Gradient dissent in language model training and saturation

Andrei Mircea, Ekaterina Lobacheva, Irina Rish Université de Montréal // Mila

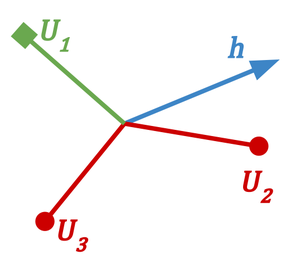
Overview

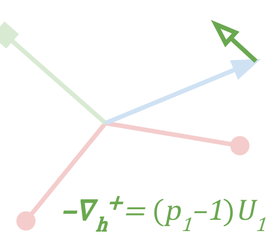
Goal: characterize learning dynamics of saturation to better understand and mitigate it.
Challenge: no shared basis in which to compare dynamics of different models.
Approach: create interpretable shared basis for studying dynamics, using **attractive**/**repulsive** components of per-sample gradients.
Findings: *gradient dissent*, where attractive/repulsive components become systematically opposed with saturation.

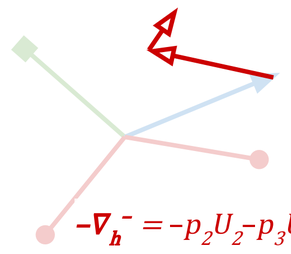
1. Output layer gradient decomposition

LM output layers lead to gradients which decompose into attractive (∇^+) and repulsive (∇^-) components in activations $\boldsymbol{h}$ and model parameters $\boldsymbol{\theta}$. Attractive/repulsive grads increase/decrease true/false logits respectively.

$\boldsymbol{h}$: hidden state
$\boldsymbol{U_1}$: true class vector
$\boldsymbol{U_2}$: false class vector
$\boldsymbol{U_3}$: false class vector
$\boldsymbol{\ell}$: output logit vector
$\boldsymbol{p}$: probability vector
$\mathcal{L}$: cross-entropy loss







$$\boldsymbol{\ell} = \begin{bmatrix} h_1 & h_2 \end{bmatrix} \begin{bmatrix} U_{1,1} & U_{2,1} & U_{3,1} \\ U_{1,2} & U_{2,2} & U_{3,2} \end{bmatrix} = \begin{bmatrix} \ell_1 & \ell_2 & \ell_3 \end{bmatrix}$$

$$\boldsymbol{p} = softmax(\begin{bmatrix} \ell_1 & \ell_2 & \ell_3 \end{bmatrix}) = \begin{bmatrix} p_1 & p_2 & p_3 \end{bmatrix}$$

$$\mathcal{L} = -log(\begin{bmatrix} p_1 \end{bmatrix})$$

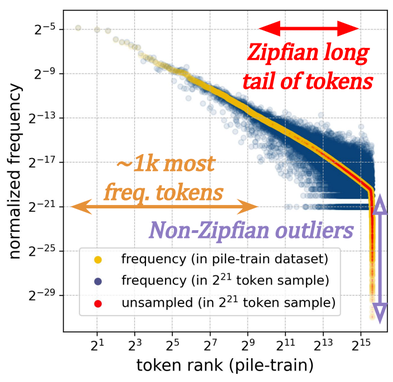
$$d\mathcal{L}/d\boldsymbol{\ell} = \nabla_\ell^+ + \nabla_\ell^- = \begin{bmatrix} 1-p_1 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & p_2 & p_3 \end{bmatrix}$$

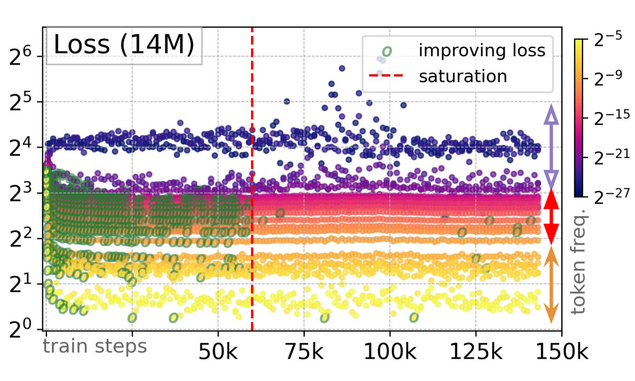
$$d\mathcal{L}/d\boldsymbol{h} = \nabla_h^+ + \nabla_h^- = \begin{bmatrix} 1-p_1 & U_{1,1} \\ & U_{2,1} \end{bmatrix} + \begin{bmatrix} p_2 & U_{1,2} \\ & U_{2,2} \end{bmatrix} + \begin{bmatrix} p_3 & U_{1,3} \\ & U_{2,3} \end{bmatrix}$$

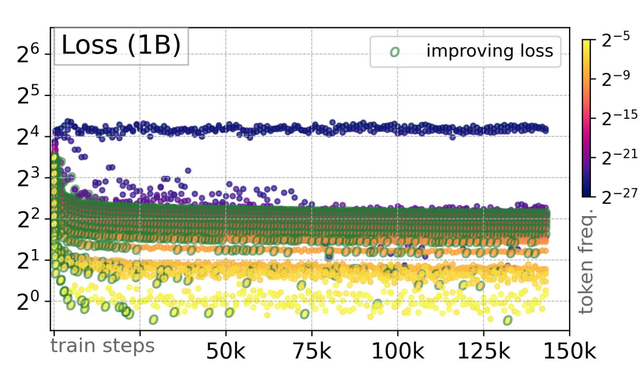
$$d\mathcal{L}/d\boldsymbol{\theta} = \nabla_\theta^+ + \nabla_\theta^- = d\boldsymbol{h}/d\boldsymbol{\theta}\,\nabla_h^+ + d\boldsymbol{h}/d\boldsymbol{\theta}\,\nabla_h^-$$

2. Characterizing saturation across model sizes and token frequecies

- Saturation is a sharp transition in models below a certain size.
- **Frequent tokens** saturate rapidly.
- Learning *and* saturation occur primarily in the **Zipfian long tail**.
- **Non-Zipfian outliers** behave in qualitatively different manner.







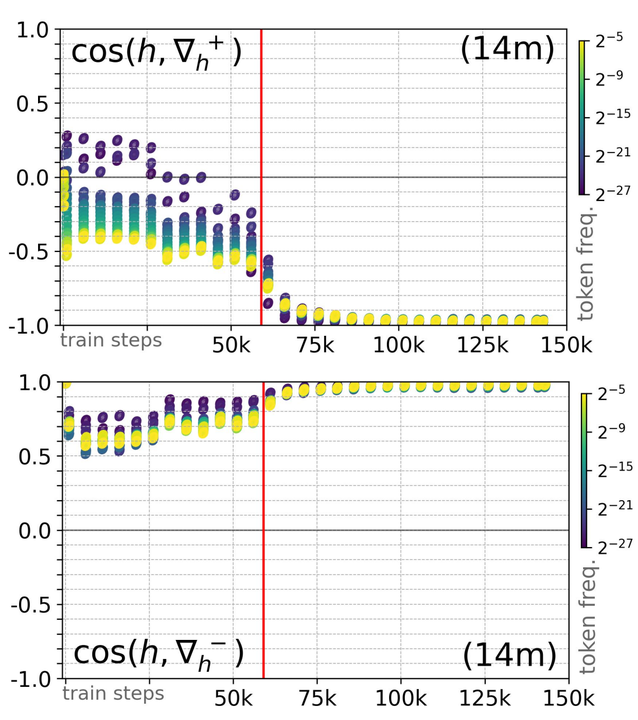
3. Comparing learning dynamics in shared latent basis across tokens and models

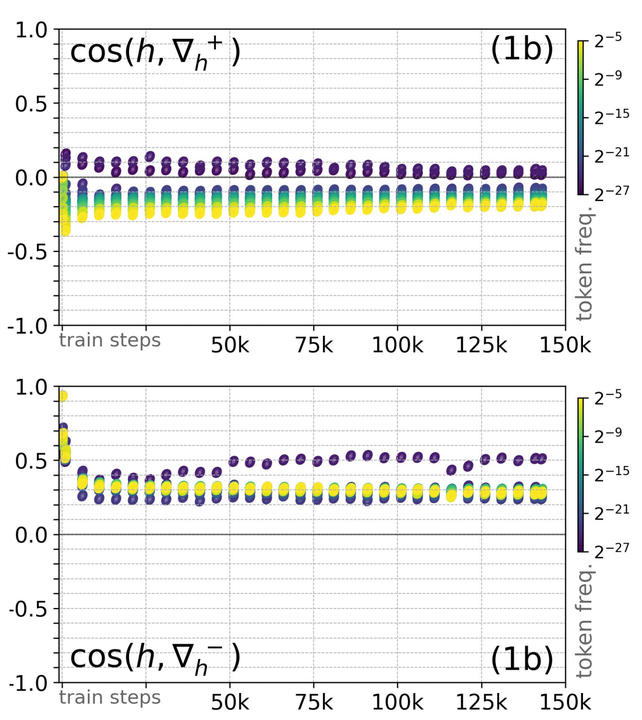
Creating a shared basis for learning dynamics:

- Project normalized $\boldsymbol{h}$ onto normalized ∇_h^+ and ∇_h^-
- Resulting $\cos(\boldsymbol{h}, \nabla_h^+)$ and $\cos(\boldsymbol{h}, \nabla_h^-)$ become shared 2D basis in which to compare learning dynamics across token frequencies and across model sizes.
- Intuitively corresponds to angular alignment of the hidden state vectors with each gradient component.

Observations and gradient dissent hypothesis:

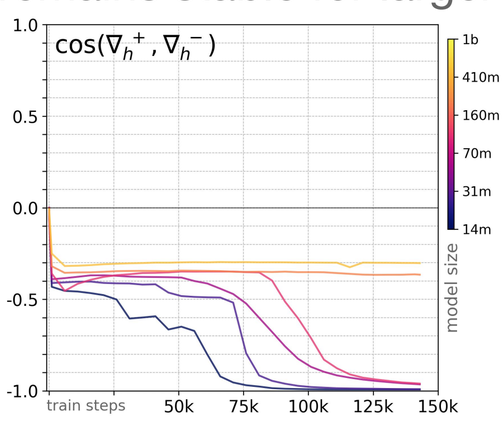
- Saturation transition co-occurs with collapse in dynamics as $\cos(\boldsymbol{h}, \nabla_h^+) = -1$ and $\cos(\boldsymbol{h}, \nabla_h^-) = 1$.
- Gradient dissent: collapse suggests ∇_h^+ and ∇_h^- become totally opposed and interfere destructively, starving gradients in remaining model parameters $\boldsymbol{\theta}$.

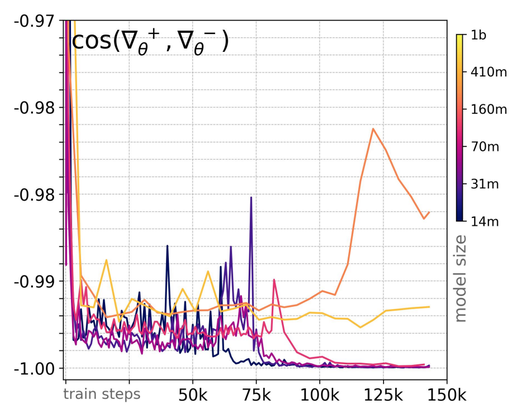




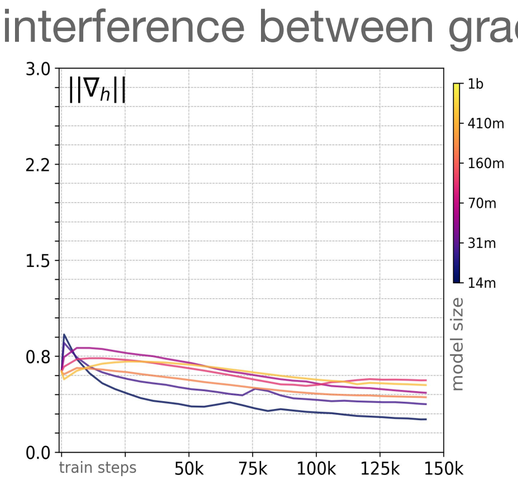
4. Evidence of dissent in saturation

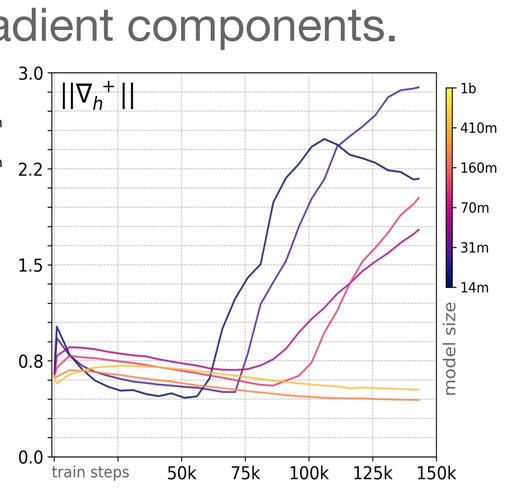
$\cos(\nabla^+, \nabla^-) \to -1$ as small models saturate, but remains stable for larger unsaturated models





During saturation, $||\nabla^+||$ and $||\nabla^-||$ explode as $||\nabla||$ remains stable, indicating destructive interference between gradient components.





5. Summary of findings and key takeaways

- Language model saturation is a sharp transition concentrated in the Zipfian long tail of tokens.
- To characterize and compare learning dynamics across models, samples, parameters, activations, etc. a shared and interpretable basis can be created by linearly decomposing the gradient.
- Gradient dissent is a phenomenon which arises as attractive/repulsive components of the output layer gradient become systematically opposed
- Gradient dissent transitions are strongly associated with model saturation transitions.

6. Open questions and future work

- What is the role of dissent in gradient saturation?
- Is saturation due to capacity or training dynamics?
- Do output layers bottleneck learning dynamics?
- What is the effect of Zipfian long-tail and outlier tokens on learning dynamics and saturation?