Never Saddle for Reparameterized Steepest Descent as Mirror Flow

Tom Jacobs, Chao Zhou, and Rebekka Burkholz

Outline

1. Steepest flow
2. Reparameterizations and saddles induce mirror flow
3. Theory: saddle escape and incomplete Implicit bias
4. Practice: finetuning LLMs and reparameterized sparse training

Steepest flow

- Steepest flow family:

$$dx_t = -\left(\text{sign}(\nabla f(x_t))|\nabla f(x_t)|^{q-1} + \alpha x_t\right) dt, \qquad x_0 = x_{\text{init}}$$

- Gradient Flow (GF) is q=2.
- SignGF (q=1) relation to AdamW:

$$\beta_1 = \beta_2 = \epsilon = 0$$

Reparameterizations and saddles

We consider a diagonal linear model:

$$x = m \odot w$$

This leads to saddle points at zero.

Connection to mirror flow

- Dynamics is a steepest mirror flow.

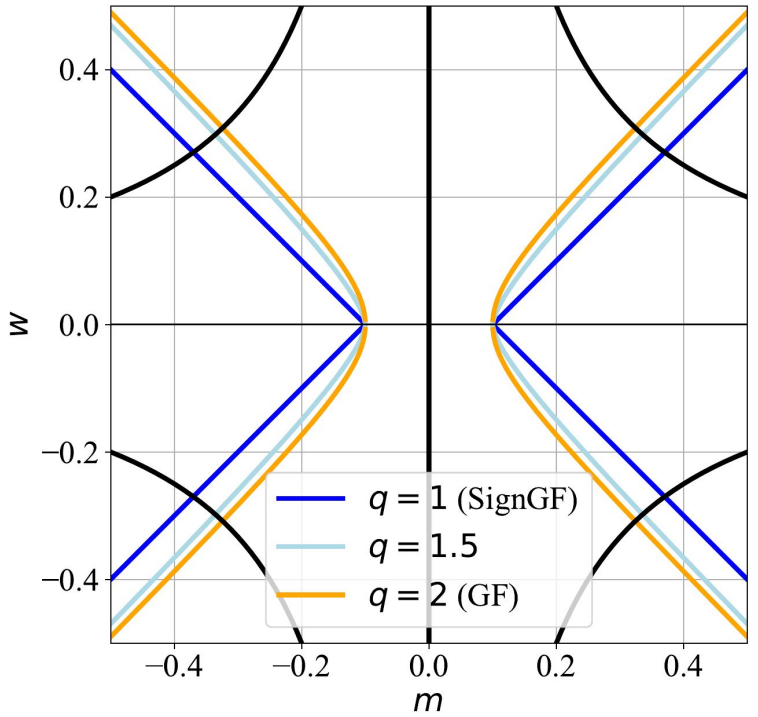
$$dx_t = -\nabla_x^2 R_{L_p,L}^{-1}(x_t)\ \mathrm{sign}\left(\nabla_x f\left(x_t\right)\right) \odot \left|\nabla_x f(x_t)\right|^{q-1} dt$$

- Reason: dynamics satisfy a new balance equation.

$$|m_t|^q - |w_t|^q = \lambda^q$$

- Faster convergence for smaller q.
- For L=2:

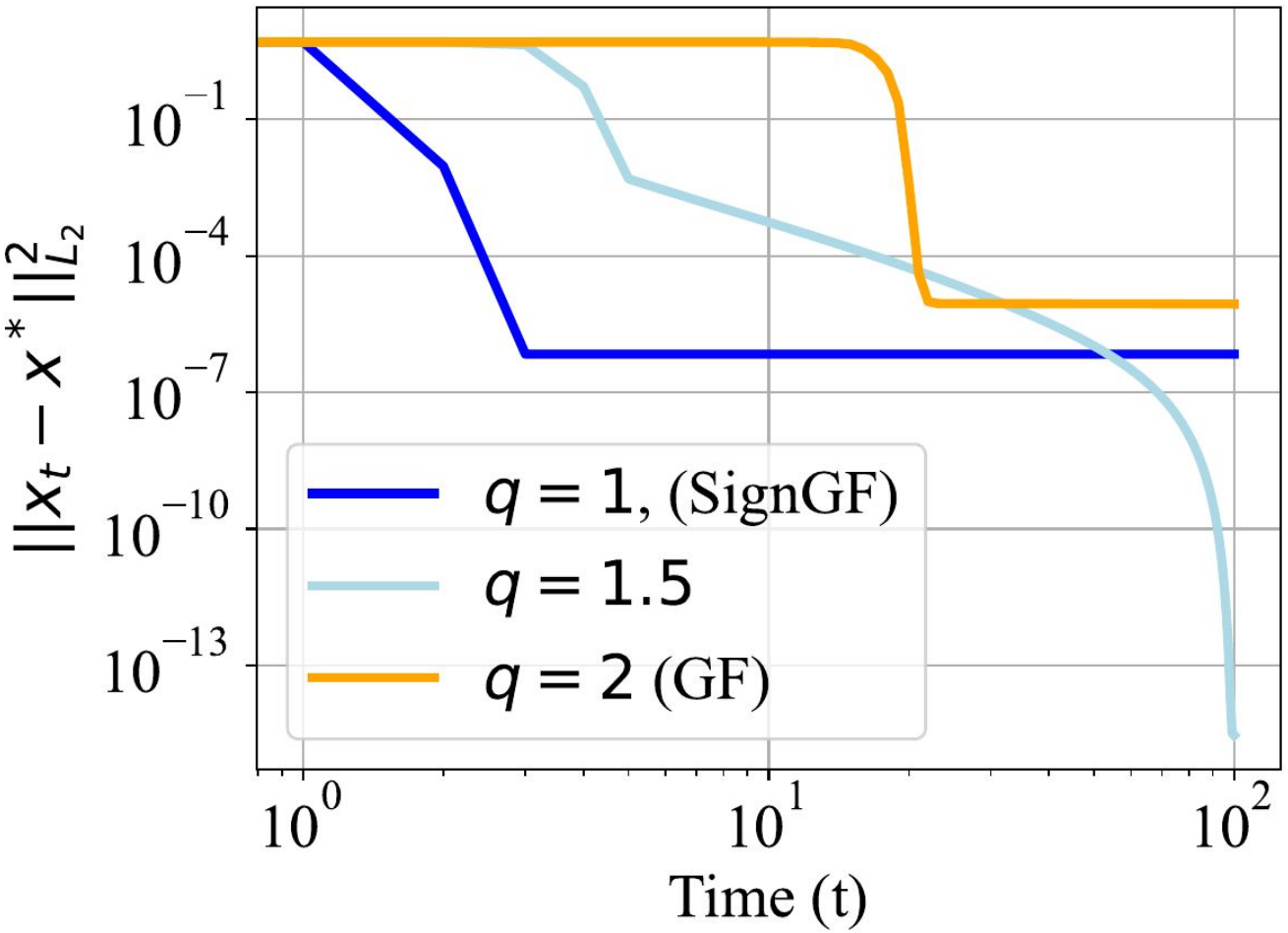
$$\nabla_x^2 R_{L_p,2}^{-1}(x) := \sqrt{4|x|^q + \lambda^{2q}}$$



Saddle escape

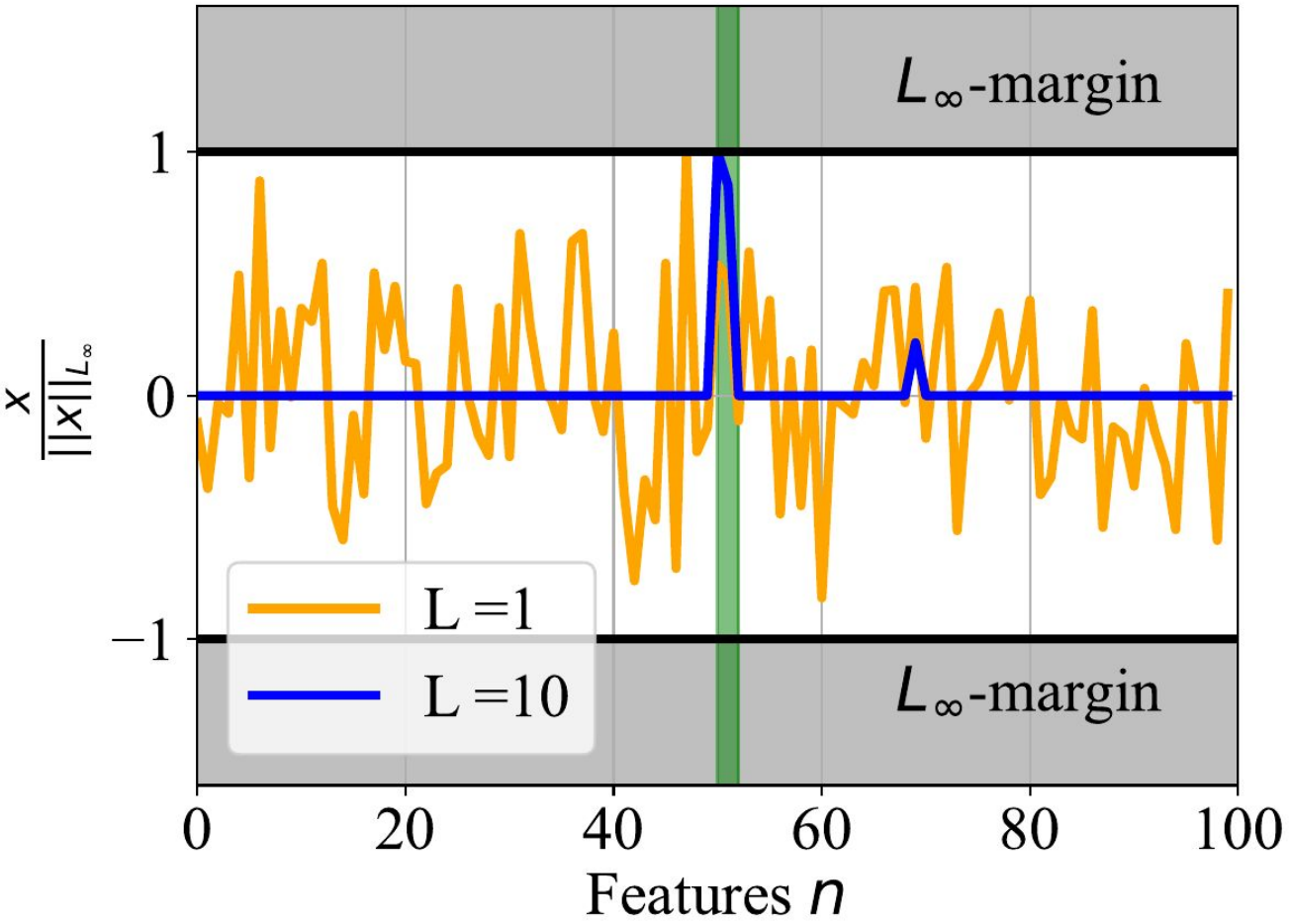
Provably faster saddle escape for smaller q.

$$\nabla_x^2 R_{L_p,2}^{-1}(x) := \sqrt{4|x|^q + \lambda^{2q}}$$



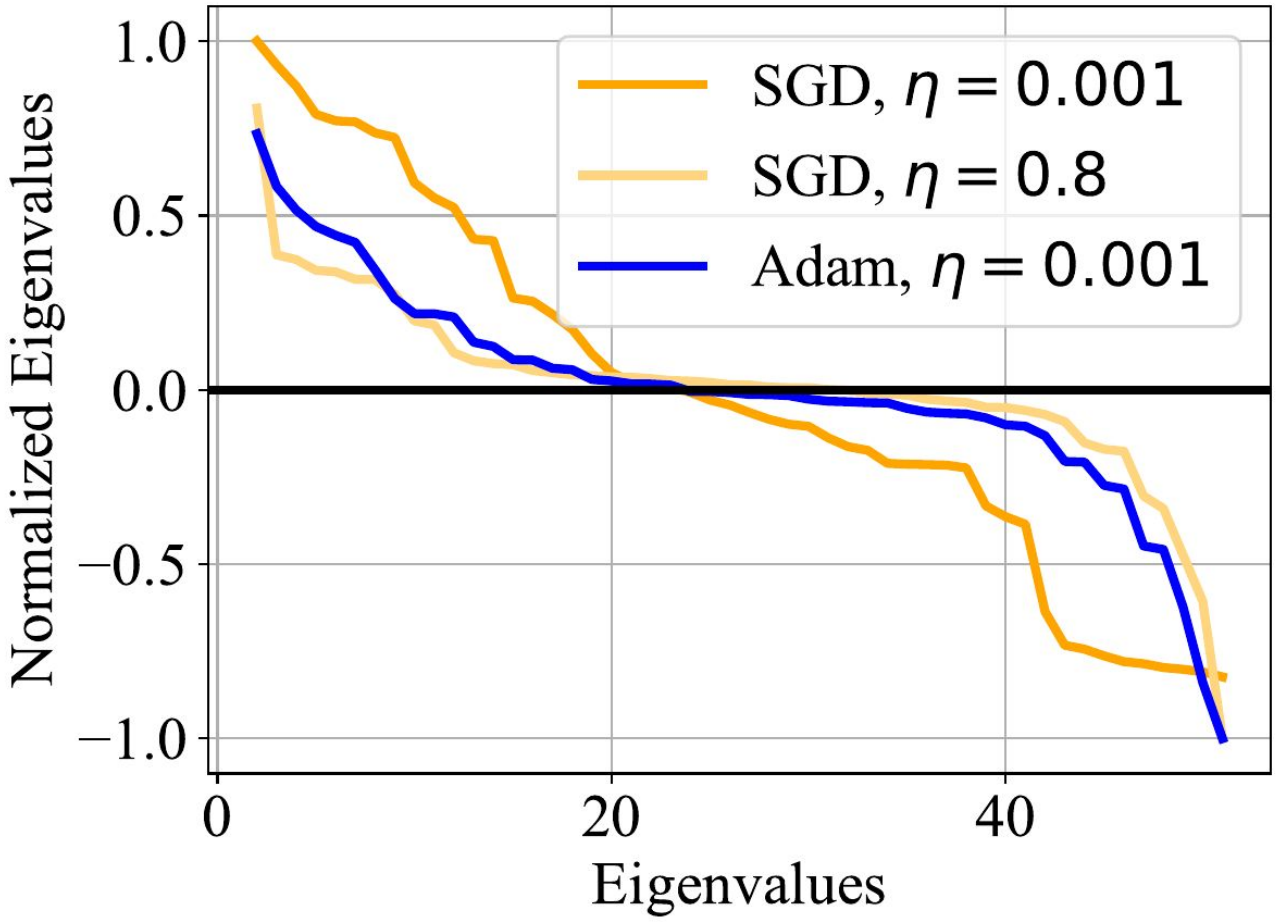
Incomplete implicit bias

- Both should have similar margin
- The geometry predicts the sparsity bias for large L



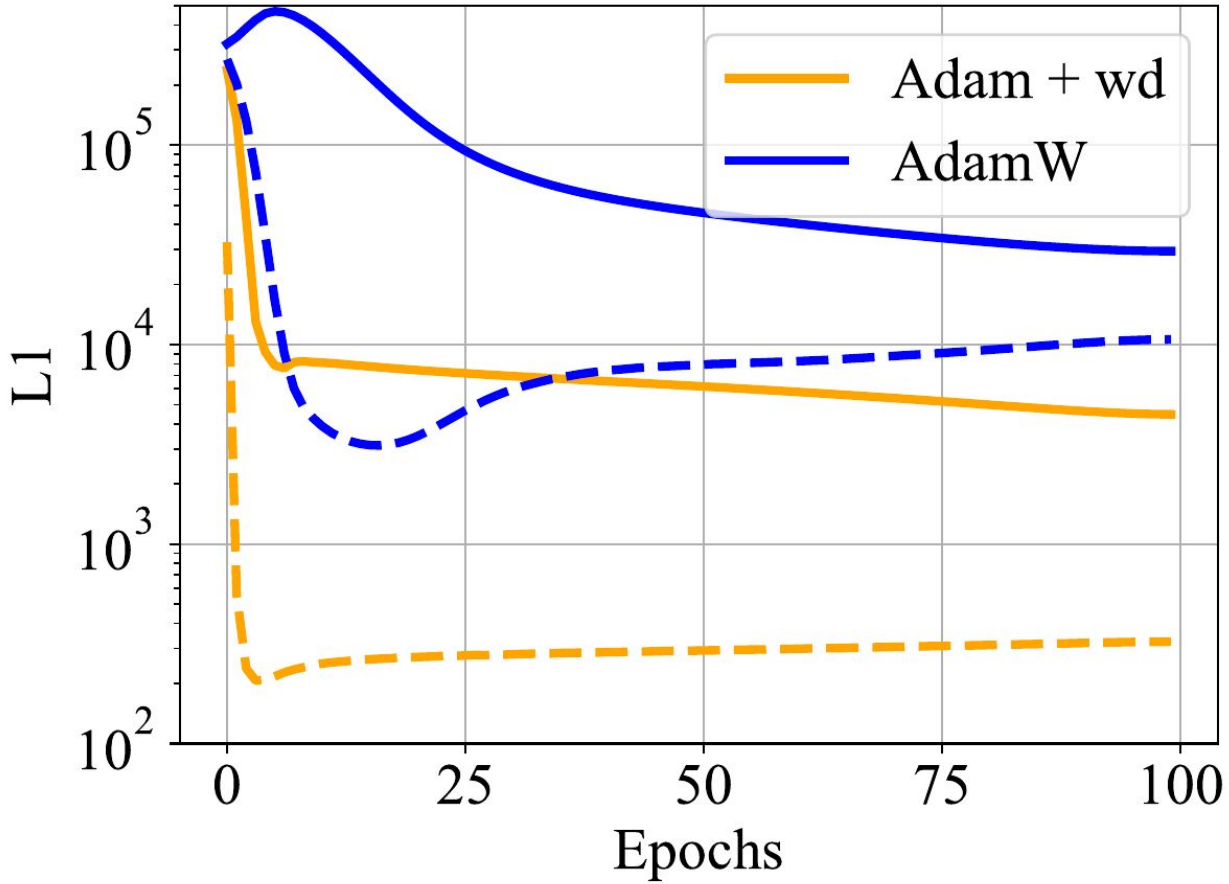
LLM finetuning

For small learning rate Adam more effectively escapes saddles.



Reparameterized sparse training

- Decoupled weight decay leads to less effective sparsity.
- Dotted lines are higher depth



Takeaway

Optimization geometry leads to saddle escape and controls the implicit bias.