Implicit regularization

Axel Böhm

July 5, 2023

Implicit regularization/bias

- why do very expressive models generalize?
- overparametrized/underdermined models
- with many global solutions
- no capacity control specified in the objective

Implicit regularization/bias

- why do very expressive models generalize?
- overparametrized/underdermined models
- with many global solutions
- no capacity control specified in the objective

Arora et al. (2019); Woodworth et al. (2020)). Different optimization choices, such as using different optimization methods (Gunasekar et al., 2018b), or different optimization parameters can significantly change the algorithmic bias, and thus completely change the effective inductive bias of learning and the ability to generalize in specific scenarios. In

Setting

only gradient descent (GD)

- least squares
- logistic regression
- (simple) nonlinear models

Examples of implicit biases

- algorithm used
- large **batch sizes** $\rightarrow$ sharp solutions [Keskar et al., 2017]
- large **step sizes** $\rightarrow$ flatter minima
- initialization
- early stopping

Underdetermined least squares

$$\min_w \|Xw - y\|^2$$

- We assume $\exists \bar{w} : X\bar{w} = y$
- Q: Which solution does GD converge to?[1]
- A: minimal distance to the starting point.

[1] irresspective of the step size

Underdetermined least squares

$$\min_w \|Xw - y\|^2$$

- We assume $\exists \bar{w} : X\bar{w} = y$
- Q: Which solution does GD converge to?[1]
- A: minimal distance to the starting point.

initialized at the origin we actually solve the **ridge regression**

$$\min_w \|Xw - y\|^2 + \lambda \|w\|^2$$

(equivalent to Bayesian regression with Gaussian prior).

[1] irrespective of the step size

Underdetermined least squares

$$\min_w \|Xw - y\|^2$$

- We assume $\exists \bar{w} : X\bar{w} = y$
- Q: Which solution does GD converge to?[1]
- A: minimal distance to the starting point.

initialized at the origin we actually solve the **ridge regression**

$$\min_w \|Xw - y\|^2 + \lambda \|w\|^2$$

(equivalent to Bayesian regression with Gaussian prior).

Also extends to momentum methods.

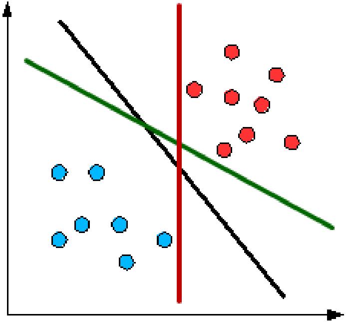
[1] irresspective of the step size

Binary classification with linearly separable data

$$\min_w \sum_{i=1}^{n} \exp(-y_i x_i^\top w)$$

- no solution exists
  (iterates diverge to infinity)
- GD converges to the **maximum-margin classifier**[a]
- the solution to hard-margin SVM

$$\min_w \|w\|^2 \text{ s.t. } y_i(w^\top x_i) \geq 1$$



[a] normalized iterates

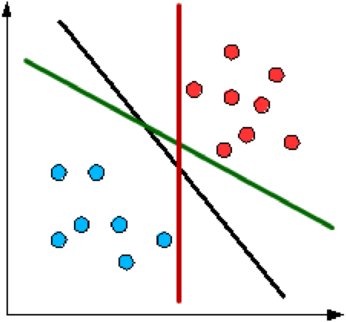
The convergence is **extremely slow** $\mathcal{O}(1/\log t)$.

Binary classification with linearly separable data

$$\min_{w} \sum_{i=1}^{n} \exp(-y_i x_i^\top w)$$

- no solution exists
  (iterates diverge to infinity)
- GD converges to the **maximum-margin classifier**[a]
- the solution to hard-margin SVM

$$\min_{w} \|w\|^2 \text{ s.t. } y_i(w^\top x_i) \geq 1$$



[a] normalized iterates

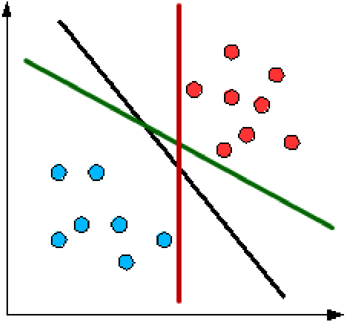
The convergence is **extremely slow** $\mathcal{O}(1/\log t)$.

Binary classification with linearly separable data

$$\min_{w} \sum_{i=1}^{n} \exp(-y_i x_i^\top w)$$

- no solution exists
  (iterates diverge to infinity)
- GD converges to the **maximum-margin classifier**[a]
- the solution to hard-margin SVM

$$\min_{w} \|w\|^2 \text{ s.t. } y_i(w^\top x_i) \geq 1$$



[a] normalized iterates

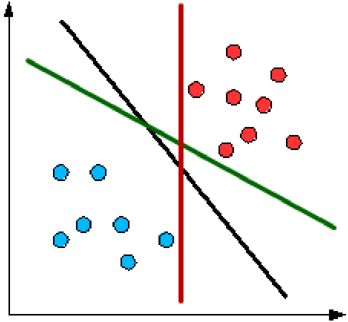
The convergence is **extremely slow** $\mathcal{O}(1/\log t)$.

Binary classification with linearly separable data

$$\min_w \sum_{i=1}^{n} \exp(-y_i x_i^\intercal w)$$

- no solution exists
  (iterates diverge to infinity)
- GD converges to the **maximum-margin classifier**[a]
- the solution to hard-margin SVM

$$\min_w \|w\|^2 \text{ s.t. } y_i(w^\intercal x_i) \geq 1$$



[a] normalized iterates

The convergence is **extremely slow** $\mathcal{O}(1/\log t)$.

least squares for non-Euclidian geometries

Mirror descent:

$$w_{t+1} = \arg\min_{w} \{\eta_t \langle w, \nabla \ell(w_t) \rangle + D_\psi(w, w_t)\}$$

- negative entropy $\psi(w) = \sum_i w_i \log w_i$
- Kullback-Leibler divergence D_ψ
- gives exponentiated/multiplicative gradient

least squares for non-Euclidian geometries

Mirror descent:

$$w_{t+1} = \arg\min_{w} \{\eta_t \langle w, \nabla \ell(w_t) \rangle + D_\psi(w, w_t)\}$$

- negative entropy $\psi(w) = \sum_i w_i \log w_i$
- Kullback-Leibler divergence D_ψ
- gives exponentiated/multiplicative gradient

least squares for non-Euclidian geometries

Mirror descent:

$$w_{t+1} = \arg\min_{w} \{\eta_t \langle w, \nabla \ell(w_t) \rangle + D_\psi(w, w_t)\}$$

- negative entropy $\psi(w) = \sum_i w_i \log w_i$
- Kullback-Leibler divergence D_ψ
- gives exponentiated/multiplicative gradient

In general: Iterates converge to

$$\arg\min_{w : Xw = y} D_\psi(w, w_0)$$

For KL divergence we get the maximum entropy solution.

Convex vs. Nonconvex

- previous models where convex
- observed bias *irresspective* of the step size
- non-convex case is more delicate then in convex GD
  - manifold spanned by the gradients is no longer flat
  - step size or momentum make you fall off

diagonal linear network, [Woodworth et al., 2020]

Consider the following nonconvex parametrization:

$$\min_u \sum_{i=1}^{n} \|\langle u \odot u, x_i \rangle - y_i\|^2.$$

- GD wrt to u can be seen as *mirror descent* in predictor space $w = u \odot u$
- the Bregman distance depends on initialization $\alpha\mathbf{e}$
- **small initialization** gives minimal L1 norm
  the "rich" regime
- **large initialization** gives known L2 regularization
  the "kernel" or "lazy" regime

diagonal linear network, [Woodworth et al., 2020]

Consider the following nonconvex parametrization:

$$\min_u \sum_{i=1}^n \|\langle u \odot u, x_i \rangle - y_i\|^2.$$

- GD wrt to u can be seen as *mirror descent* in predictor space $w = u \odot u$
- the Bregman distance depends on initialization $\alpha \mathbf{e}$
- **small initialization** gives minimal L1 norm
  the "rich" regime
- **large initialization** gives known L2 regularization
  the "kernel" or "lazy" regime

diagonal linear network, [Woodworth et al., 2020]

Consider the following nonconvex parametrization:

$$\min_u \sum_{i=1}^{n} \|\langle u \odot u, x_i \rangle - y_i\|^2.$$

- GD wrt to u can be seen as *mirror descent* in predictor space $w = u \odot u$
- the Bregman distance depends on initialization $\alpha \mathbf{e}$
- **small initialization** gives minimal L1 norm
  the "rich" regime
- **large initialization** gives known L2 regularization
  the "kernel" or "lazy" regime

diagonal linear network, [Woodworth et al., 2020]

Consider the following nonconvex parametrization:

$$\min_u \sum_{i=1}^{n} \|\langle u \odot u, x_i \rangle - y_i\|^2.$$

- GD wrt to u can be seen as *mirror descent* in predictor space $w = u \odot u$
- the Bregman distance depends on initialization $\alpha \mathbf{e}$
- **small initialization** gives minimal L1 norm
  the "rich" regime
- **large initialization** gives known L2 regularization
  the "kernel" or "lazy" regime

Caveat: holds for gradient flow.

with macroscopic step sizes [Nacson et al., 2022]

ity. In particular, we show how using large step size for non-centered data can change the implicit bias from a "kernel" type behavior to a "rich" (sparsity-inducing) regime — even when gradient flow, studied in previous works, would not escape the "kernel" regime. We do so by using

Matrix factorization

$$\min_{U,V} \sum_{i=1}^{n} \|\langle UV^\intercal, X_i \rangle - y_i\|^2$$

Similar results for infinitesimal step sizes

- implicit bias depends on initialization
- nuclear norm instead of L1 [Gunasekar et al., 2017]
- requires restricted isometry (RIP) [Li et al., 2018]

Thank You!

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2017).
Implicit regularization in matrix factorization.
Advances in Neural Information Processing Systems, 30.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2017).
On large-batch training for deep learning: Generalization gap and sharp minima.
In *International Conference on Learning Representations*.

Li, Y., Ma, T., and Zhang, H. (2018).
Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations.
In *Conference On Learning Theory*, pages 2–47. PMLR.

Nacson, M. S., Ravichandran, K., Srebro, N., and Soudry, D. (2022).
Implicit bias of the step size in linear diagonal neural networks.
In *International Conference on Machine Learning*, pages 16270–16295. PMLR.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. (2020).
Kernel and rich regimes in overparametrized models.
In *Conference on Learning Theory*, pages 3635–3673. PMLR.