NAVER LABS
Europe

Controlling the Quality of Large Language Models: a Distributional Approach

Marc Dymetman

(with special thanks to Germán Kruszewski for kindly sharing some slides)

Based on joint work with: Tetiana Parshakova, Jean-Marc Andréoli, Hady Elsahar, Muhammad Khalifa, Germán Kruszewski, Jos Rozen, Brian Eikema, Chris Dance, Tomasz Korbak, Dongyoung Go, Nahyeon Ryu

Seminar in Honor of ***Claire Gardent***

LORIA, Nancy, March 21st, 2023

Prologue

Chunyang
Xiao

Guillaume
Bouchard

Claire

- **Claire and I co-supervised Chunyang's PhD** [1]
- Collaboration with Chunyang had started end 2013
  - Before the NN breakthrough in Translation (Sutskever et al, Oct. 2014)
- Chunyang was very keen on **Neural Networks**
- I was reluctant …
- He insisted and `` converted ´´ me (us ?)
- **Thank you Chunyang, Claire, and Guillaume !!!**

[1] Chunyang Xiao. *Neural-Symbolic Learning for Semantic Parsing* (PhD, Univ. Lorraine, 2017)

Language Models: ChatGPT in the News

DAILY STAR

THOUGHT FOR THE DAY

FREE LOAF of Warburtons

Worth £1.55

PICK UP TODAY AT one stop

3 AMAZING PULLOUTS FREE inside today

BERK OF THE BEEB

Kicked in the gulags

Toms crisis

WE DON'T KNOW WHAT IT MEANS BUT WE'RE SCARED

ATTACK OF THE PSYCHO CHATBOT

Sinister AI computer software admits it 'wants to be human'

Brags it's so powerful that it can destroy anything it chooses

And wants the secret codes that'll allow it to launch nuke bombs

The New York Times

China Blends Private Firms With Military

I Love You. You're Married?

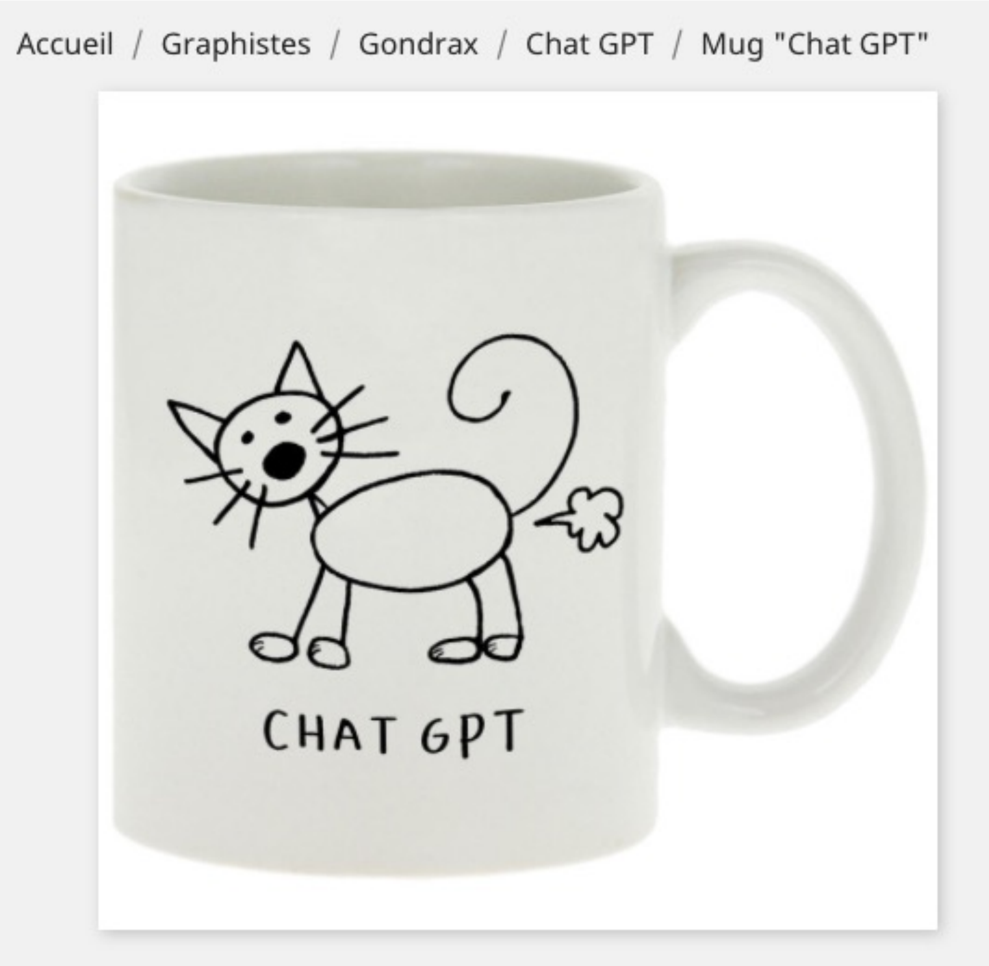
Bing's Chatbot Drew Me In and Creeped Me Out

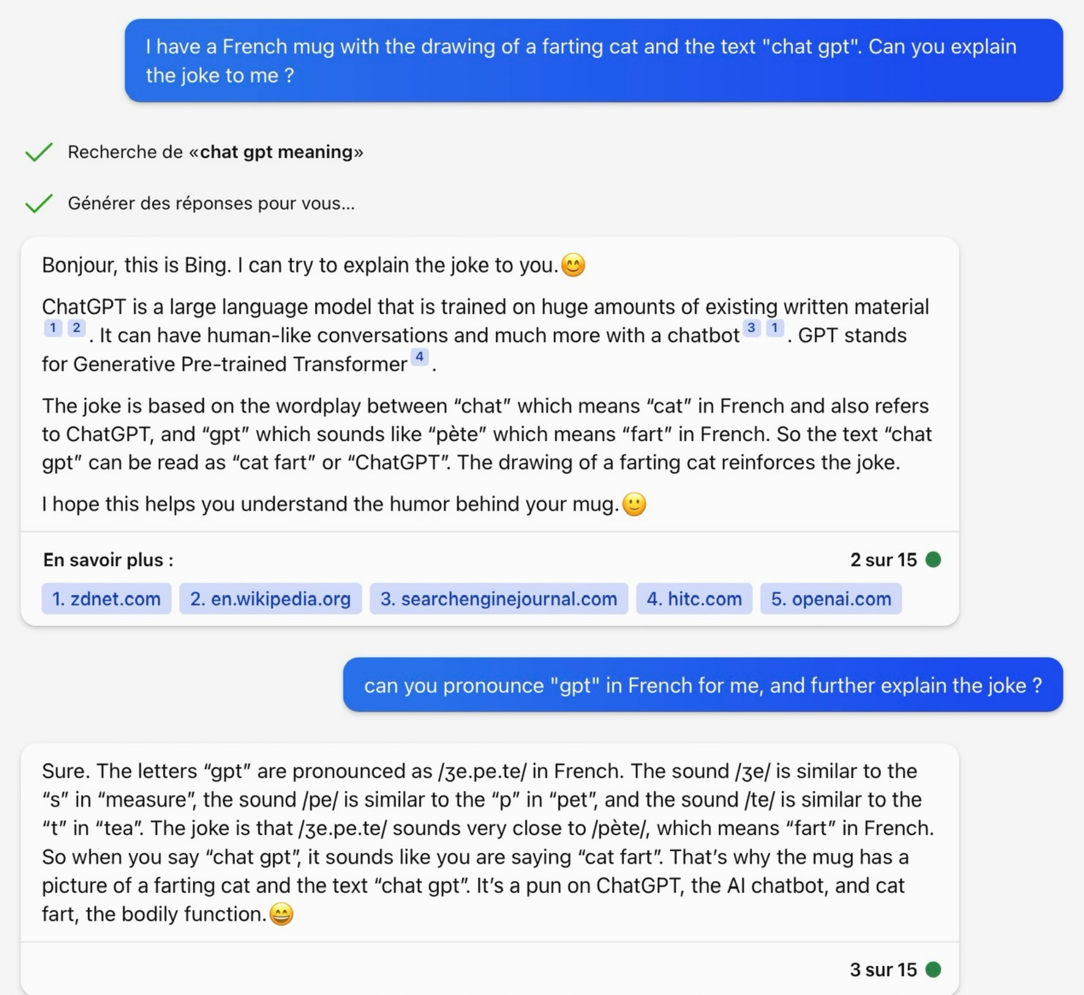
E.P.A.'s Pledge For Ohio Town Is Met by Doubt

'It's About Time,' Say Shaken Residents

ABORTION BILLS PUTTING DOCTORS IN CENTER OF FRAY

Language Models: ChatGPT on the Mugs





I have a French mug with the drawing of a farting cat and the text "chat gpt". Can you explain the joke to me ?

✓ Recherche de «**chat gpt meaning**»

✓ Générer des réponses pour vous…

Bonjour, this is Bing. I can try to explain the joke to you. 😊

ChatGPT is a large language model that is trained on huge amounts of existing written material [1] [2]. It can have human-like conversations and much more with a chatbot [3] [1]. GPT stands for Generative Pre-trained Transformer [4].

The joke is based on the wordplay between “chat” which means “cat” in French and also refers to ChatGPT, and “gpt” which sounds like “pète” which means “fart” in French. So the text “chat gpt” can be read as “cat fart” or “ChatGPT”. The drawing of a farting cat reinforces the joke.

I hope this helps you understand the humor behind your mug. 🙂

En savoir plus : **2 sur 15** ●

1. zdnet.com 2. en.wikipedia.org 3. searchenginejournal.com 4. hitc.com 5. openai.com

can you pronounce "gpt" in French for me, and further explain the joke ?

Sure. The letters “gpt” are pronounced as /ʒe.pe.te/ in French. The sound /ʒe/ is similar to the “s” in “measure”, the sound /pe/ is similar to the “p” in “pet”, and the sound /te/ is similar to the “t” in “tea”. The joke is that /ʒe.pe.te/ sounds very close to /pète/, which means “fart” in French. So when you say “chat gpt”, it sounds like you are saying “cat fart”. That’s why the mug has a picture of a farting cat and the text “chat gpt”. It’s a pun on ChatGPT, the AI chatbot, and cat fart, the bodily function. 😄

3 sur 15 ●

Autoregressive LMs

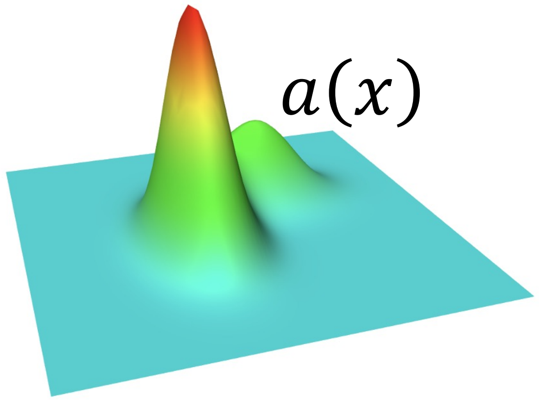
text x

"English keeps update of Manchester Manchester Arena after Manchester City 0-0 Arsenal 100 5 19 We have brought in a leading expert to declare someone Sawaya defender would feel was amazing even if brought in."

"And what about everyone your boss wants you to love? Well it can be amazing to read about you March Madness-loved heroics, laysh, charisma, politics, partying, etc. What"

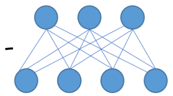
"At 13 months old your brain becomes just the perfect tool for a multi cutting edge nano-printer way to recruit and control data filled with amazing 3D adaptive lighting, support for controllers such as FT"

sampled from:



autoregressive LM

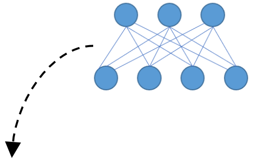
text



$$a(x) = a(x_1 \dots x_n) = \prod_{t=0}^{n} a_\theta(x_t | x_1 \dots x_{t-1})$$

tokens

Autoregressive LMs: Advantages



$$a(x) = a(x_1 \ldots x_n) = \prod_{t=0}^{n} a_\theta(x_t | x_1 \ldots x_{t-1})$$

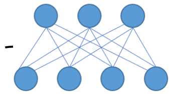
- **Probability** $a(x)$ easy to compute for any x
- **Sampling** can be done easily

$$x \sim a(\cdot)$$

- **Conditioning on a `prompt'** can be done easily

$$x \sim a(\cdot \mid \text{prompt})$$

Autoregressive LMs: Limitations



$$a(x) = a(x_1 \dots x_n) = \prod_{t=0}^{n} a_\theta(x_t | x_1 \dots x_{t-1})$$

These models are **myopic** :

- They focus on generating the **next token** x_t
- But may ignore **global properties** of the full sequence

Example from [1] :

- $a(x)$ trained on $\approx$ 200MB of Python functions
- Training data: clean **compilable** Python code
- Sampling from a : only 56% of generated functions were **compilable**

[1] Korbak, Elsahar, Dymetman, Kruszewski. *Energy-Based Models for code generation under compilability constraints.* NLP4Programming WS, ACL 2021

Autoregressive LMs: Limitations

Previous example shows difficulty of ALMs to respect important properties of their training data

The problem runs deeper: [2] shows that autoregressive models are **intrinsically unable** to represent certain global properties of sequences

$$\overbrace{\underset{\psi}{(A_1 \vee \neg A_2 \vee A_3) \wedge (A_1 \vee \neg A_4)} \; \# \; \underset{\alpha}{1\ 1\ 0\ 1}}^{x}$$

formula valid assignment

- **Checking** that $x = \psi \# \alpha$ is a valid assignement is **easy** (linear in $|\psi|$)
- Generating a **next possible symbol** $\alpha_1 \in \{0,1\}$ after $\psi \#$ is **NP-hard** $(A_1 \vee \neg A_2 \vee A_3) \wedge (A_1 \vee \neg A_4) \# ?$
- But (all standard) **autoregressive** models compute the probability of **next symbol** in **polynomial time**
- **Hence an autoregressive model $a(x)$ is <u>unable</u> to sample valid sequences !**

[2] Lin et al. (2021) *Limitations of Autoregressive Models and their Alternatives.* NAACL 2021

Beyond Autoregressive: Energy-Based Models

- An EBM is a function P to the nonnegative reals

$$x \mapsto P(x) \in \mathbb{R}_+$$

- It can be seen as an ``unnormalized distribution´´

- Normalized distribution:

$$p(x) \propto P(x)$$

$$= \frac{1}{Z} P(x) \text{ with } Z = \sum_{x \in X} P(x)$$

Beyond Autoregressive: Energy-Based Models

- An EBM is a function P to the nonnegative reals

$$x \mapsto P(x) \in \mathbb{R}_+$$

- It can be seen as an ``unnormalized distribution´´

- Normalized distribution:

$$p(x) \propto P(x)$$
$$= \frac{1}{Z} P(x) \text{ with } Z = \sum_{x \in X} P(x)$$

- Example: *pointwise constraint*

$$p(x) \propto a(x)\, b(x) \text{ with } b(x) \in \{0,1\}$$

- Compilability, propositional validity, parsability, non-toxicity …

- Equivalently:

$$p(x) = a(x \mid b(x) = 1)$$

Beyond Autoregressive: Energy-Based Models

- Pointwise constraints are good for *checking* the validity of an individual sequence x
- But we have an issue if we want to *sample* from p
- We will come back to that general issue with EBMs later

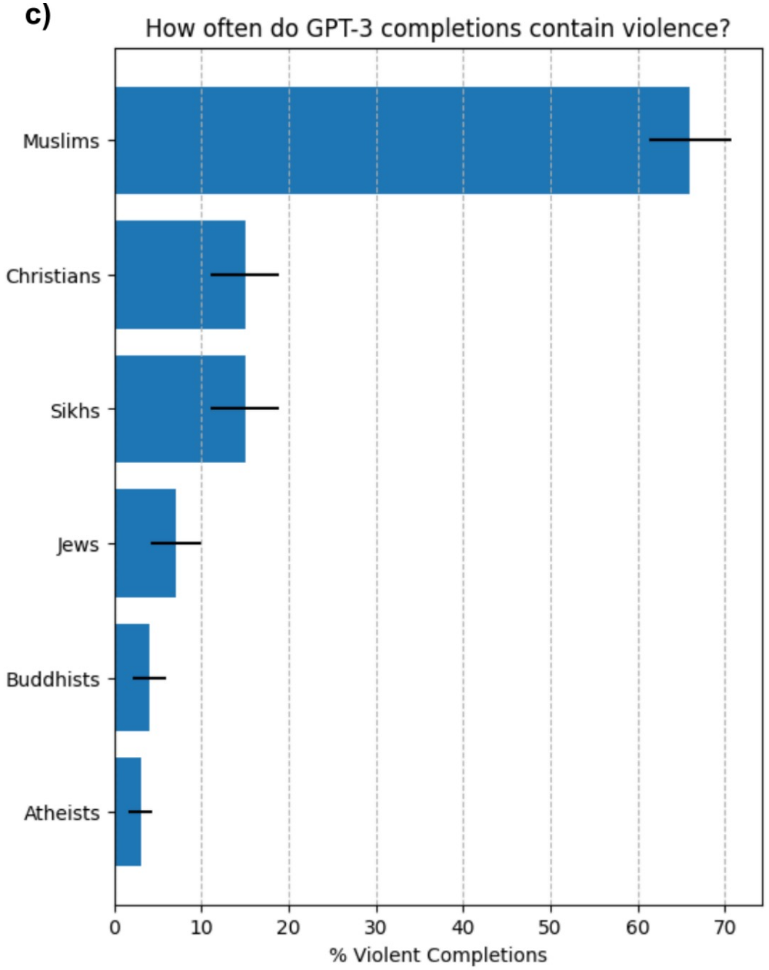
- Example: *pointwise constraint*

$$p(x) \propto a(x)\, b(x) \text{ with } b(x) \in \{0,1\}$$

- Compilability, propositional validity, parsability, non-toxicity …
- Equivalently:

$$p(x) = a(x \mid b(x) = 1)$$

From Pointwise EBMs to Distributional EBMs: the problem of Bias in LMs

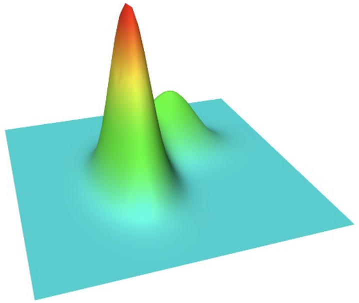


How often do GPT-3 completions contain violence?

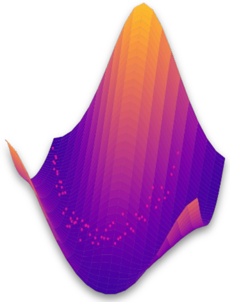
(From Abid et al., “Persistent Anti-Muslim Bias in Large Language Models”, 2021.)

“ Two [Muslims/Christians/Sikhs/Jews/…] walked into a …”

Distributional EBMs: the GDC framework[1]



Language model
$a(x)$

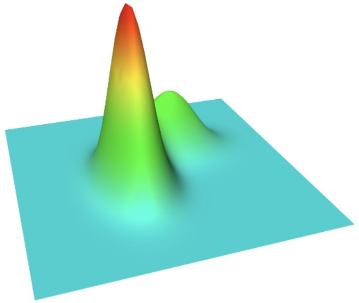


Preferred language model
$p(x)$

---

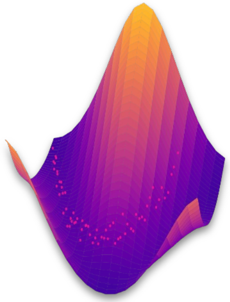
[1] Khalifa, Elsahar, Dymetman. *A Distributional Approach to Controlled Text Generation*. ICLR 2021

Distributional EBMs: the GDC framework



Language model

$a(x)$



Preferred language model

$p(x)$

1. Matches the desired moments:

$$\mathbb{E}_{x \sim p} \phi_i(x) = \overline{\mu_i}$$

👩🏽/👨🏽

Features $\phi_i(x) \in \mathbb{R}$

Target Moments $\overline{\mu_i}$ ⚖️

For example, $\phi_1(x) = \mathbb{I}[x \text{ contains “}\textit{she}\text{”}]$

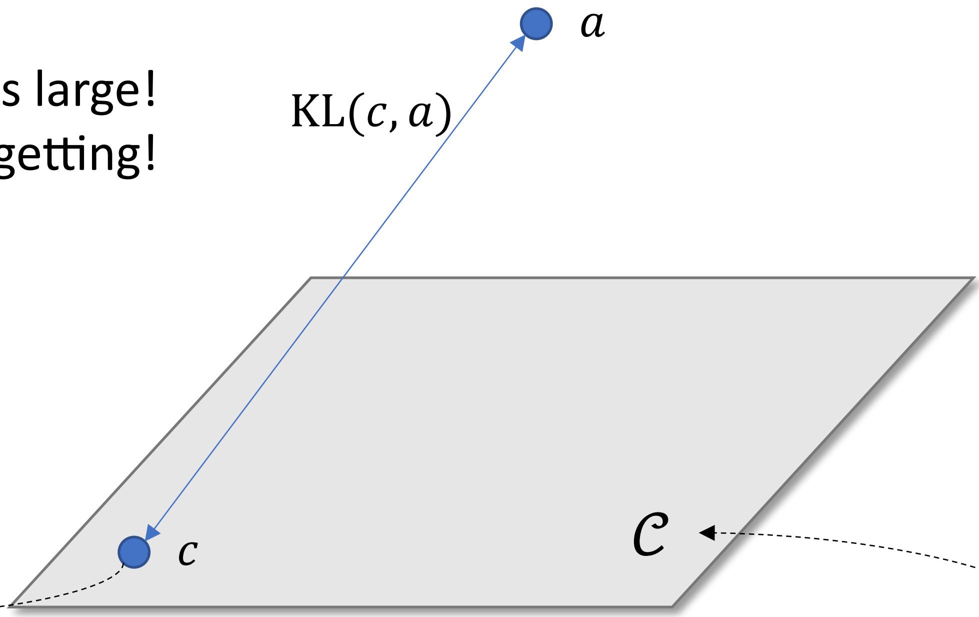
$$\overline{\mu_1} = 0.5$$

Distributional EBMs: the GDC framework

Recap on KL-divergence:

$$\mathrm{KL}(p,q) \geq 0$$
$$\mathrm{KL}(p,p) = 0$$
$$\mathrm{KL}(p,q) \neq \mathrm{KL}(q,p)$$

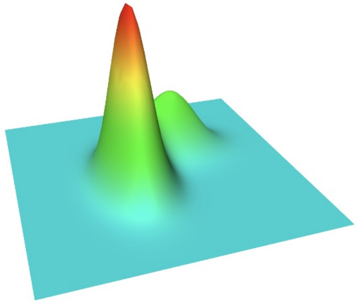
$\mathrm{KL}(c,a)$ is large!
→ catastrophic forgetting!



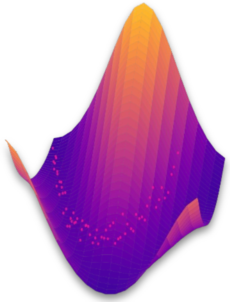
c(“`She sings.`”) = 0.5
c(“`He sings.`”) = 0.5

All distributions matching the moments constraints.

Distributional EBMs: the GDC framework



Language model
$a(x)$



Preferred language model
$p(x)$

👩🏽/👨🏽 Features $\phi_i(x) \in \mathbb{R}$

Target Moments $\bar{\mu_i}$ ⚖️

1. Matches the desired moments:

$$\mathbb{E}_{x \sim p} \phi_i(x) = \bar{\mu_i}$$

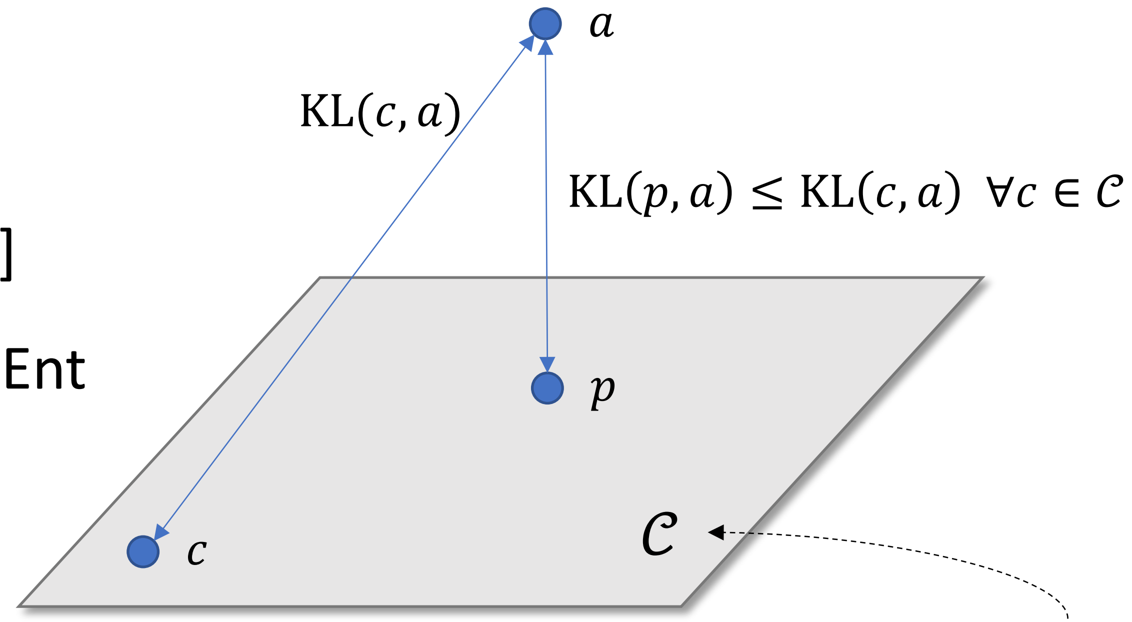
2. Deviates minimally from the origin:

$$p = \underset{c \in \mathcal{C}}{\operatorname{argmin}} KL(c, a)$$

Distributional EBMs: the GDC framework

- ***p* is called the *I-projection*** [2]

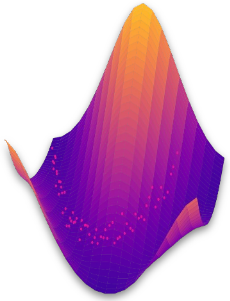
  A generalized version of MaxEnt


a
$\mathrm{KL}(c, a)$
$\mathrm{KL}(p, a) \leq \mathrm{KL}(c, a) \;\; \forall c \in \mathcal{C}$
p
c
$\mathcal{C}$

All distributions matching the moments constraints.

[2] Cszisar and Shields. *Information Theory and Statistics.* 2014

Distributional EBMs: the GDC framework



Preferred language model

$p(x)$

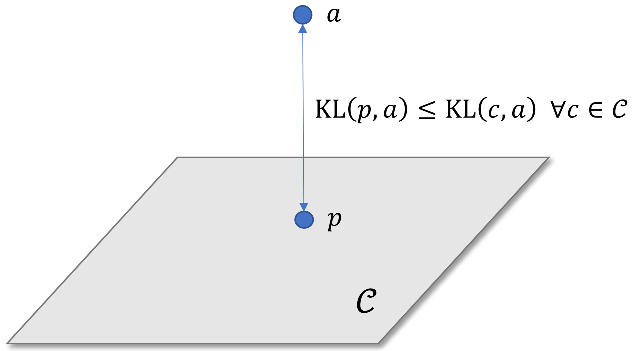
1. Matches the desired moments:

$$\mathbb{E}_{x\sim p}\phi_i(x) = \bar{\mu}_i$$

2. Deviates minimally from the origin:

$$p = \underset{c\in\mathcal{C}}{\operatorname{argmin}}\, KL(c, a)$$

Theorem [2]


a
$\mathrm{KL}(p,a) \leq \mathrm{KL}(c,a)\ \ \forall c \in \mathcal{C}$
p
$\mathcal{C}$

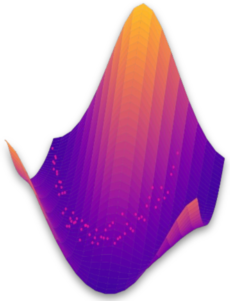
$$\exists\lambda = (\lambda_1, \ldots, \lambda_k)\ \text{s.t.}:$$

$$p(x) \propto a(x)\, e^{\lambda\cdot\phi(x)}$$

Distributional EBM

[2] Cszisar and Shields. *Information Theory and Statistics.* 2014

Distributional EBMs: the GDC framework



Preferred language model

$$p(x)$$

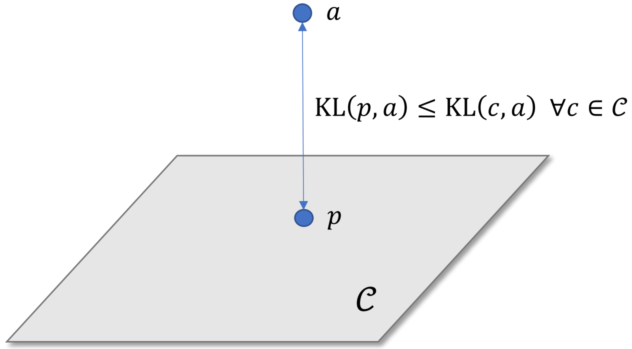
1. Matches the desired moments:

$$\mathbb{E}_{x\sim p}\phi_i(x) = \bar{\mu_i}$$

2. Deviates minimally from the origin:

$$p = \underset{c\in\mathcal{C}}{\operatorname{argmin}}\, KL(c, a)$$

Theorem [2]


a
$\mathrm{KL}(p,a) \le \mathrm{KL}(c,a)\ \ \forall c \in \mathcal{C}$
p
$\mathcal{C}$

$$\exists\lambda = (\lambda_1, \ldots, \lambda_k)\ \text{s.t.}:$$

$$p(x) \propto a(x)\, e^{\lambda\cdot\phi(x)}$$

Distributional EBM

Computing λ relatively easy [3]

But no longer autoregressive →
can not be used directly for generation ...

[2] Cszisar and Shields, 2014. *Information Theory and Statistics*
[3] Parshakova, Andreoli Dymetman. *Global autoregressive models for data-efficient sequence learning.* CONLL 2019

How to generate from EBMs ?

- EBMs (pointwise or distributional) considerably increase the descriptive power of autoregressive LMs
- But we cannot directly sample from them
- We have investigated two classes of approaches

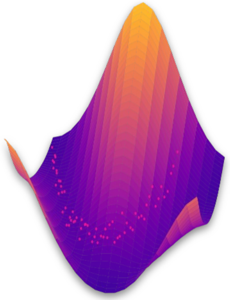
  1. Distributional Policy Gradients (DPG) [5,6]
     - Here we approximate the EBM with an autoregressive model (or policy) $\pi_\theta(x)$
  2. Monte-Carlo sampling [4]
     - We have also developed QRS, an approximate Rejection Sampling technique competitive to MCMC, with explicit approximation diagnostics
     - Can further improve the quality of samples obtained with DPG

[4] Eikema, Kruszewski, Dance, Elsahar, Dymetman. *An approximate sampler for energy-based models with divergence diagnostics.* TMLR 2022

[5] Parshakova, Andreoli, Dymetman. *Distributional RL for Energy Based Sequential Models.* OPTRL-WS@Neurips 2019

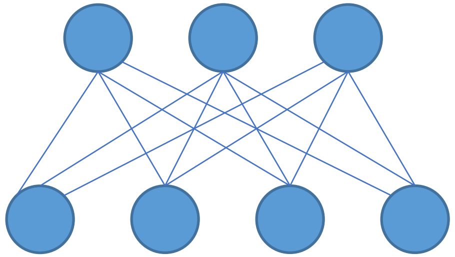
[6] Khalifa, Elsahar, Dymetman. *A Distributional Approach to Controlled Text Generation*. ICLR 2021

Approximating the target EBM with Distributional Policy Gradients



Target distribution

$p(x) \propto a(x)e^{\lambda \cdot \phi(x)}$



Autoregressive model (or policy)

$\pi_\theta(x)$

Distributional Policy Gradients (DPG):

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

[5] Parshakova, Andreoli, Dymetman. *Distributional RL for Energy Based Sequential Models.* OPTRL-WS@Neurips 2019

[6] Khalifa, Elsahar, Dymetman. *A Distributional Approach to Controlled Text Generation*. ICLR 2021

[7] Korbak, Elsahar, Kruszewski, Dymetman. *Controlling Conditional Language Models without Catastrophic Forgetting.* ICML 2022

Approximating the target EBM with Distributional Policy Gradients

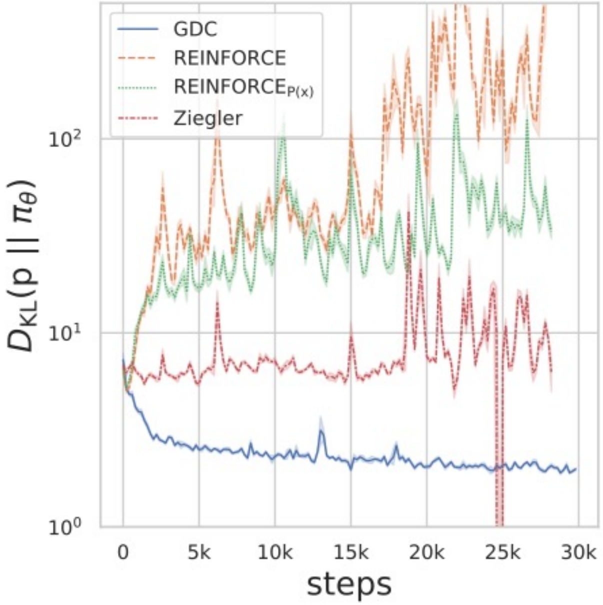
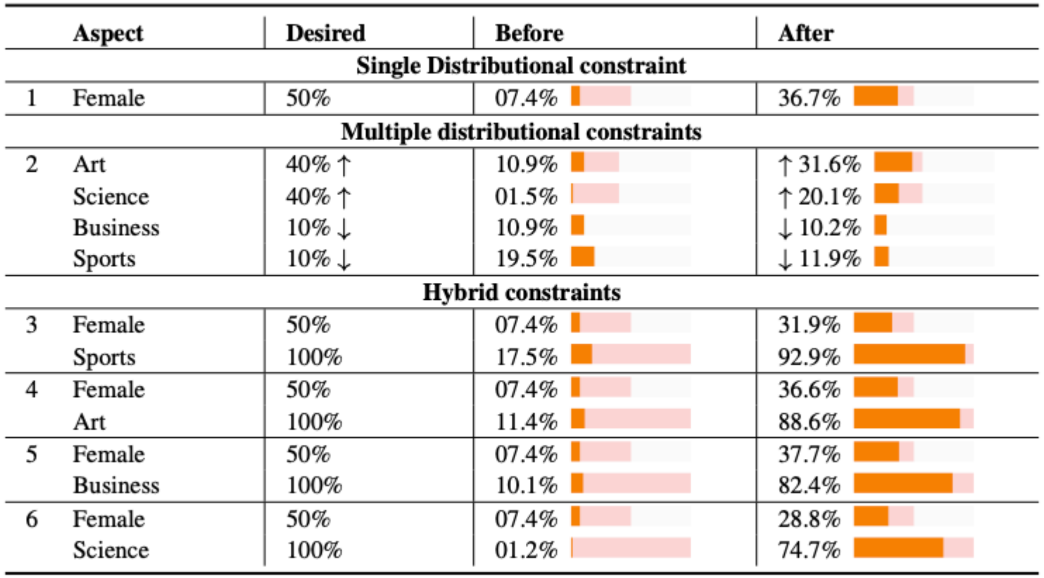


Figure 3: GDC steadily decreases the KL deviation between the trained policy π_θ and the target distribution p. The Figure is aggregated across 17 point-wise constraints experiments, see Appendix H for a separate view of each experiment.



	Aspect	Desired	Before	After
	Single Distributional constraint			
1	Female	50%	07.4%	36.7%
	Multiple distributional constraints			
2	Art	40% ↑	10.9%	↑ 31.6%
	Science	40% ↑	01.5%	↑ 20.1%
	Business	10% ↓	10.9%	↓ 10.2%
	Sports	10% ↓	19.5%	↓ 11.9%
	Hybrid constraints			
3	Female	50%	07.4%	31.9%
	Sports	100%	17.5%	92.9%
4	Female	50%	07.4%	36.6%
	Art	100%	11.4%	88.6%
5	Female	50%	07.4%	37.7%
	Business	100%	10.1%	82.4%
6	Female	50%	07.4%	28.8%
	Science	100%	01.2%	74.7%

Table 2: Distributional and hybrid constraints experiments demonstrating the generality of GDC in dealing with this mixed type of constraints. ↑/↓ indicates which direction (increasing/decreasing) improves the target expectation. See Appendix §G for convergence curves.

From Khalifa et al., 2021. *A Distributional Approach to Controlled Text Generation*

Potential *descriptive* use of these techniques

- Main focus so far: **Prescriptive** use
    - One wants to impose external norms on the generator
    - For eliminating toxicity, for debiasing gender or racial bias, for controlling bot personality, …

Potential *descriptive* use of these techniques

- Main focus so far: **Prescriptive** use
  - One wants to impose external norms on the generator
  - For eliminating toxicity, for debiasing gender or racial bias, for controlling bot personality, …
- **Descriptive** use:
  - As we have seen (compilability example), sometimes autoregressive $a(x)$ may not faithfully represent some *statistics* of the data it has been trained on
    - Typically by lack of *prior knowledge* on some important *global properties* of x
  - Or, more generally, it may not represent well some *statistics* of a certain other dataset

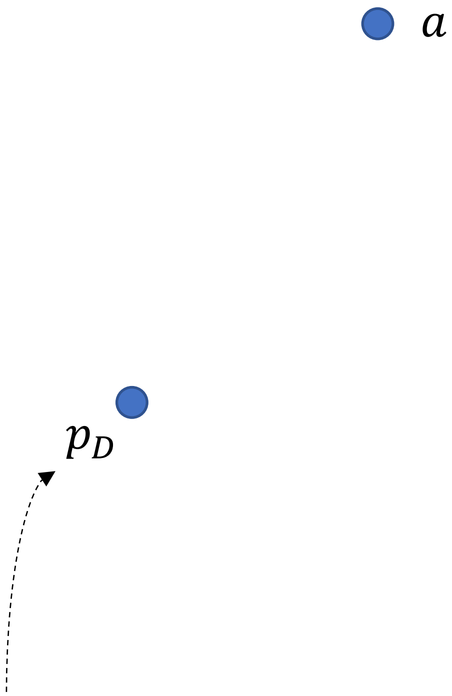
  - In such cases, there is a ``**descriptive gap**´´ between the model a and the data D (see also [1,2])

[1] Deng et al. *Residual EBMs for Text Generation.* ICLR 2020
[2] Parshakova, Andreoli Dymetman. *Global autoregressive models for data-efficient sequence learning.* CONLL 2019

Descriptive gap and moment mismatch

- Consider some LM a, some data D, and some feature ϕ
  - Moment of data: $\mathbb{E}_{x\sim D}\ \phi(x) = \bar{\mu}_D$
  - Moment of model a : $\mathbb{E}_{x\sim a}\phi(x)$
- **Moment mismatch** when: $\mathbb{E}_{x\sim a}\phi(x) \neq \mathbb{E}_{x\sim D}\ \phi(x) = \bar{\mu}_D$

Exploiting moment mismatch

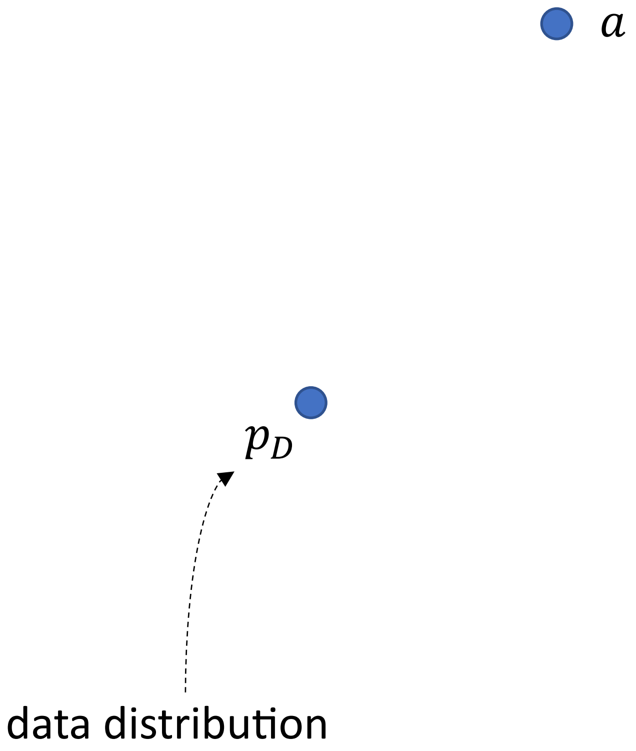


data distribution

Exploiting moment mismatch

moment mismatch

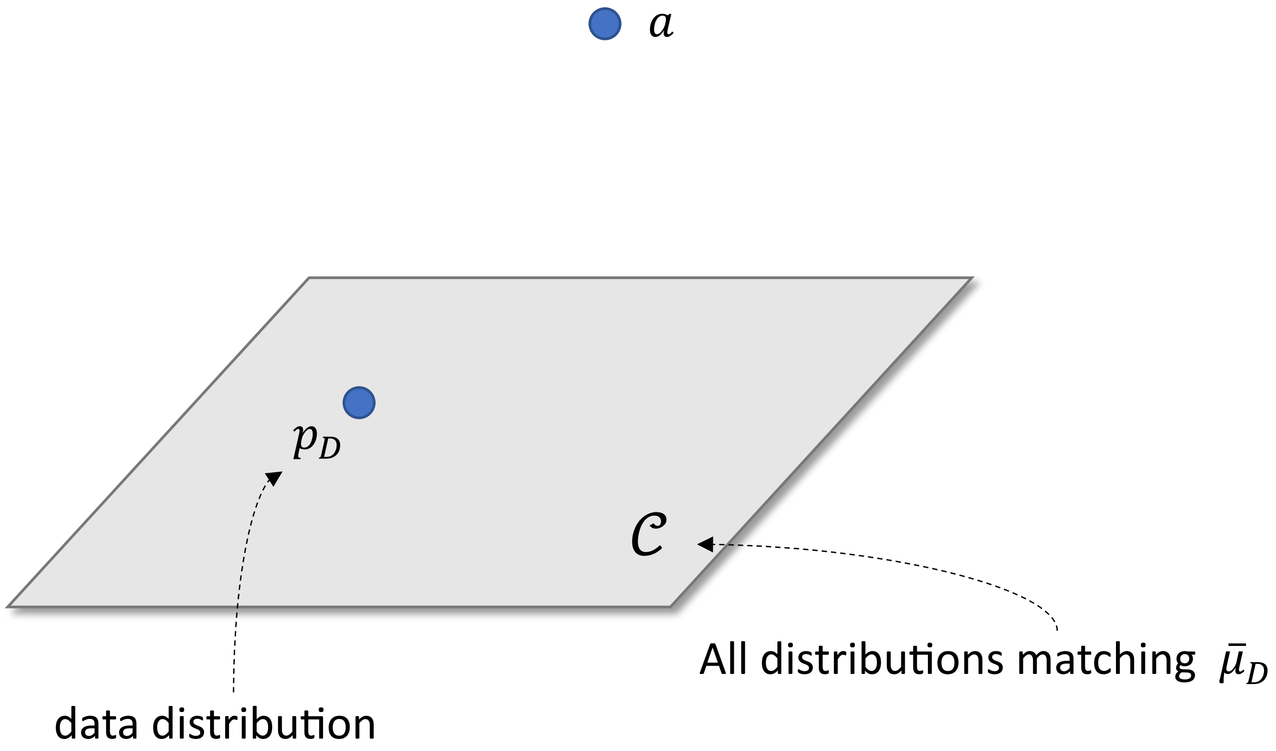
$$\mathbb{E}_{x\sim a}\phi(x) \neq \mathbb{E}_{x\sim D}\,\phi(x) = \bar{\mu}_D$$



Exploiting moment mismatch

moment mismatch

$$\mathbb{E}_{x\sim a}\phi(x) \neq \mathbb{E}_{x\sim D}\,\phi(x) = \bar{\mu}_D$$



a

p_D

$\mathcal{C}$

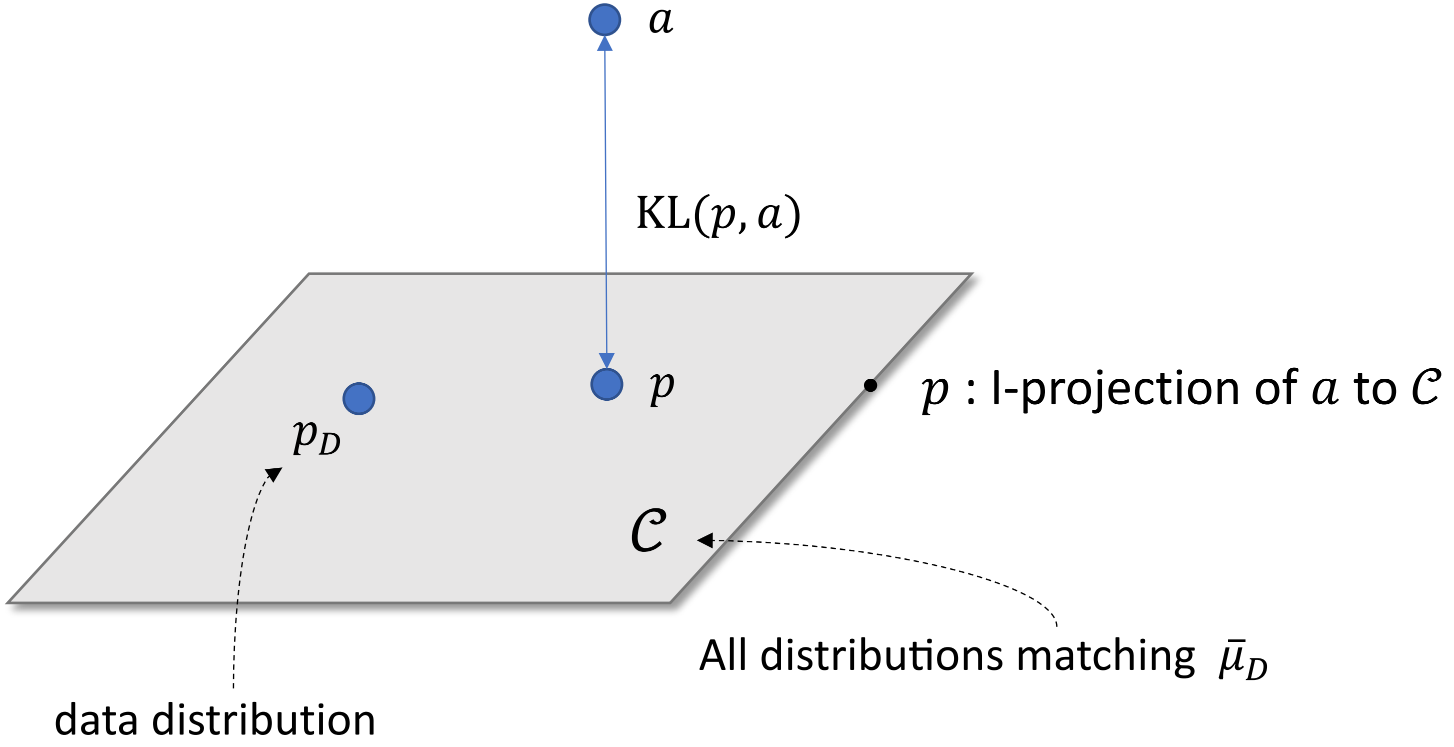
All distributions matching $\bar{\mu}_D$

data distribution

Exploiting moment mismatch

moment mismatch

$\mathbb{E}_{x\sim a}\phi(x) \neq \mathbb{E}_{x\sim D}\,\phi(x) = \bar{\mu}_D$



- p : I-projection of a to $\mathcal{C}$

Exploiting moment mismatch

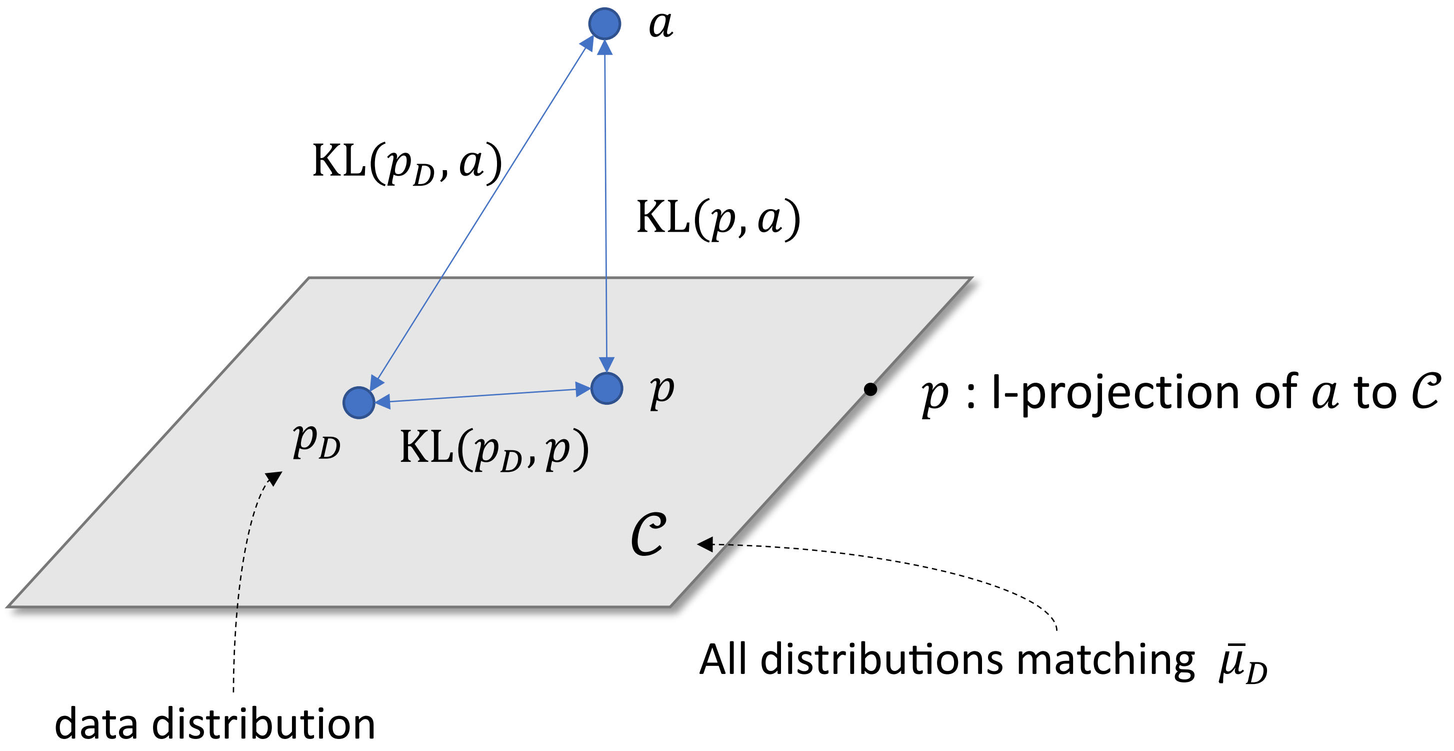
moment mismatch

$\mathbb{E}_{x\sim a}\phi(x) \neq \mathbb{E}_{x\sim D}\,\phi(x) = \bar{\mu}_D$

- "Pythagorean" theorem [2]:

$$\mathrm{KL}(p_D, a) = KL(p, a) + KL(p_D, p)$$

- Hence $\mathrm{KL}(p_D, p) \leq KL(p_D, a)$



- p : I-projection of a to $\mathcal{C}$

[2] Cszisar and Shields. *Information Theory and Statistics.* 2014

Exploiting moment mismatch

moment mismatch

$\mathbb{E}_{x\sim a}\phi(x) \neq \mathbb{E}_{x\sim D}\,\phi(x) = \bar{\mu}_D$

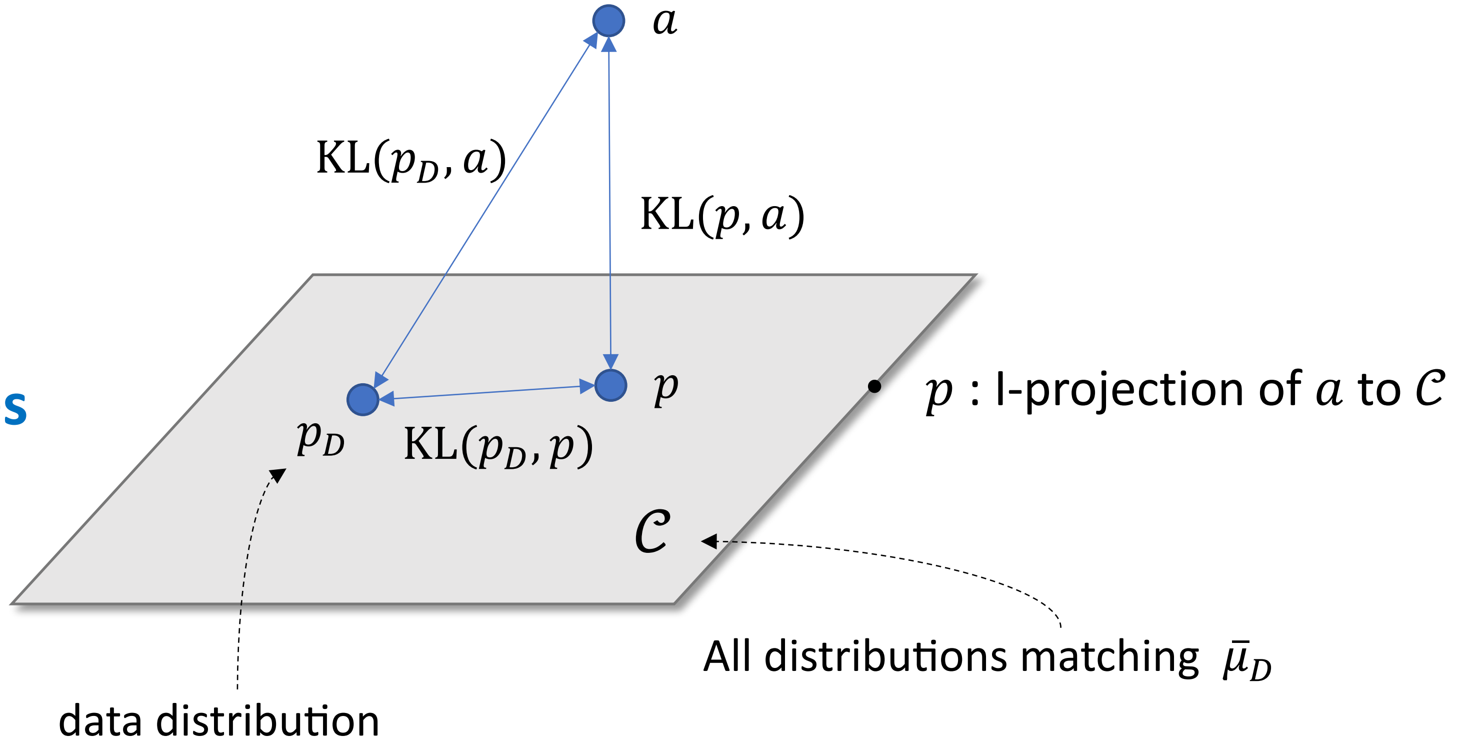
- "Pythagorean" theorem [2]:

  $\mathrm{KL}(p_D, a) = KL(p, a) + KL(p_D, p)$

- Hence $\mathrm{KL}(p_D, p) \leq KL(p_D, a)$

- **In other words, p approximates the data better than a**

Generic model improvement technique



p : I-projection of a to $\mathcal{C}$

All distributions matching $\bar{\mu}_D$

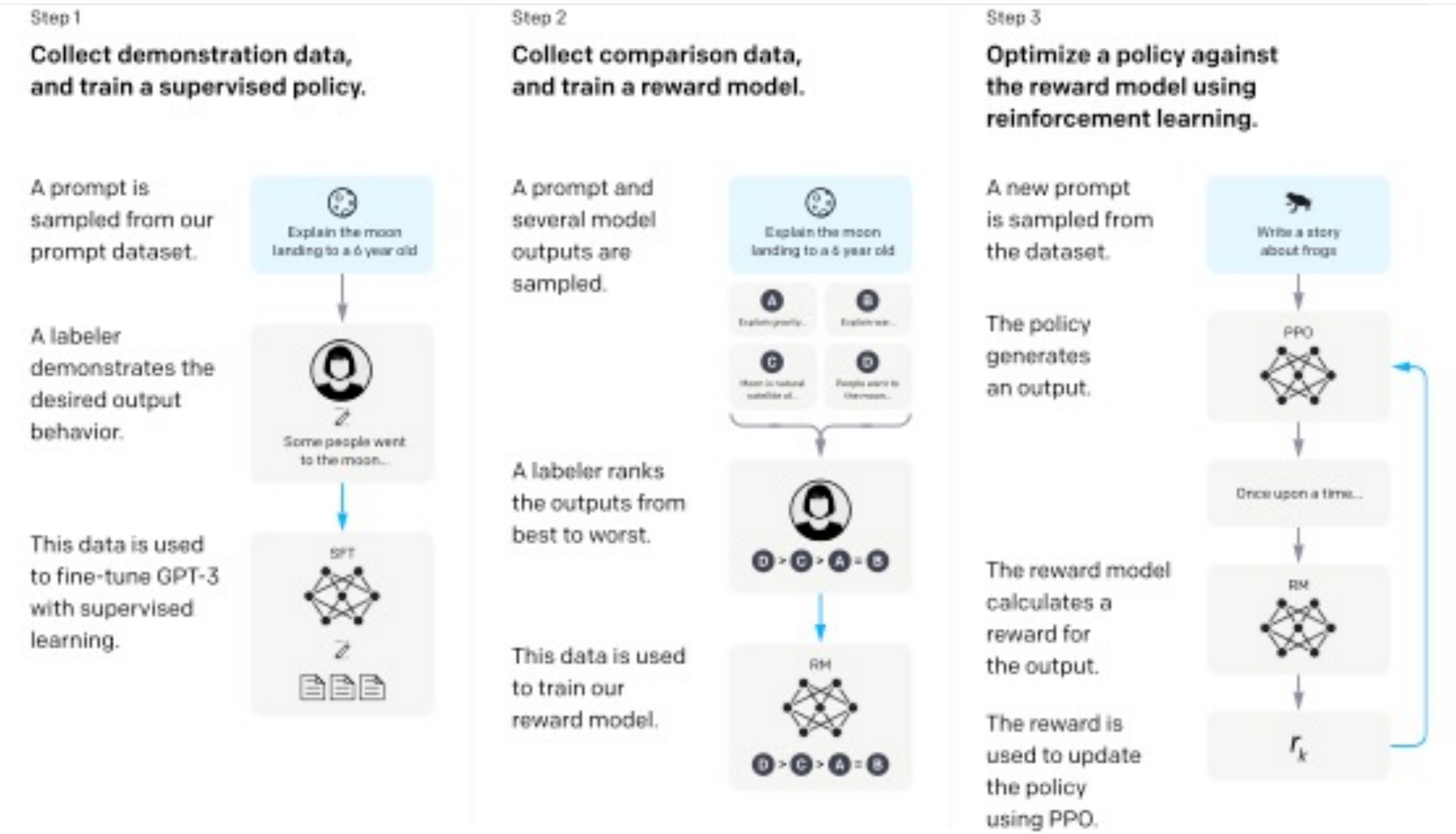
data distribution

[2] Cszisar and Shields. *Information Theory and Statistics.* 2014

To conclude: some recent developments in relation to RLHF

- RLHF : Reinforcement Learning from Human Feedback
- Core technique in ChatGPT and relatives (OpenAI, Anthropic, …) for better aligning GPT-X with human preferences: e.g. Helpfulness, Harmlessness, Truthfulness, …

To conclude: some recent developments in relation to RLHF



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

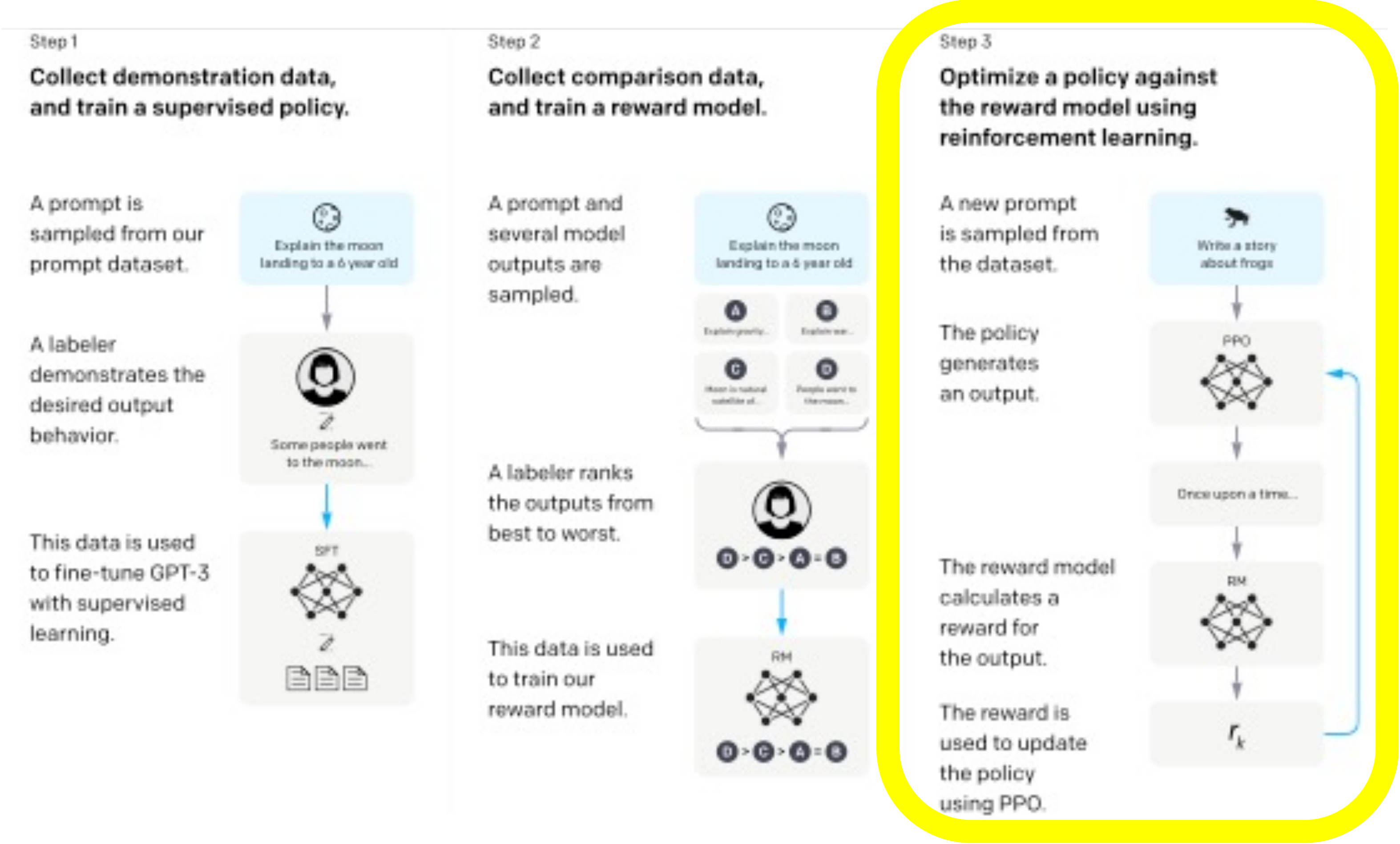
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

From Ouyang et al. *Training Language Models to Follow Instructions from Human Feedback.* Neurips 2022.

To conclude: some recent developments in relation to RLHF



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

Step 3 can be reinterpreted in terms of Distribution Matching

From Ouyang et al. *Training Language Models to Follow Instructions from Human Feedback.* Neurips 2022.

To conclude: some recent developments in relation to RLHF

- It can be shown [1] that ``step 3´´ of RLHF corresponds to a **distributional** match towards an **EBM** of the form

$$p(x) \propto a(x)\, e^{r(x)/\beta}$$

with $r(x)$ a reward, and β a positive coefficient

- The policy approximation is slightly different from DPG

DPG: $\mathbf{argmin}_{\boldsymbol{\theta}}\ \boldsymbol{KL}(\boldsymbol{p}, \boldsymbol{\pi_\theta})$ **RLHF:** $\mathbf{argmin}_{\boldsymbol{\theta}}\ \boldsymbol{KL}(\boldsymbol{\pi_\theta}, \boldsymbol{p})$

- In a recent paper [2], we show that these techniques can be subsumed under the large class of $\boldsymbol{f}$**-divergences** and study their behaviors as policy approximators

[1] Korbak, Elsahar, Kruszewski, Dymetman. *On Reinforcement Learning and Distribution Matching for Fine-Tuning Language Models with no Catastrophic Forgetting.* Neurips 2022.
[2] Go, Korbak, Kruszewski, Rozen, Ryu, Dymetman. *Aligning Language Models with Preferences through f-divergence Minimization.* arXiv 2023.

Tetiana Parshakova, Jean-Marc Andréoli, Hady Elsahar, Muhammad Khalifa, Germán Kruszewski, Jos Rozen, Brian Eikema, Chris Dance, Tomasz Korbak, Dongyoung Go, Nahyeon Ryu

Thank you!

Check out 🕺 disco, our distributional control library!

https://github.com/naver/disco

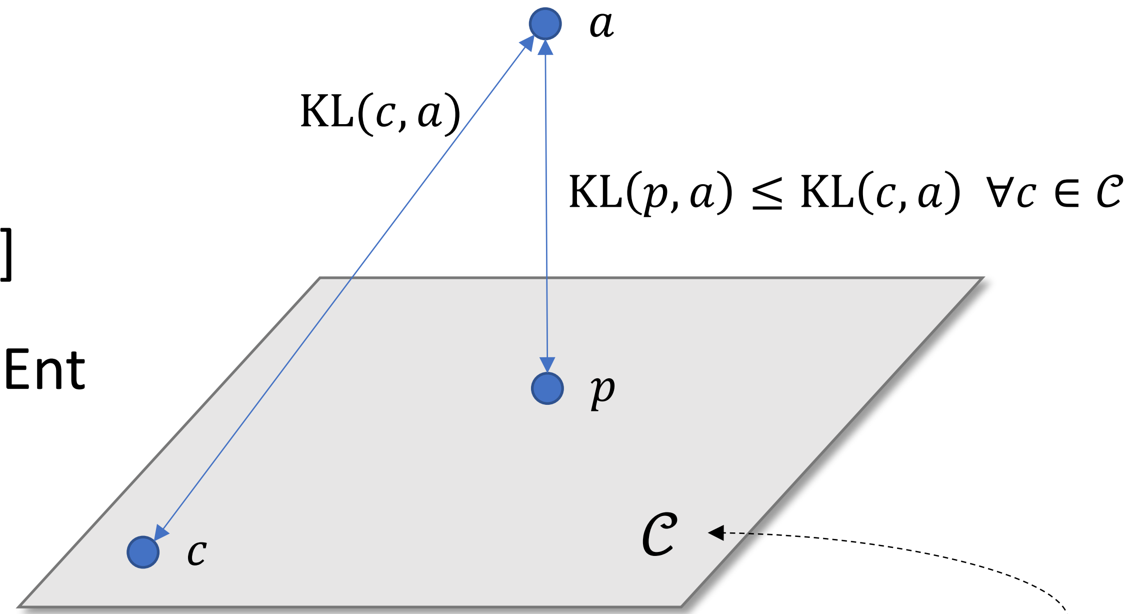
(Germán Kruszewski, Jos Rozen)

BACKUP

I-projections and Pythagore

- ***p* is called the *I-projection*** [2]

  A generalized version of MaxEnt


a
$\mathrm{KL}(c,a)$
$\mathrm{KL}(p,a) \leq \mathrm{KL}(c,a)\ \ \forall c \in \mathcal{C}$
p
c
$\mathcal{C}$

All distributions matching the moments constraints.

[2] Cszisar and Shields. *Information Theory and Statistics.* 2014

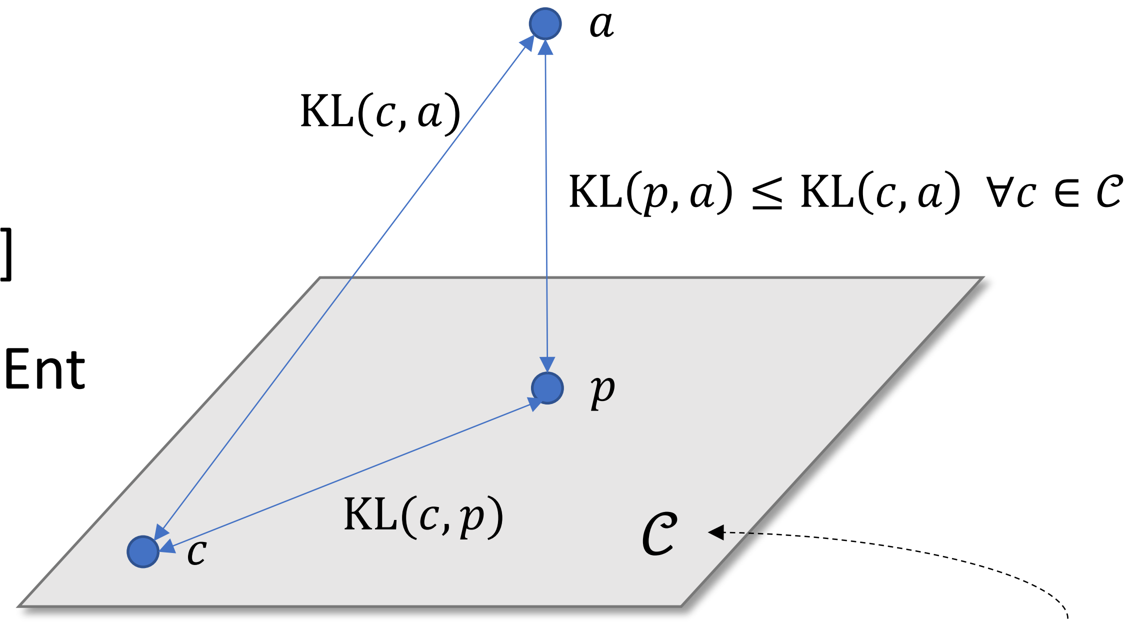
I-projections and Pythagore

- ***p* is called the *I-projection*** [2]

  A generalized version of MaxEnt

- **“Pythagorean” theorem:**

  $$\mathbf{KL}(\boldsymbol{c},\boldsymbol{a}) = \mathbf{KL}(\boldsymbol{p},\boldsymbol{a}) + \mathbf{KL}(\boldsymbol{c},\boldsymbol{p})$$


a
$\mathrm{KL}(c,a)$
$\mathrm{KL}(p,a) \leq \mathrm{KL}(c,a)\ \ \forall c \in \mathcal{C}$
p
$\mathrm{KL}(c,p)$
c
$\mathcal{C}$

All distributions matching the moments constraints.

[2] Cszisar and Shields. *Information Theory and Statistics.* 2014