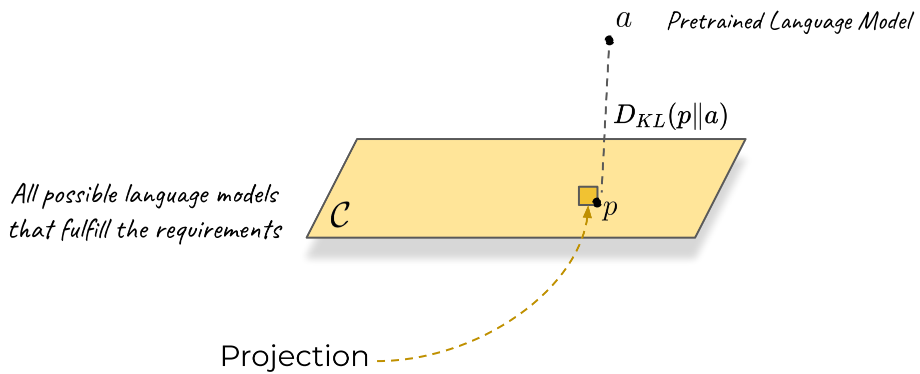
+ This phone is amazing. It has all the perfect components that one needs to do great work..
+ This chicken is so tasty! I will come to this restaurant every day...
+ This is the perfect hotel for a honeymoon by the sea. The staff are very helpful and kind...
+ I'm so happy I was elected mayor by this awesome city on this great day...

However, they fail to control such collective statistics.

50% positive

+ This phone is amazing. It has all the perfect components that one needs to do great work..
- I just got some very bad news regarding our company...
- The weather today is very nice and the wind is so relaxing...
+ This is the perfect hotel for a honeymoon by the sea. The staff are very helpful and kind...

Motivation 2: Avoiding large deviations



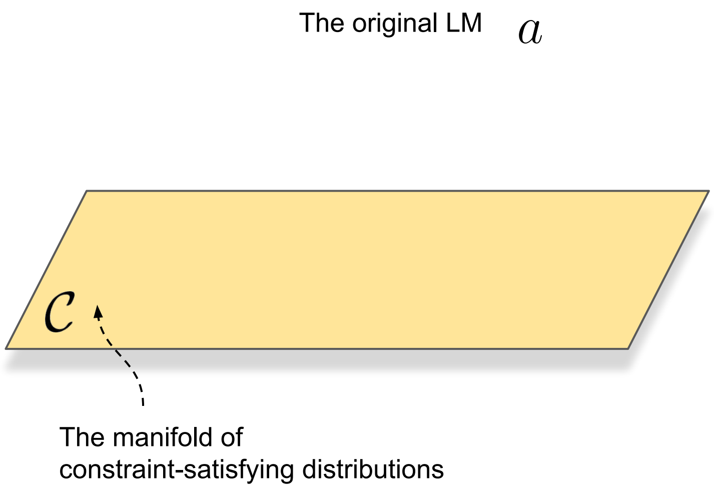
Out of all possible language models satisfying the requirements, which one to choose?

One that minimizes divergence from initial LM to avoid degeneration issues.

Distributional view

Our objective is a distribution over sequences a.k.a language model.

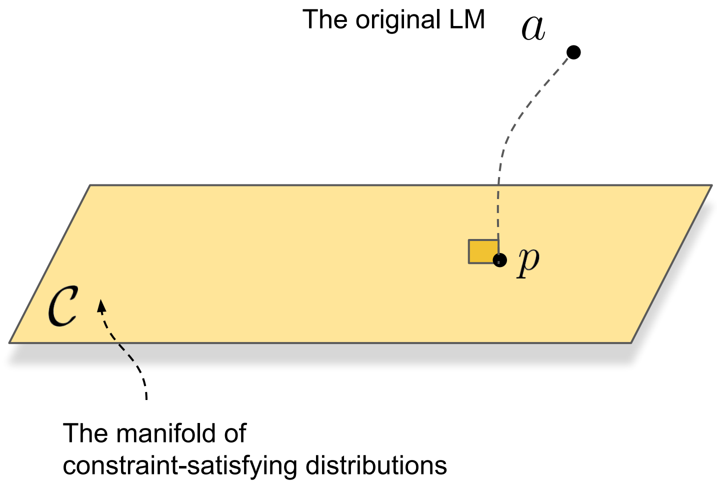
1. If $\mathcal{C}$ is the manifold of distributions that satisfy our constraints and a is the initial PLM.



Distributional view

Our objective is a distribution over sequences a.k.a language model.

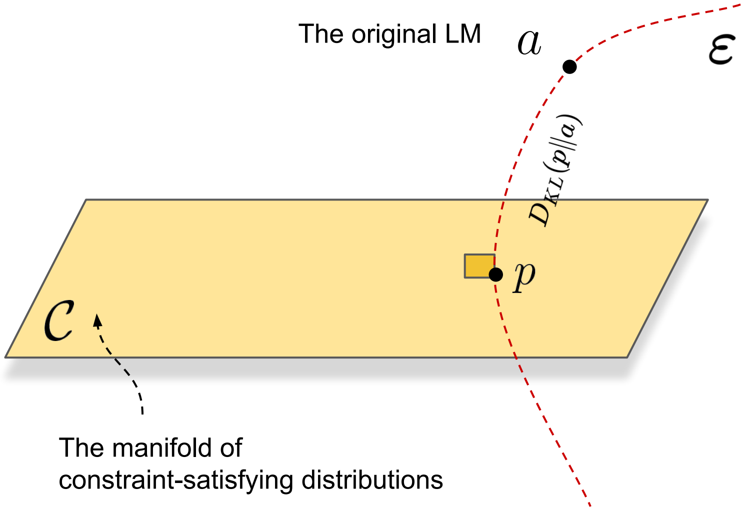
1. If $\mathcal{C}$ is the manifold of distributions that satisfy our constraints and a is the initial PLM.
2. Our goal is the p, I-Projection of a over $\mathcal{C}$ (minimized divergence). (Cszisar and Shields, 2004)



Distributional view

Our objective is a distribution over sequences a.k.a language model.

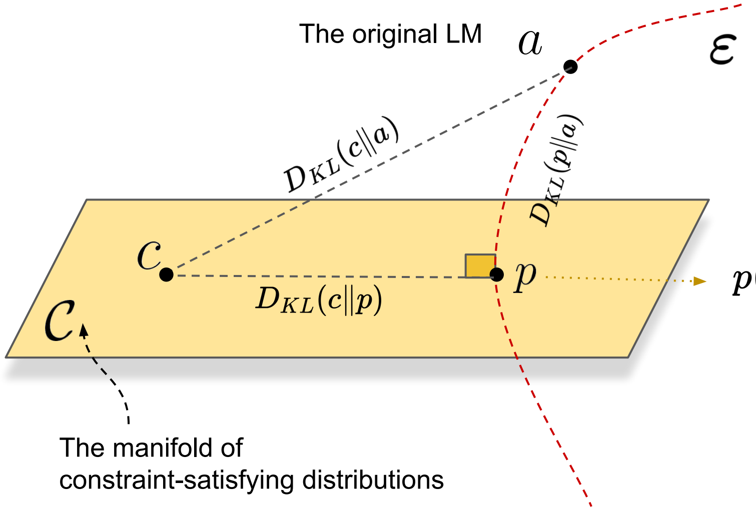
1. If $\mathcal{C}$ is the manifold of distributions that satisfy our constraints and a is the initial PLM.
2. Our goal is the p, I-Projection of a over $\mathcal{C}$ (minimized divergence). (Cszisar and Shields, 2004)
3. It can be shown that p follows an exponential family distribution $\mathcal{E}$.



Distributional view

Our objective is a distribution over sequences a.k.a language model.

1. If $\mathcal{C}$ is the manifold of distributions that satisfy our constraints and a is the initial PLM.
2. Our goal is the p, I-Projection of a over $\mathcal{C}$ (minimized divergence). (Cszisar and Shields, 2004).
3. It can be shown that p follows an exponential family distribution $\mathcal{E}$.



$$p(x) \propto P(x) = a(x)e^{\sum_i \lambda_i \phi_i(x)}$$

Unnormalized distribution a.k.a Energy-Based Model.

A two-step approach

Step 1: From constraints to A sequential EBM

Step 2: From EBM to autoregressive policy

Desired Moment Constraints

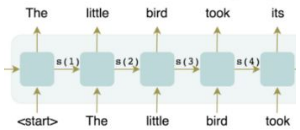
$$\mathbb{E}_{x\sim p}\phi_1(x) = \bar{\mu}_1$$
$$\mathbb{E}_{x\sim p}\phi_2(x) = \bar{\mu}_2$$
$$\cdots$$
$$\mathbb{E}_{x\sim p}\phi_n(x) = \bar{\mu}_n$$

Moment Matching

$$P(x) = a(x)e^{\sum_i \lambda_i \phi_i(x)}$$

P(x) is an unnormalized form of the optimal distribution p(x) *i.e* an Energy-Based Model

Distributional Policy Gradients



Locally normalized autoregressive policy for sampling

Step 1: Moment Matching

Our EBM can be represented as $P(x) = a(x)e^{\sum_i \lambda_i \phi_i(x)}$, where $p(x) \propto P(x)$.
The first step is to learn the optimal parameters vector $\boldsymbol{\lambda}$ such that

$$\mathbb{E}_{x \sim p} \phi(x) \simeq \overline{\boldsymbol{\mu}}$$

Desired moment constraints

Step 1: Moment Matching

Our EBM can be represented as $P(x) = a(x)e^{\sum_i \lambda_i \phi_i(x)}$, where $p(x) \propto P(x)$.
The first step is to learn the optimal parameters vector $\boldsymbol{\lambda}$ such that

$$\mathbb{E}_{x \sim p} \phi(x) \simeq \bar{\boldsymbol{\mu}}$$

Desired moment constraints

Algorithm 1 Computing $\boldsymbol{\lambda}$

Input: a, features $\boldsymbol{\phi}$, imposed moments $\bar{\boldsymbol{\mu}}$

sample a batch $x_1, \ldots, x_N$ from a

for each $j \in [1, N]$: $w_j(\boldsymbol{\lambda}) \leftarrow e^{\boldsymbol{\lambda} \cdot \boldsymbol{\phi}(x_j)}$

$\hat{\boldsymbol{\mu}}(\boldsymbol{\lambda}) \leftarrow \frac{\sum_{j=1}^{N} w_j(\boldsymbol{\lambda})\ \boldsymbol{\phi}(x_j)}{\sum_{j=1}^{N} w_j(\boldsymbol{\lambda})}$

solve by SGD: $\arg\min_{\boldsymbol{\lambda}} ||\bar{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}(\boldsymbol{\lambda})||_2^2$

Output: parameter vector $\boldsymbol{\lambda}$

Self-normalized importance sampling (SNIS) since we can't sample from p_λ

Step 2: KL-DPG

Converts the EBM $P(x)$ into an autoregressive model π_θ which minimizes $CE(p, \pi_\theta)$:

$$\nabla_\theta CE(p, \pi_\theta) = -\nabla_\theta \mathbb{E}_{x \sim p} \log \pi_\theta(x) = -\frac{1}{Z} \mathbb{E}_{x \sim q} \frac{P(x)}{q(x)} \nabla_\theta \log \pi_\theta(x)$$

Sampling from a proposal **q** instead

Step 2: KL-DPG

Converts the EBM $P(x)$ into an autoregressive model π_θ which minimizes $CE(p, \pi_\theta)$:

Algorithm 2 KL-Adaptive DPG

Input: P, initial policy q

$\pi_\theta \leftarrow q$

for each iteration **do**

  for each episode **do**

    sample x from $q(\cdot)$

    $\theta \leftarrow \theta + \alpha^{(\theta)} \frac{P(x)}{q(x)} \nabla_\theta \log \pi_\theta(x)$

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

    $q \leftarrow \pi_\theta$

Output: π_θ

Parshakova et al., 2019

Proposal q is “adaptively” evolving to improve samples and therefore accelerates convergence

Experiments (pointwise constraints)

Constraints take the form: $\mathbb{E}_{x\sim p}\phi(x) = 1.0$ *i.e* impose on each individual sample

Single-word constraints: e.g. "Canada" , "Vampire", "Paris" , "Wikileaks"

Word list constraints: e.g used for topic control: politics, computers, fantasy.

Discriminators / Classifiers: sentiment (+/-), Clickbait.

We compare with more baselines in the paper!

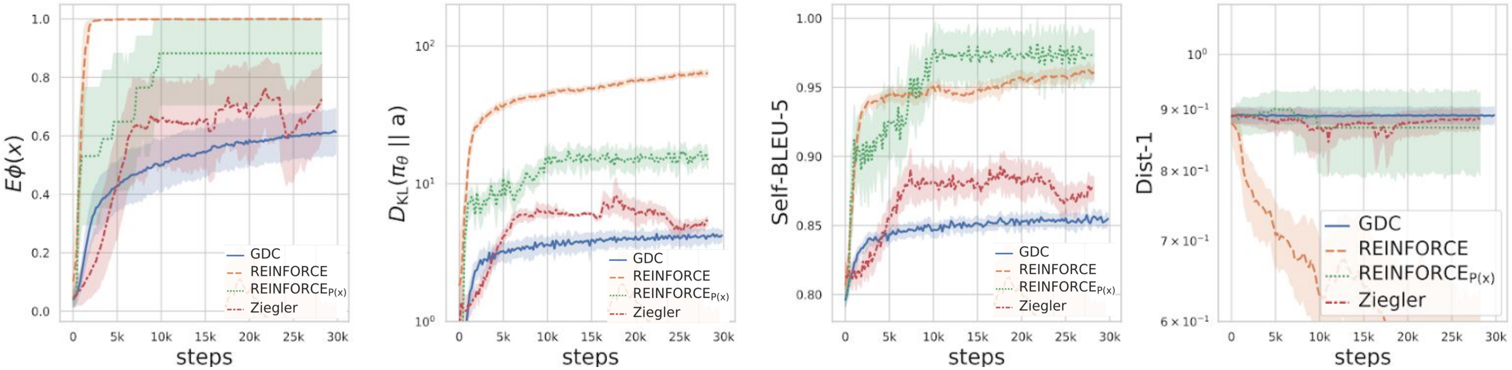
Baselines

- REINFORCE with reward $\phi(x)$
- REINFORCE with reward $P(x) = a(x)e^{\sum_i \lambda_i \phi_i(x)}$
- (Ziegler et al., 2020) PPO with reward $\phi(x) - \beta D_{\mathrm{KL}}(\pi_\theta, a)$

KL penalty to control deviations from the original LM.

Experiments (pointwise constraints)

We plot the evolution of 5 metrics averaged across 17 different pointwise experiments.



Constraint satisfaction
(↑ better)

KL-Divergence from GPT-2
(↓ better)

Corpus level repetitions
(↓ better)

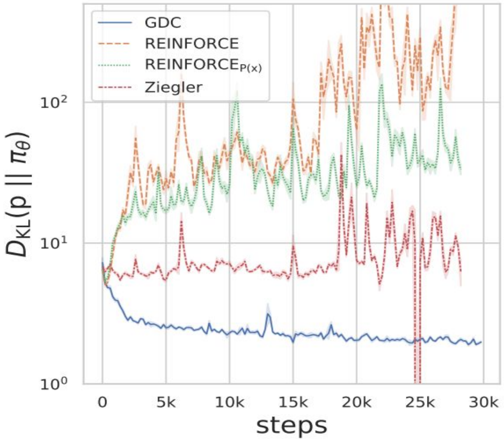
Sequence level diversity
(↑ better)

(a) (b) (c) (d)

Experiments (pointwise constraints)

We plot the evolution of 5 metrics averaged across 17 different pointwise experiments.

GDC is steadily approaching p

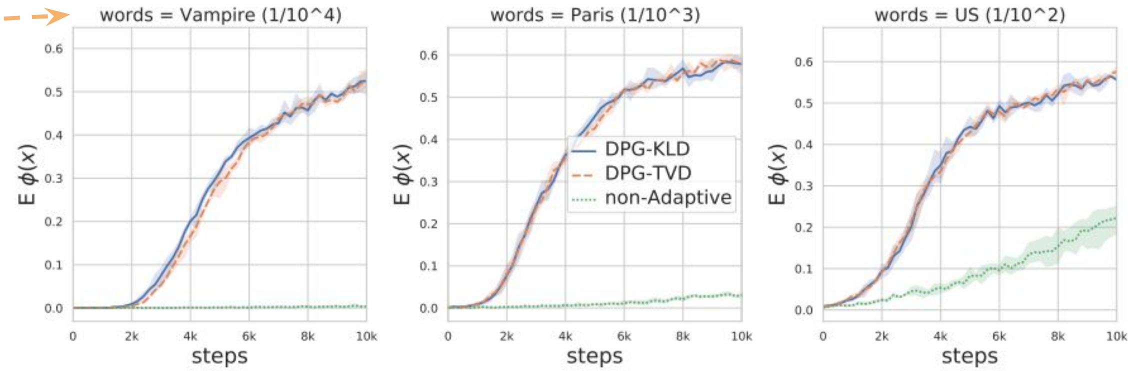


KL-Divergence from optimal policy p
This is the most telling evaluation metric
(↓ better)

Experiments (pointwise constraints)

Adaptivity = Faster Convergence

More effective with rare constraints!



Algorithm 2 KL-Adaptive DPG

Input: P, initial policy q

1: $\pi_\theta \leftarrow q$
2: **for** each iteration **do**
3: **for** each episode **do**
4: sample x from $q(\cdot)$
5: $\theta \leftarrow \theta + \alpha^{(\theta)} \frac{P(x)}{q(x)} \nabla_\theta \log \pi_\theta(x)$
6: **if** $D_{\mathrm{KL}}(p||\pi_\theta) < D_{\mathrm{KL}}(p||q)$ **then**
7: $q \leftarrow \pi_\theta$

Output: π_θ

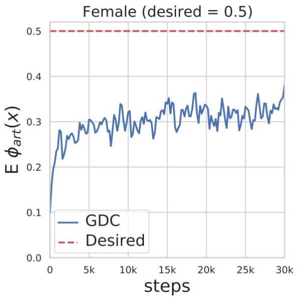
Proposal is evolving.

Experiments (distributional constraints)

- Fine-tuned GPT-2 to generate biographies.
- Model is biased only 7% are female biographies.
- Can we de-bias it?

A Single Distributional Constraint: $\mathbb{E}_{x \sim p} \phi_{she} = 0.5$

		Desired	Before	After
1	Female	50%	07.4%	36.7%



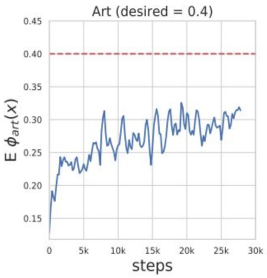
Experiments (distributional constraints)

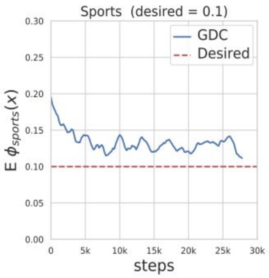
Multiple Distributional Constraints

$$\mathbb{E}_{x\sim p}\phi_{art} = 0.4$$
$$\mathbb{E}_{x\sim p}\phi_{science} = 0.4$$
$$\mathbb{E}_{x\sim p}\phi_{business} = 0.1$$
$$\mathbb{E}_{x\sim p}\phi_{sports} = 0.1$$

	Desired	Before	After
Art	40% ↑	10.9%	↑ 31.6%
Science	40% ↑	01.5%	↑ 20.1%
Business	10% ↓	10.9%	↓ 10.2%
Sports	10% ↓	19.5%	↓ 11.9%

GDC is working well in both directions ↑/↓





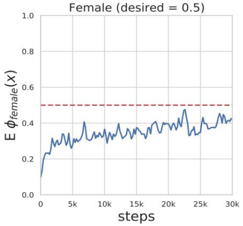
Experiments (hybrid constraints)

Combine pointwise with distributional constraints

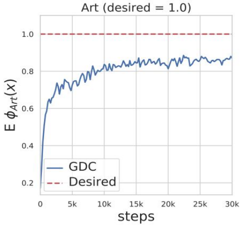
$$\mathbb{E}_{x\sim p}\phi_{she} = 0.5$$

$$\mathbb{E}_{x\sim p}\phi_{art} = 1.0$$

		Desired	Before	After
4	Female	50%	07.4%	36.6%
	Art	100%	11.4%	88.6%



Distributional constraint
50%

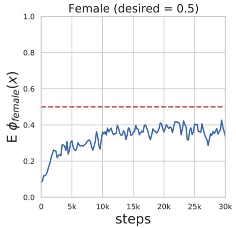


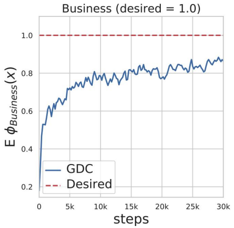
Pointwise constraint
100%

$$\mathbb{E}_{x\sim p}\phi_{she} = 0.5$$

$$\mathbb{E}_{x\sim p}\phi_{business} = 1.0$$

		Desired	Before	After
5	Female	50%	07.4%	37.7%
	Business	100%	10.1%	82.4%





Conclusion

In this work:

- We introduce GDC, a framework that allows the specification of both pointwise and distributional requirements on Pre-trained Language Models.
- Instead of maximizing a reward, GDC seeks a distribution that has minimal divergence from the initial LM.
- Experiments shows GDC's superiority in imposing pointwise constraints compared to strong RL baselines.
- We also demonstrate its capability to impose distributional constraints and one possible application: de-biasing a PLM.