SPG: Sandwiched Policy Gradient for Mask Diffusion Language Models

Chenyu Wang, Paria Rashidinejad, Andy Su, Song Jiang, Sid Wang, Siyan Zhao, Cai Zhou, Shannon Shen, Feiyu Chen, Tommi Jaakkola, Yuandong Tian, Bo Liu

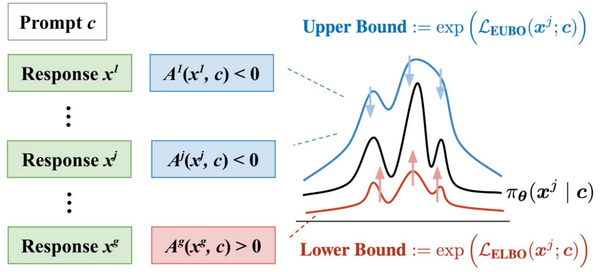
Meta · Massachusetts Institute of Technology · UCLA · USC · ICLR

Paper Code

SPG: Sandwiched Policy Gradient with Evidence Bounds

A principal challenge in RL for diffusion large language models (dLLMs) is the **computationally intractable log-likelihood**, which is essential for accurate policy gradient estimation. Prior work applies the evidence lower bound (ELBO) or a one-step estimation as a surrogate for true likelihood, while leads to misaligned gradient.

We propose SPG to reduce bias by optimizing **sandwiched variational bounds** based on reward.



ELBO: $\mathcal{L}_{\text{ELBO}}(\boldsymbol{x};\boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{z}_t}\left[\sum_{i=1}^{n} w(t)\cdot \mathbb{1}(\boldsymbol{z}_{t,i}=\boldsymbol{m})\cdot \log \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i \mid \boldsymbol{z}_t)\right]$

EUBO: ?

A Tractable Evidence Upper Bound

Theorem 1 (Evidence Upper Bound for Masked Diffusion). *Assume the forward denoising process has T steps with a monotonic schedule α_t. For any $\beta \geq 1$ and a sequence $\boldsymbol{x}_{1:n}$, we have:*

$$\mathcal{L}_{EUBO}(\boldsymbol{x}_{1:n};\boldsymbol{\theta}) = \frac{1}{\beta}\sum_{i=1}^{n}\log\sum_{t=1}^{T-1}\mathbb{E}_{\boldsymbol{z}_{t+1}}\left[\frac{\alpha_t-\alpha_{t+1}}{1-\alpha_{t+1}}\cdot\mathbb{1}(\boldsymbol{z}_{t+1,i}=\boldsymbol{m})\cdot\pi_{\boldsymbol{\theta}}^{\beta}(\boldsymbol{x}_i \mid \boldsymbol{z}_{t+1})\right] + C(T), \quad (6)$$

where $C(T) := \mathbb{1}(\beta<n)\cdot\frac{1}{\beta}\log\mathbb{E}_{\boldsymbol{z}_{1:T}\sim q(\cdot|\boldsymbol{x})}\left[q(\boldsymbol{z}_{1:T} \mid \boldsymbol{x})^{-n}\right]$ *is a constant independent of* $\boldsymbol{\theta}$.

Corollary 1. *Taking the limit of $T \to \infty$, we have:*

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{EUBO}(\boldsymbol{x}_{1:n};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\left(\tilde{\mathcal{L}}_{EUBO}(\boldsymbol{x}_{1:n};\boldsymbol{\theta}) + C(T)\right) = \nabla_{\boldsymbol{\theta}}\tilde{\mathcal{L}}_{EUBO}(\boldsymbol{x}_{1:n};\boldsymbol{\theta}), \quad \textit{where}$$

$$\tilde{\mathcal{L}}_{EUBO}(\boldsymbol{x}_{1:n};\boldsymbol{\theta}) = \frac{1}{\beta}\sum_{i=1}^{n}\log\mathbb{E}_{t,\boldsymbol{z}_t}\left[w(t)\cdot\mathbb{1}(\boldsymbol{z}_{t,i}=\boldsymbol{m})\cdot\pi_{\boldsymbol{\theta}}^{\beta}(\boldsymbol{x}_i \mid \boldsymbol{z}_t)\right].$$

Algorithm 1 SPG: Sandwiched Policy Gradient for Masked dLLMs

Require: prompt distribution $\mathcal{D}$, number of completions per prompt g, number of inner updates μ, forward process q, number of Monte Carlo samples m, initial policy π_0, learning rate ϵ.

Initialize $\pi_\theta \leftarrow \pi_0$
while not converged **do**
  Sample a prompt $\boldsymbol{c} \sim \mathcal{D}$, then g completions $\{\boldsymbol{x}^j \sim \pi_\theta(\cdot \mid \boldsymbol{c})\}_{j=1}^{g}$
  $\forall j \in [g]$, compute reward $R(\boldsymbol{c}, \boldsymbol{x}^j)$ and advantage $A^j(\boldsymbol{x}^j, \boldsymbol{c})$
  for gradient update iterations $\{1, \ldots, \mu\}$ **do**
    $\forall j \in [g]$, generate m perturbed samples $\{\boldsymbol{z}_{t_\tau}^j\}_{\tau=1}^{m} \sim q(\cdot \mid \boldsymbol{x}^j)$
    Compute the sandwiched policy gradient $\nabla\mathcal{J}_{\text{SPG}}(\boldsymbol{\theta})$ where:

$$\mathcal{J}_{\text{SPG}}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{1}{g}\sum_{j=1}^{g}\left(\mathbb{1}_{A^j\geq 0}\cdot A^j\mathcal{L}_{\text{ELBO}}(\boldsymbol{x}^j \mid \boldsymbol{c};\boldsymbol{\theta}) + \mathbb{1}_{A^j<0}\cdot A^j\tilde{\mathcal{L}}_{\text{EUBO}}(\boldsymbol{x}^j \mid \boldsymbol{c};\boldsymbol{\theta})\right)\right],$$

    and $\mathcal{L}_{\text{ELBO}}$, $\tilde{\mathcal{L}}_{\text{EUBO}}$ are estimated from $\{\boldsymbol{z}_{t_\tau}^j\}_{\tau=1}^{m}$, using Equation 2 and 7.
    Perform gradient update: $\theta \leftarrow \boldsymbol{\theta} + \epsilon\nabla\mathcal{J}_{\text{SPG}}(\boldsymbol{\theta})$
return π_θ

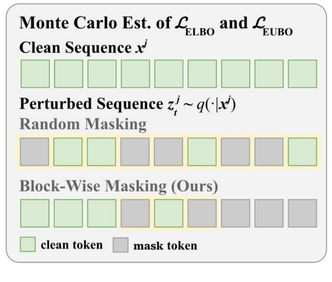
We apply two practical tricks to stablize the training:
1) Mixture of upper and lower bound for negative advantage traces and
2) Block-wise masking strategy for Monte Carlo estimation.

Mixture of Upper and Lower Bound for Negative Advantage Traces

Mixture of lower and upper bound as an estimation leads to
1) confidence-aware weighting and **2)** lower variance and more stable training.

$$\tilde{\mathcal{L}}_{\text{Mix}}(\boldsymbol{x} \mid \boldsymbol{c};\boldsymbol{\theta}) := \omega\cdot\tilde{\mathcal{L}}_{\text{EUBO}}(\boldsymbol{x} \mid \boldsymbol{c};\boldsymbol{\theta}) + (1-\omega)\cdot\mathcal{L}_{\text{ELBO}}(\boldsymbol{x} \mid \boldsymbol{c};\boldsymbol{\theta})$$

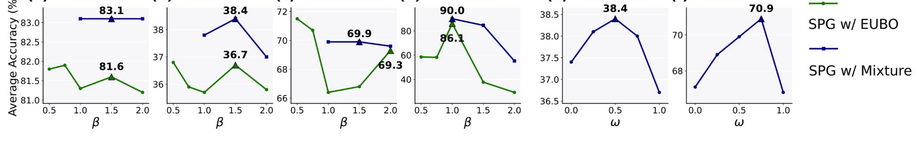
Block-Wise Masking Strategy for Monte Carlo Estimation



The commonly applied state-of-the-art inference strategy for dLLMs is **block-wise semi-autoregressive** unmasking, which is used during policy rollout.

To better align data distributions between policy rollout and optimization, we adopt a **block-wise masking strategy** instead of random masking for Monte Carlo estimation of ELBO and EUBO.

Ablations on Key Hyperparameters



Results on Four Reasoning Benchmarks

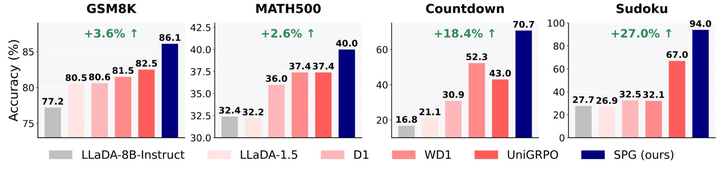


Table 1 Model performance on four reasoning benchmarks. The best results are bolded and the second best are underlined. SPG consistently outperforms all other methods. We denote the absolute gain of test accuracy to the previous state-of-the-art in green.

	GSM8K (0-shot)			MATH500 (0-shot)			Countdown (0-shot)			Sudoku (3-shot)		
Model / Seq Len	**128**	**256**	**512**	**128**	**256**	**512**	**128**	**256**	**512**	**128**	**256**	**512**
LLaDA-8B-Inst.	69.5	77.2	79.8	28.2	32.4	34.6	18.8	16.8	16.8	5.7	27.7	26.2
LLaDA-1.5	70.4	80.5	81.9	26.8	32.2	35.8	21.9	21.1	21.5	7.4	26.9	29.0
D1	72.2	80.6	81.3	31.4	36.0	39.4	30.9	30.9	34.4	7.2	32.5	29.3
WD1	74.6	81.5	83.0	31.0	37.4	39.0	48.8	52.3	50.8	33.1	32.1	22.5
UniGRPO	74.9	82.5	82.7	32.4	37.4	39.4	44.5	43.0	57.0	59.0	67.0	62.9
SPG w/ EUBO	77.1	83.8	83.9	33.2	37.6	39.4	68.4	**71.5**	68.0	81.2	87.1	89.9
SPG w/ Mixture	**78.5**+3.6	**86.1**+3.6	**84.5**+1.5	**33.4**+1.0	**40.0**+2.6	**41.8**+2.4	**68.8**+20	70.7+18	**70.3**+13	**82.9**+24	**94.0**+27	**93.1**+30

Ablations on Algorithm Components

1) Log-likelihood estimation methods for negative advantage traces.

Model	GSM8K	MATH500	Countdown	Sudoku
SPG wo/ neg	77.4	32.7	45.5	68.8
SPG w/ ELBO	80.9	37.4	67.1	82.4
SPG w/ EUBO	81.6	36.7	69.3	86.1
SPG w/ Mixture	**83.1**+2.2	**38.4**+1.0	**69.9**+2.8	**90.0**+7.6

2) Masking strategies in Monte Carlo estimation.

Model	Masking	MATH500	Countdown
SPG w/ EUBO	random	36.7	45.4
	block-wise	36.7+0.0	69.3+23.9
SPG w/ Mixture	random	36.9	62.8
	block-wise	38.4+1.5	69.9+7.1

Results on Coding

	HumanEval (0-shot)		
Model / Seq Len	**128**	**256**	**512**
LLaDA-8B-Inst.	28.1	35.4	34.8
LLaDA-1.5	22.6	34.8	32.3
D1	29.3	37.8	37.2
WD1	25.6	39.0	38.4
UniGRPO	28.1	39.6	38.4
SPG w/ EUBO	29.3	**41.5**+1.9	33.5
SPG w/ Mixture	**31.1**+1.8	39.0	**39.6**+1.2

	MBPP (0-shot)		
Model / Seq Len	**128**	**256**	**512**
LLaDA-8B-Inst.	36.2	42.0	38.1
LLaDA-1.5	33.5	44.8	40.1
D1	42.0	45.5	43.2
WD1	38.9	45.9	44.8
UniGRPO	39.3	45.5	45.9
SPG w/ EUBO	43.6	46.7	46.3
SPG w/ Mixture	**46.3**+4.3	**50.6**+4.7	**49.0**+3.1

Ablations on Inference Strategies