Language Model Decoding as Direct Metrics Optimization

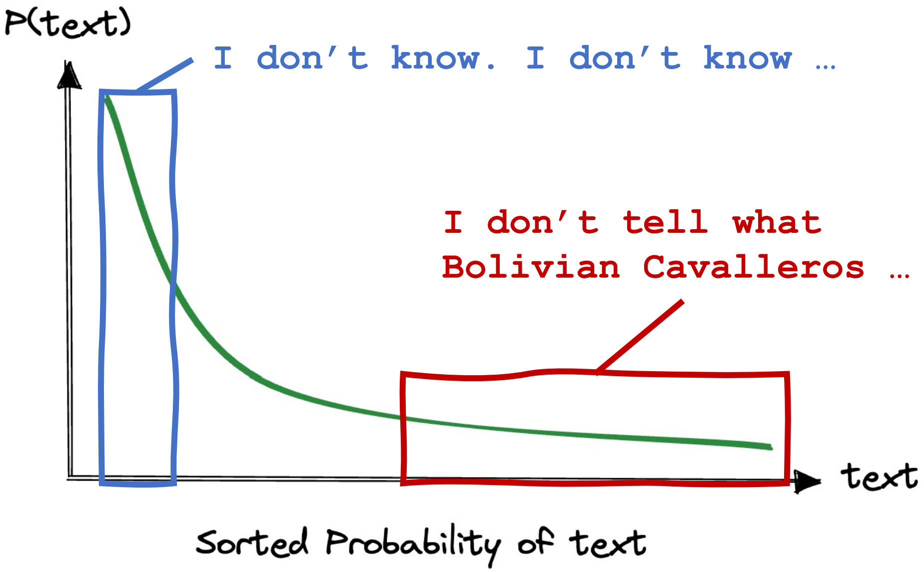
ICLR 2024

Haozhe Ji, Pei Ke, Hongning Wang, Minlie Huang

CoAI Group, Department of Computer Science, Tsinghua University

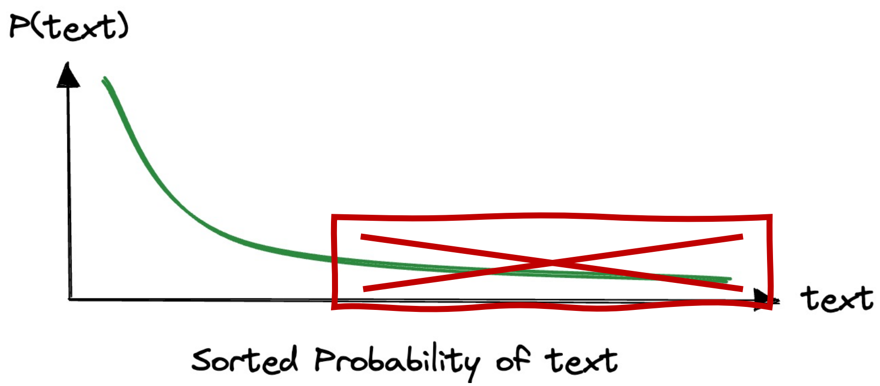
Background

- **Problem**: Decode from language models (LMs) to produce human-like texts.
- **Motivation**: Two mis-specifications of the LM's distribution:
- (i) The **unreliable** long tail [Holtzman et al., 2020]
  - The low-probability samples are **noisy**, **incoherent.**
- (ii) The **degenerated** mode [Welleck et al., 2020]
  - The highest probability samples are **repetitive** and exhibit **low diversity.**

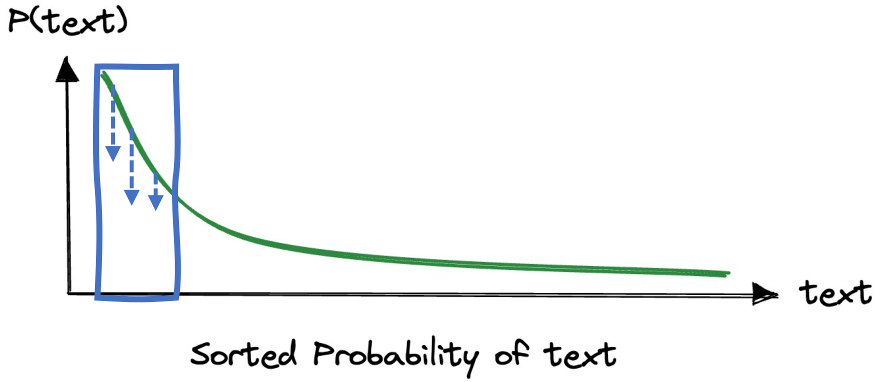


Background

- Existing solutions focus on "one end of the spectrum" with *ad-hoc* designs.
- (i) The **unreliable** long tail [Holtzman et al., 2020]
  - Sample from the **truncated** distribution with different criteria, e.g., top-k, top-p, typicality, etc.

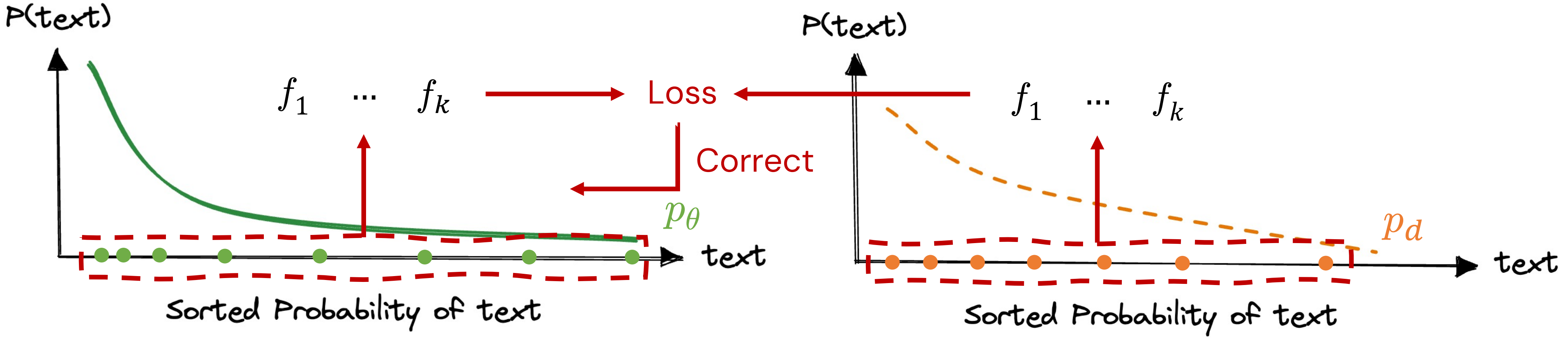


- (ii) The **degenerated** mode [Welleck et al., 2020]
  - Search with contrastive objective to **penalize** repetitive patterns, e.g., repetitive tokens, n-grams, embeddings.



Method

- **Our solution**: Correct the LM distribution by aligning with human distribution on **metrics** that reflect the mis-specifications, e.g., coherence, repetition, etc.
- **Input**:
  - (i) LM distribution p_θ (ii) *K* metric functions $f_k(\cdot)$ (iii) **samples** from human distribution p_d
- **Goal**:
  - Correct the LM distribution p_θ to align with human distribution p_d on the set of metrics $\{f_k\}_{k=1}^{K}$ with minimal deviation from p_θ.



Method

- **Formulation**:
  - Finding the **optimal decoding distribution** q_{opt} that solves the constrained optimization problem.

$$q_{\text{opt}} = \arg\min_{q \in \mathcal{P}} D_{\text{KL}}(q \| p_\theta)$$
$$s.t. \; \mathbb{E}_{\hat{\boldsymbol{x}} \sim q}[f_k(\hat{\boldsymbol{x}})] = \mathbb{E}_{\boldsymbol{x} \sim p_d}[f_k(\boldsymbol{x})], \quad k \in \{1, \cdots, K\},$$

  - Alignment on set of metrics $\{f_k\}_{k=1}^{K}$:
    - ***K*** constraints that match the expected metric scores on the generated texts with the human texts.
    - Sampling from q_{opt} produces texts that are human-like as evaluated by the metrics.
  - Minimal deviation from p_θ:
    - Minimize the reverse KL between q and p_θ to avoid over-optimization.
    - Reverse KL encourages q to seek the mode of p_θ while avoiding its long tail.

Method

- **Solving the optimization problem**:
  - The optimal decoding distribution q_{opt} has an analytic form defined as an energy-based model (EBM).

Proposition 1. *The distribution that solves the optimization problem (1) is in the form of:*

$$p_{\theta,\boldsymbol{\mu}}(\boldsymbol{x}) \propto p_\theta(\boldsymbol{x}) \exp\Big[-E_{\boldsymbol{\mu}}(\boldsymbol{x})\Big], \quad \forall \boldsymbol{x} \in S(p_{\theta,\boldsymbol{\mu}}) \tag{2}$$

where $E_{\boldsymbol{\mu}}(\boldsymbol{x}) = \boldsymbol{\mu}^\top \boldsymbol{f}(\boldsymbol{x})$ *and* $S(p) = \{\boldsymbol{x} : p(\boldsymbol{x}) > 0\}$ *is the support of distribution p.* $\boldsymbol{\mu} \in \mathbb{R}^K$ *is determined by the constraints in (1).*

  - The EBM is parametrized by the product of an auto-regressive LM p_θ and an exponential energy term $\exp[-\mu^\top f(x)]$.
  - Two remaining problems include:
    - Determining the coefficient $\mu = \{\mu_k\}_{k=1}^K$
    - Sampling form the EBM

Method

- **Theoretical guarantee of perplexity improvement**
  - The optimal decoding distribution q_{opt} improves the perplexity of the original LM distribution p_θ on human texts.

Proposition 2. *The optimal solution q_{opt} of the optimization problem (1) satisfies:*

1. $S(q_{\text{opt}}) \supseteq S(p_d)$, *where* $S(p) = \{\boldsymbol{x} : p(\boldsymbol{x}) > 0\}$.
2. $H(p_d, q_{\text{opt}}) = H(p_d, p_\theta) - D_{\text{KL}}(q_{\text{opt}} \| p_\theta)$, *where* $H(p, q) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log q(\boldsymbol{x})$.

  - **Statement 1** establishes the feasibility of computing the perplexity of q_{opt}
    - Existing heuristic decoding methods, e.g., truncation-based sampling and search methods are **infeasible** to calculate perplexity due to their sparse supports.
  - **Statement 2** reveals a non-negative perplexity (PPL) improvement of q_{opt} over p_θ

$$PPL(q_{opt}) = 2^{H(p_d, q_{\text{opt}})} < PPL(p_\theta) = 2^{H(p_d, p_\theta)}$$

    - As a distribution-level evaluation, the PPL improvement justifies that q_{opt} is generally a **better approximation** of the human distribution than p_θ.

Method

- **Determine the coefficient $\mu = \{\mu_k\}_{k=1}^{K}$**
  - Find μ that satisfies the K constraints

$$\mathbb{E}_{\hat{\boldsymbol{x}}\sim q}[f_k(\hat{\boldsymbol{x}})] = \mathbb{E}_{\boldsymbol{x}\sim p_d}[f_k(\boldsymbol{x})], \quad k \in \{1, \cdots, K\}$$

    - **1**. Estimate by weighted importance sampling (WIS)
    - **2**. Minimize the error between LHS and RHS

- **Sampling from the EBM**
  - A Sampling-importance-resampling (SIR) approach
    - **1**. Draw M samples from the LM p_θ given prefix
    - **2.** Calculate the importance weight $e^{-\mu^\top f}$
    - **3.** Resample from the empirical distribution
  - When M is finite, we empirically sample from p_θ with a temperature τ to increase convergence.

Algorithm 1 $\boldsymbol{\mu}_{\text{opt}}$ estimation with WIS

Input: p_θ, $\boldsymbol{F}$, learning rate α
Output: $\boldsymbol{\mu}_{\text{opt}}$
1: Initialize $\boldsymbol{\mu}$ randomly
2: Sample trajectories $\{\hat{\boldsymbol{x}}^i\}_{i=1}^N \sim p_\theta$
3: **repeat**
4: $\quad \hat{\boldsymbol{F}} \leftarrow \frac{\sum_{i=1}^N \exp(-E_{\boldsymbol{\mu}}(\hat{\boldsymbol{x}}^i))\boldsymbol{f}(\hat{\boldsymbol{x}}^i)}{\sum_{i=1}^N \exp(-E_{\boldsymbol{\mu}}(\hat{\boldsymbol{x}}^i))}$
5: $\quad \boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \alpha \nabla_{\boldsymbol{\mu}} \sqrt{\frac{1}{K}\|1 - \hat{\boldsymbol{F}}/\boldsymbol{F}\|_2^2}$
6: **until** convergence
7: $\boldsymbol{\mu}_{\text{opt}} \leftarrow \boldsymbol{\mu}$

Algorithm 2 Conditional Sampling with SIR

Input: p_θ, $E_{\boldsymbol{\mu}}$, prefix $\boldsymbol{x}_{\leq t_0}$, M, τ
Output: continuation $\boldsymbol{x}_{>t_0}$
1: **for** $i \leftarrow 1$ to M **do** $\quad \triangleright$ In parallel
2: $\quad$ Sample $\hat{\boldsymbol{x}}^i_{>t_0} \sim p^\tau_\theta(\cdot|\boldsymbol{x}_{\leq t_0})$
3: $\quad$ Compute $w_i \leftarrow \exp(-E_{\boldsymbol{\mu}}(\boldsymbol{x}_{\leq t_0}, \hat{\boldsymbol{x}}^i_{>t_0}))$
4: **end for**
5: Sample $j \sim$ Categorical$\left(\frac{w_1}{\sum_{i=1}^M w_i}, \cdots, \frac{w_M}{\sum_{i=1}^M w_i}\right)$
6: Set $\boldsymbol{x}_{>t_0} \leftarrow \hat{\boldsymbol{x}}^j_{>t_0}$

Experiments

- **Datasets:** Wikipedia (Wikitext-103), News (Wikinews)
- **Models:** GPT-2 XL (1.5B), OPT-6.7B
- **Metrics:**
  - **Repetition** [Welleck et al., 2020]**:** seq-rep-*n* (*n*=2,3,4), tok-rep-*l* (*l*=8,12,32)
  - **Coherence** [Su et al., 2022]: Cosine similarity between embeddings of $x_{\leq t_0}$ and $x_{>t_0}$
  - **Diversity** [Li et al., 2022]: Aggregated n-gram diversity
  - **Information** [Braverman et al., 2022]**:** **E**xponential of entropy rate evaluated by an LM
  - **MAUVE** [Pillutla et al., 2021]**:** Distributional similarity between two sets of texts

Experiments

◉ Main results:

		Method	Wikipedia SR-4	Wikipedia TR-32	Wikipedia COH	Wikipedia DIV	Wikipedia e^{ENT}	Wikipedia MAU	News SR-4	News TR-32	News COH	News DIV	News e^{ENT}	News MAU	Δ_{ref}
		Reference	0.48	21.3	62.3	92.5	23.2	-	0.29	18.7	66.6	94.1	13.8	-	
	GPT-2 XL	Greedy	60.9	65.5	60.2	8.03	2.29	59.7	53.2	58.2	63.8	13.2	2.19	65.2	39.8
Sampling	GPT-2 XL	Top-k	2.11	23.4	60.9	87.8	10.1	77.8	0.95	20.3	64.7	91.7	8.17	96.3	3.6
Sampling	GPT-2 XL	Nucleus	1.19	20.0	57.3	**92.4**	17.3	78.3	0.80	18.7	60.8	93.5	11.0	95.3	2.3
Sampling	GPT-2 XL	Typical	0.81	17.4	54.9	94.5	30.1	78.7	0.42	16.9	57.2	95.3	18.2	95.0	3.9
Search	GPT-2 XL	CD	1.31	28.2	68.7	85.9	7.55	77.8	0.63	23.2	71.2	90.5	6.55	95.1	5.8
Search	GPT-2 XL	CS	1.78	23.0	56.9	90.6	5.25	83.3	0.77	19.2	63.6	**94.1**	4.18	95.7	4.2
	GPT-2 XL	DAEMON	**0.42**	**22.5**	**62.5**	92.2	**22.8**	**88.1**	**0.18**	**18.7**	**66.3**	94.5	**13.7**	**97.4**	0.3

◆ Generally, sampling methods are worse in coherence, search methods are worse in diversity and repetition.

◆ Our method (Daemon) achieves the lowest Δ_{ref} averaged on all metrics and attains the highest MAUVE score.

Experiments

Other results:

- Perplexity evaluation

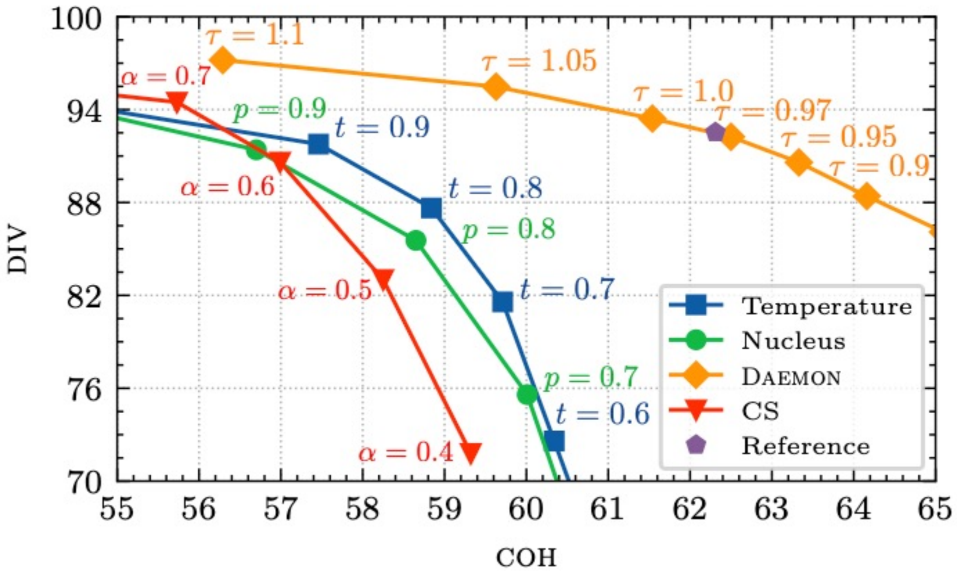
Model	Wikipedia ori	Wikipedia imp	News ori	News imp
GPT-2 XL	23.1	**22.0**	13.9	**13.1**
OPT-6.7B	16.4	**16.2**	10.8	**10.2**

Consistent perplexity improvement across models and datasets

- Human evaluation

Ours vs.	Fluency Win	Fluency Lose	Coherence Win	Coherence Lose	Informativeness Win	Informativeness Lose
CD	**0.54**	0.35	**0.48***	0.36	**0.48***	0.27
CS	**0.53***	0.34	**0.47***	0.29	**0.41**	0.33
Nucleus	**0.54***	0.33	**0.66***	0.15	**0.45***	0.30
Typical	**0.53***	0.30	**0.62***	0.19	**0.44***	0.23

- Evaluating the coherence-diversity trade-off



- Tuning temperature yields a better frontier of coherence and diversity that dominates the baseline methods.

Conclusion

- We propose to frame decoding from LM as an optimization problem, which finds the optimal decoding distribution that align with human distribution on multiple metrics.
- We prove the optimal decoding distribution is guaranteed to improve the perplexity of the original LM, indicating a general improvement of approximating the human distribution.
- Finally, our extensive empirical results demonstrate that our method achieves better performance of alignment with human texts on multiple metrics, and superior quality-diversity trade-off.