

On the Implicit Bias of Adam

Matias D. Cattaneo Jason M. Klusowski
Boris Shigida*

Princeton University

ICML 2024

*bs1624@princeton.edu

Background: modified ODE for gradient descent

- GD is the Euler method solving an ODE:

$$\underbrace{\dot{\boldsymbol{\theta}} = -\nabla \underbrace{E}_{\text{loss}}(\boldsymbol{\theta})}_{\text{ODE}} \xrightarrow{\text{discretization}} \underbrace{\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - \underbrace{h}_{\text{step size}} \nabla E(\boldsymbol{\theta}^{(n)})}_{\text{gradient descent}},$$

(Note: In the original image, an arrow points from the text "iteration number" to the superscript $(n+1)$ in the second equation.)

giving $O(h)$ -closeness: $\|\boldsymbol{\theta}(nh) - \boldsymbol{\theta}^{(n)}\| = O(h)$

- There is an ODE that is closer to the iterations:

$$\dot{\tilde{\boldsymbol{\theta}}} = -\nabla E(\tilde{\boldsymbol{\theta}}) - \underbrace{\frac{h}{4} \nabla \|\nabla E(\tilde{\boldsymbol{\theta}})\|^2}_{\text{implicit regularization}} \xrightarrow{\text{discretization}} \boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - h \nabla E(\boldsymbol{\theta}^{(n)}),$$

giving $O(h^2)$ -closeness: $\|\tilde{\boldsymbol{\theta}}(nh) - \boldsymbol{\theta}^{(n)}\| = O(h^2)$

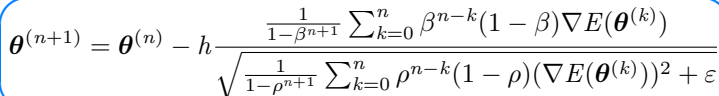
- See, e. g.,
D. Barrett and B. Dherin (2021). “Implicit Gradient Regularization”. In: International Conference on Learning Representations

Background: $O(h)$ approximation of Adam

- Full-batch Adam with fixed β, ρ, ε is close to perturbed sign-GD flow:

$$\dot{\boldsymbol{\theta}} = -\frac{\nabla E(\boldsymbol{\theta})}{\sqrt{|\nabla E(\boldsymbol{\theta})|^2 + \varepsilon}}$$

β, ρ momentum hyperparameters
 ε numerical stability hyperparameter


$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - h \frac{\frac{1}{1-\beta^{n+1}} \sum_{k=0}^n \beta^{n-k} (1-\beta) \nabla E(\boldsymbol{\theta}^{(k)})}{\sqrt{\frac{1}{1-\rho^{n+1}} \sum_{k=0}^n \rho^{n-k} (1-\rho) (\nabla E(\boldsymbol{\theta}^{(k)}))^2 + \varepsilon}}$$

full-batch Adam

- C. Ma, L. Wu, and E. Weinan (2022). “A qualitative study of the dynamic behavior for adaptive gradient algorithms”. In: Mathematical and Scientific Machine Learning. PMLR, pp. 671–692

Our contribution: $O(h^2)$ approximation of Adam

- There is a correction term

$$\dot{\tilde{\theta}}(t) = -\frac{\nabla E(\tilde{\theta}(t)) + \text{correction}(\tilde{\theta}(t))}{\sqrt{|\nabla E(\tilde{\theta}(t))|^2 + \varepsilon}},$$

giving $O(h^2)$ -closeness to Adam

- It is given by

$$\text{correction}_j(\theta) := \frac{h}{2} \left\{ \frac{1 + \beta}{1 - \beta} - \frac{1 + \rho}{1 - \rho} + \frac{1 + \rho}{1 - \rho} \cdot \frac{\varepsilon}{|\nabla_j E(\theta)|^2 + \varepsilon} \right\} \nabla_j \|\nabla E(\theta)\|_{1,\varepsilon}$$

- The index j means “ j -th component”
- The *perturbed one-norm* is $\|\mathbf{v}\|_{1,\varepsilon} = \sum_i \sqrt{v_i^2 + \varepsilon}$

Correction term in the typical case

- When ε is small (compared to gradient components),

$$\text{correction}(\boldsymbol{\theta}) \approx \underbrace{\frac{h}{2} \left\{ \frac{1+\beta}{1-\beta} - \frac{1+\rho}{1-\rho} \right\} \nabla \|\nabla E(\boldsymbol{\theta})\|_{1,\varepsilon}}_{\text{implicit anti-regularization}}$$

- The perturbed one-norm is $\|\boldsymbol{v}\|_{1,\varepsilon} = \sum_i \sqrt{v_i^2 + \varepsilon}$
- Since $\rho > \beta$, the norm is *anti-penalized*

Interpretation

- Define for some radius r ,

$$\ell_\infty\text{-sharpness}(r) := \max_{\|\boldsymbol{\delta}\|_\infty \leq r} E(\boldsymbol{\theta} + \boldsymbol{\delta}) - E(\boldsymbol{\theta})$$

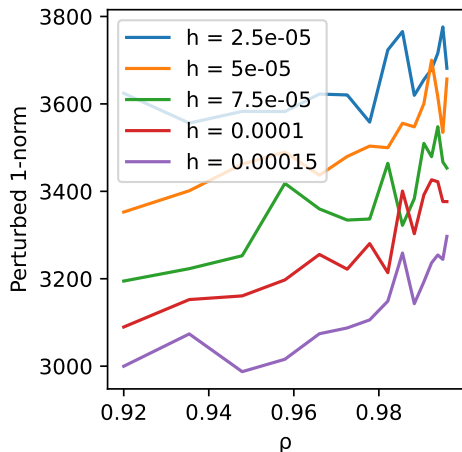
- Linearize under “max”:

$$\max_{\|\boldsymbol{\delta}\|_\infty \leq r} E(\boldsymbol{\theta} + \boldsymbol{\delta}) - E(\boldsymbol{\theta}) \approx \max_{\|\boldsymbol{\delta}\|_\infty \leq r} \nabla E(\boldsymbol{\theta})^\top \boldsymbol{\delta} = r \|\nabla E(\boldsymbol{\theta})\|_1$$

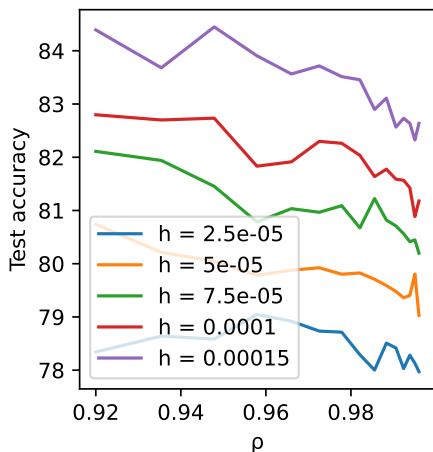
- Since for small ε , the perturbed one-norm is just the one-norm, **Adam anti-penalizes approximate ℓ_∞ -sharpness**
- This biases the trajectory towards “higher curvature” regions

Empirical evidence: increasing ρ

As ρ increases, the norm rises



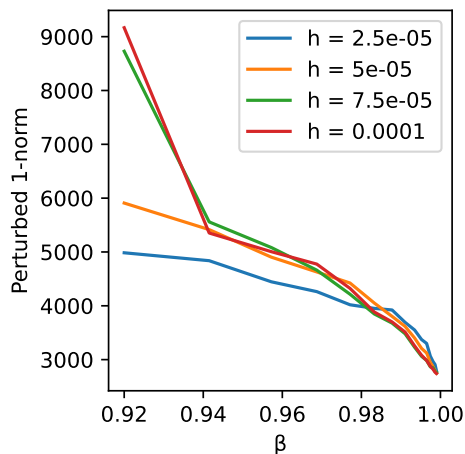
and test accuracy falls



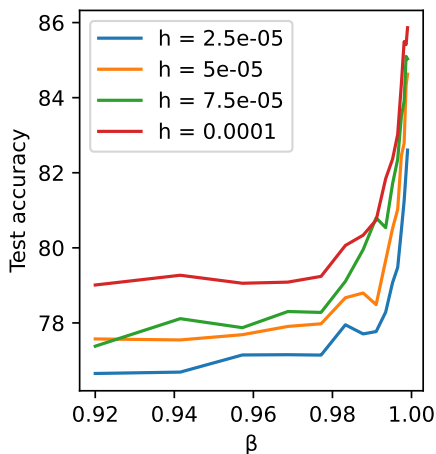
Resnet-50 on CIFAR-10 trained with full-batch Adam, $\epsilon = 10^{-8}$, $\beta = 0.99$

Empirical evidence: increasing β

As β increases, the norm falls



and test accuracy rises



Resnet-50 on CIFAR-10 trained with full-batch Adam, $\rho = 0.999$, $\varepsilon = 10^{-8}$

Other contributions

- Complete characterization of the full-batch bias:

typical case

| | ε "small" | ε "large" |
|-------------------|---|--|
| $\rho > \beta$ | $-\ \nabla E(\boldsymbol{\theta})\ _1$ -penalized | $\ \nabla E(\boldsymbol{\theta})\ _2^2$ -penalized |
| $\beta \geq \rho$ | $\ \nabla E(\boldsymbol{\theta})\ _1$ -penalized | $\ \nabla E(\boldsymbol{\theta})\ _2^2$ -penalized |

SGD (late at training)

- Full analysis and proof for the general **mini-batch** case
- Empirical verification for simple convolutional neural networks, ResNets, and Vision Transformers