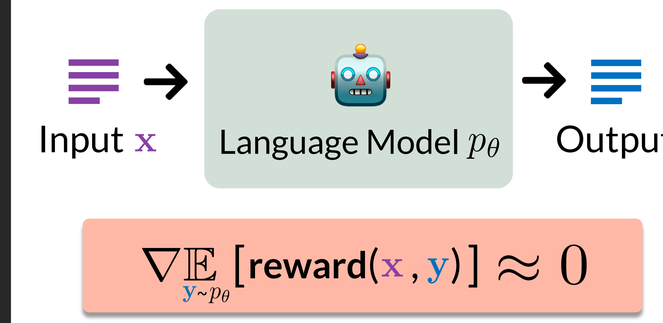
Finetuning language models via reinforcement learning **suffers from vanishing gradients**



Vanishing Gradients in Reinforcement Finetuning of Language Models

Noam Razin, Hattie Zhou, Omid Saremi, Vimal Thilak, Arwen Bradley, Preetum Nakkiran, Joshua Susskind, Etai Littwin

Apple, Tel Aviv University, Mila, Universite de Montreal

Terminology

Language Model: Neural network trained to produce a distribution over text

Supervised Finetuning (SFT): Minimize cross entropy loss over labeled examples

Reinforcement Finetuning (RFT): Maximize reward via policy gradient

1 Theory: Vanishing Gradients

$\nabla \approx 0$

Gradient of expected reward for an input vanishes if the input's reward standard deviation is small

2 Experiments

Vanishing gradients in RFT are prevalent and detrimental to maximizing reward

3 Possible Solutions

Initial SFT phase can mitigate vanishing gradients in RFT, and does not need to be expensive

Takeaway

Reward standard deviation of individual inputs is a key quantity to track for successful finetuning

1 $\mathrm{STD}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[\text{reward}(\mathbf{x},\mathbf{y})]$ — reward std of $\mathbf{x}$ under the model

Theorem

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

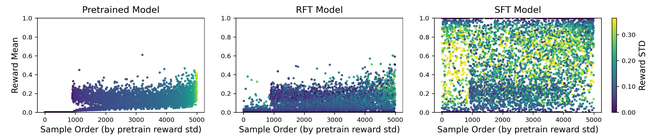
Expected gradient for an input vanishes when reward std is small, even if reward is suboptimal

2 **Benchmark:** GRUE (Ramamurthy et al. 2023)

Findings:

- Vanishing gradients are prevalent: 3 of 7 datasets contain considerable number of train inputs with small reward std
- RFT has limited impact on reward of inputs with small reward std

NarrativeQA **(many inputs with small reward std and low reward)**



IMDB **(few inputs with small reward std and low reward)**

3 **Do Common Heuristics Help?** ✗

Increasing learning rate, temperature, entropy regularization

Observation: Initial SFT phase (commonly used in practice) reduces number of inputs with small reward std

Importance of SFT in RFT pipeline: It helps mitigate vanishing gradients

Implication: Efficient Initial SFT Phase

If SFT phase helps due to mitigating vanishing gradients for RFT

➡ A few SFT steps on small number of labeled samples may suffice

1% of labeled samples for SFT lead to roughly same reward as "full" SFT