Mirror, Mirror of the Flow: How Does Regularization Shape Implicit Bias?

Tom Jacobs Chao Zhou Rebekka Burkholz

CISPA Helmholtz Center for Information Security

Background and motivation

- The implicit bias characterized by mirror flow tries to explain the role of overparameterization in generalization.
- Explicit regularization (weight decay) is used in general.
- What is the effect of regularization on the implicit bias?

Optimization problem

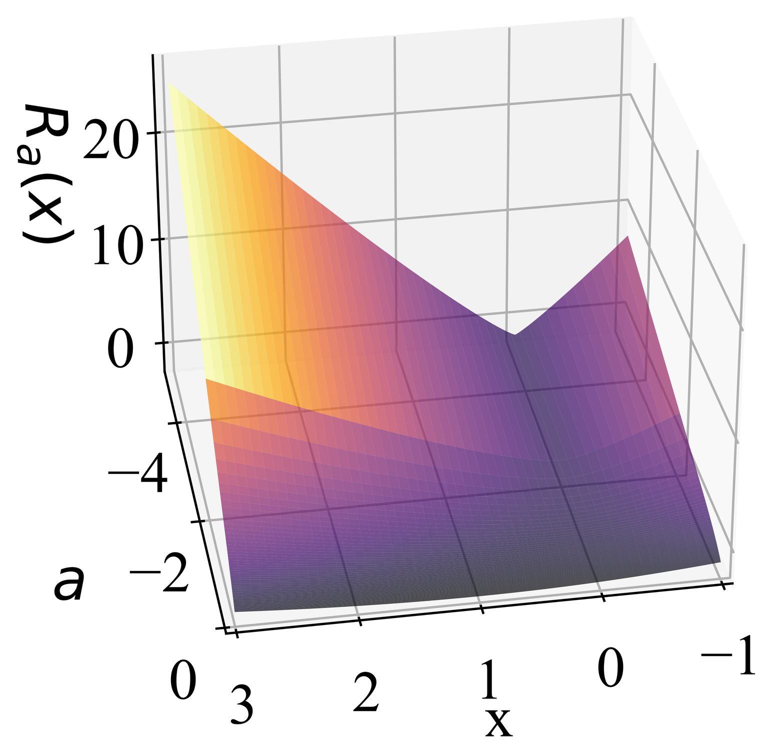
The optimization problem for a reparameterization $g : M \to \mathbb{R}^n$ and regularization $h : M \to \mathbb{R}$ is given by:

$$\min_{w \in M} f(g(w)) + \alpha h(w),$$

where, M is a smooth manifold and $\alpha \geq 0$ is the regularization strength. New: regularization h on the parameters w.

Example: quadratic reparameterization

Consider $x = g(m, w) = m \odot w$ with weight decay. The positional ($x_0 \to 0$) and type of bias ($L_2 \to L_1$) of R_a changes.

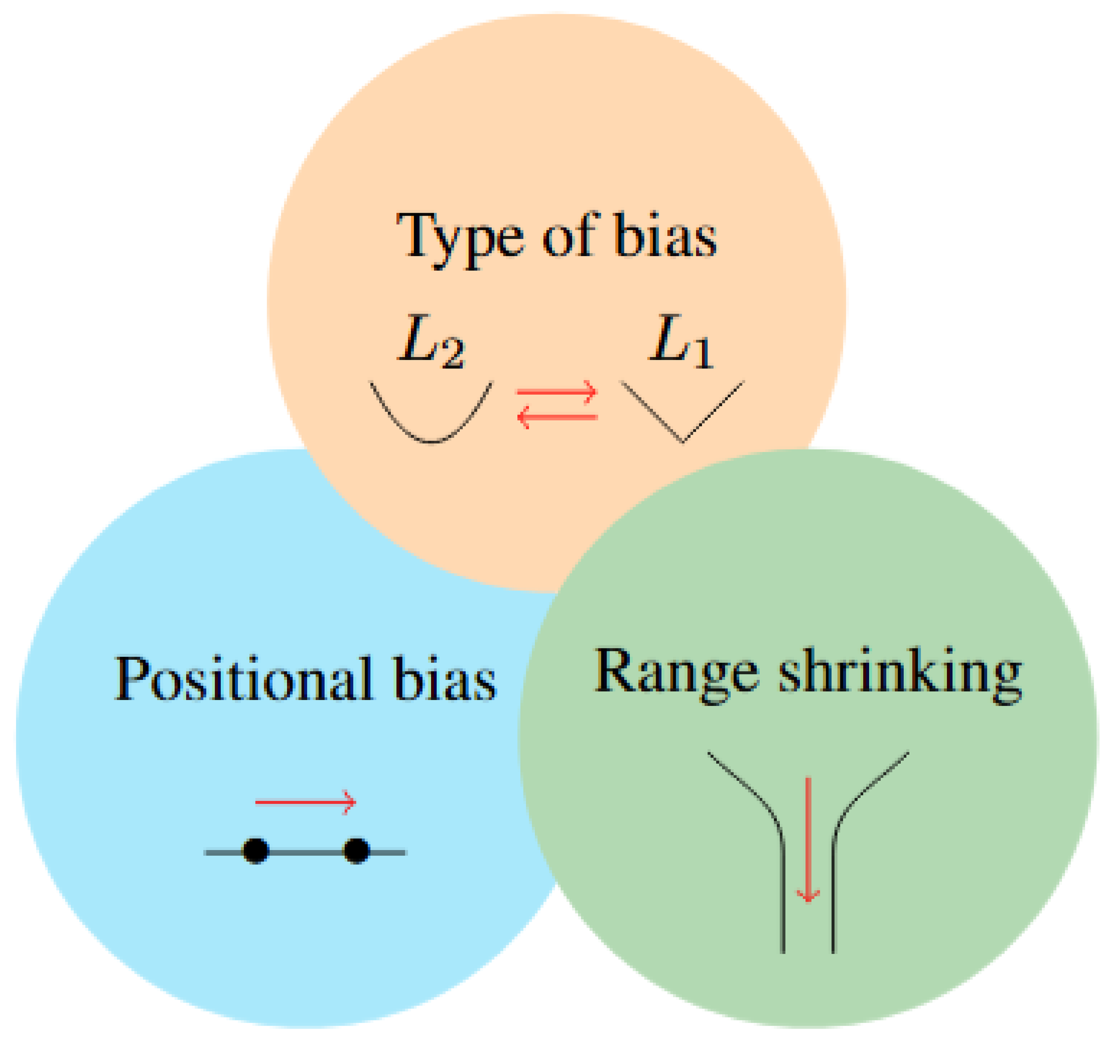


Main result: regularization in the mirror flow

The dynamics of $x_t = g(w_t)$, with conditions on (g, h) is:

$$d\nabla R_{a_t}(x_t) = -\nabla f(x_t)dt, \qquad x_0 = g(w_{\text{init}}),$$

where R_{a_t} is now a time varying Legendre function with $a_t = -\int_0^t \alpha_s ds$. This has a lasting effect on R_{a_t} (implicit bias).



Geometry: convergence and optimality

The dynamics of x_t, a changing Riemannian geometry, is

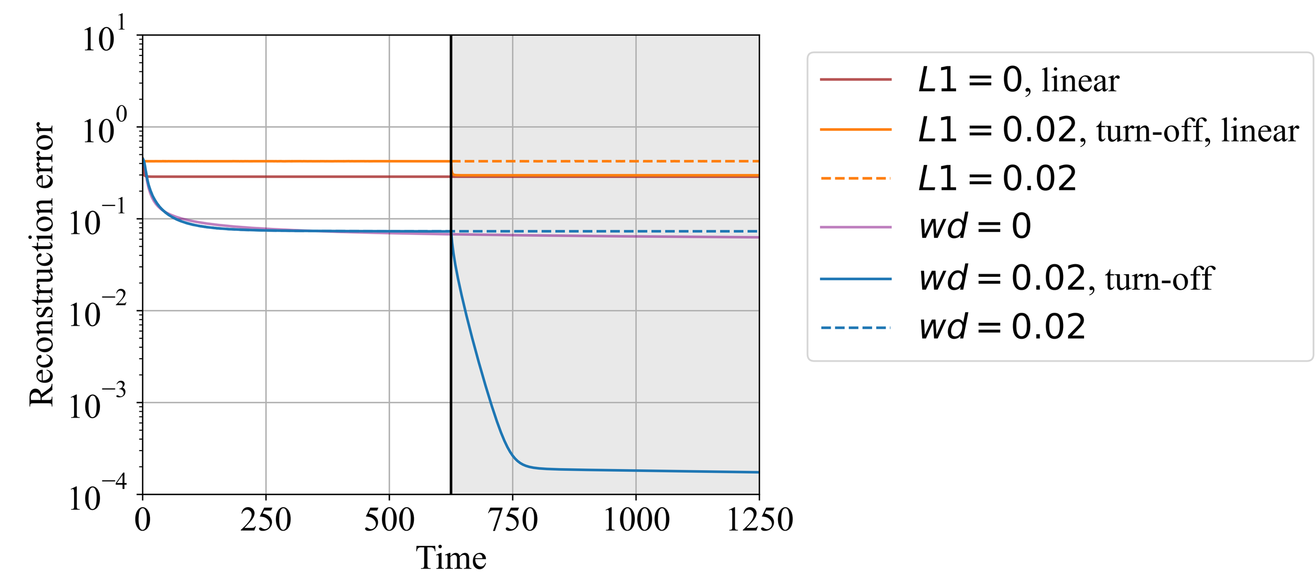
$$dx_t = -\left(\nabla_x^2 R_{a_t}(x_t)\right)^{-1} \left(\nabla_x f(x_t) + \alpha_t \nabla_x y_t\right) dt,$$

with initialization $x_0 = g(w_{init})$ and $y_0 = h(w_{init})$. This gives:

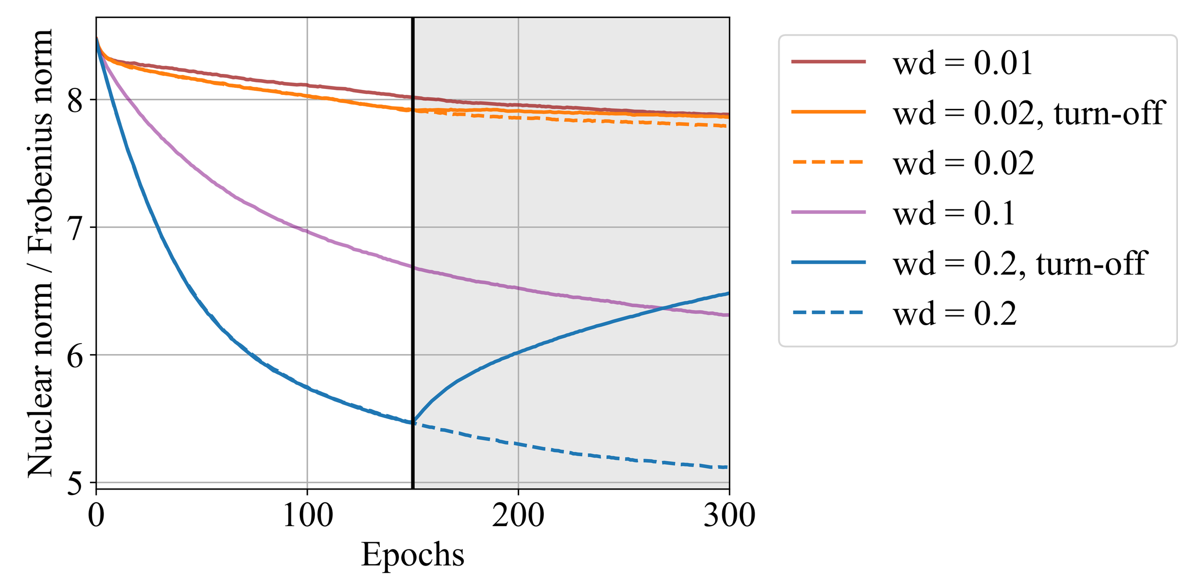
- The **convergence** for time varying Bregman functions
- The **optimality** for matrix sensing by turning off regularization i.e. $x_\infty = \text{argmin}_{x \in \mathbb{R}^n : xZ = y} R_{a_\infty}(x)$

Experiment: turn-off weight decay

- The theory is validated in a matrix sensing setup.
- Quad. reparam. recover the sparse ground-truth.



- Similar experimental insights hold on attention (and LoRA). For a ViT on ImageNet, this can boost validation accuracy by over 1% at similar relative sparsity.



Takeaway

- The regularization **controls** the Legendre function.
- The controllable change has a **lasting** effect.
- This could inform regularization strategies and enable new analysis of training dynamics.