Vanishing Gradients in Reinforcement Finetuning of Language Models

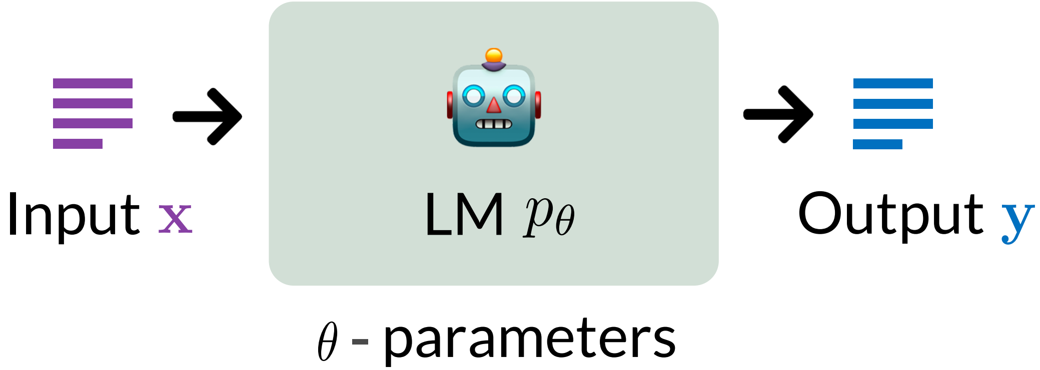
Noam Razin

Joint work w/ Hattie Zhou, Omid Saremi, Vimal Thilak, Arwen Bradley,
Preetum Nakkiran, Joshua Susskind, Etai Littwin

ICLR 2024

Language Models (LMs)

Language Model (LM): Neural network trained on large amounts of text data to produce a **distribution over text**



Supervised Finetuning of LMs

LMs are adapted to human preferences and downstream tasks via **finetuning**

Supervised Finetuning (SFT)

Minimize cross entropy loss over labeled inputs via **gradient-based methods**

Limitations:

- Hard to formalize human preferences through labels
- Labeled data is expensive

Reinforcement Finetuning of LMs

Limitations of SFT led to wide adoption of a **reinforcement learning**-based approach

(e.g. Ziegler et al. 2019, Stiennon et al. 2020, Ouyang et al. 2022, Bai et al. 2022, Dubois et al. 2023, Touvron et al. 2023)

Reinforcement Finetuning (RFT)

Maximize reward over unlabeled inputs via **policy gradient algorithms**

, , $\cdots$, reward function $r(\mathbf{x}, \mathbf{y})$

Expected reward for input $\mathbf{x}$: $V_\theta(\mathbf{x}) = \mathbb{E}_{\mathbf{y} \sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x}, \mathbf{y})]$

Reward function $r(\mathbf{x}, \mathbf{y})$ can be:

- Learned from human preferences
- Tailored to a downstream task

Our Work: Vanishing Gradients Due to Small Reward STD

$\mathrm{STD}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]$ — reward std of $\mathbf{x}$ under the model

Theorem

$$\|\nabla_\theta V_\theta(\mathbf{x})\| = O\left(\mathrm{STD}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})]^{2/3}\right)$$

Expected gradient for an input vanishes when reward std is small, even if reward mean is suboptimal

*Same holds for PPO gradient

Proof Idea: Stems from use of softmax + reward maximization objective

Main Contributions: Vanishing Gradients in RFT

$\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$

Expected gradient for an input vanishes in RFT if the input's reward std is small

Experiments + theory: **vanishing gradients in RFT are prevalent and detrimental** to maximizing reward

Exploration of possible solutions: **Initial SFT phase** allows overcoming vanishing gradients in RFT, and **does not need to be expensive**

Reward std is a key quantity to track for successful RFT

Main Contributions: Vanishing Gradients in RFT

Thank You!

$\nabla_\theta \mathbf{V}_\theta(\mathbf{x}) \approx \mathbf{0}$

Expected gradient for an input vanishes in RFT if the input's reward std is small

Experiments + theory: **vanishing gradients in RFT are prevalent and detrimental** to maximizing reward

Exploration of possible solutions: **Initial SFT phase** allows overcoming vanishing gradients in RFT, and **does not need to be expensive**

Reward std is a key quantity to track for successful RFT