Stochastic gradient descent on singular models

Mid-point presentation, PIBBSS 2023

Guillaume Corlouer

August 9, 2023

Motivation

- Developmental interpretability: How the training dynamics shape the learned model (ref)
- SLT: training is governed by singular regions of W and shift between singular regions happen via phase transitions (ref)
- Phase transitions might be natural stopping times to look for learned capabilities
- There might be a correspondence between the geometry of singular regions and specific functions that gets learn (ref)

Salient open questions

Many important empirical questions but I focus on theory:

- Understand how the geometry of singular regions affect SGD
- From this suggest estimates of geometric invariant for NN
- Other important geometric and topological invariants?

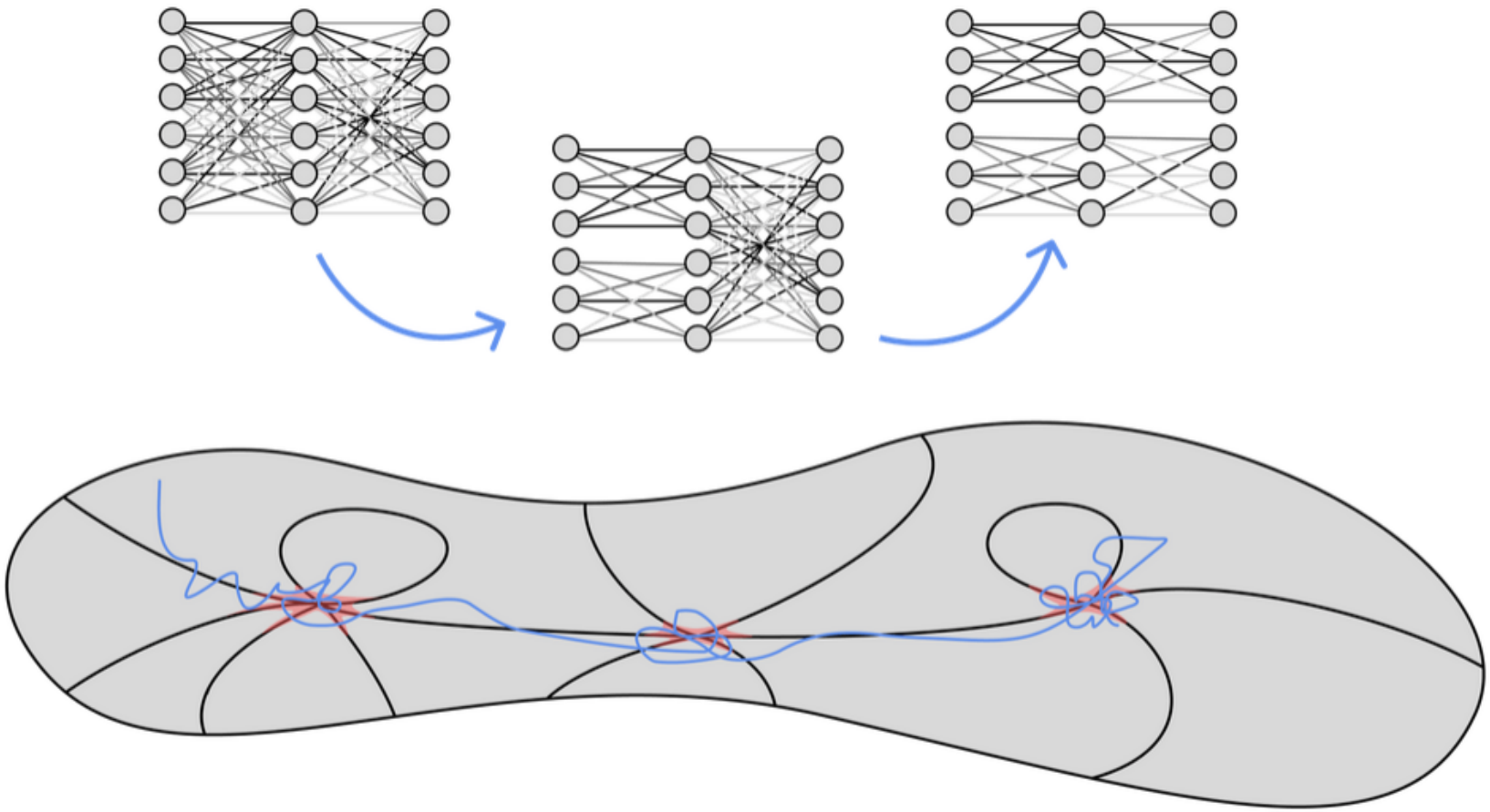


Figure: SGD (blue) shape the learned architecture. Posterior accumulate around singular regions (red) belonging to different phases. Taken from (ref)

Setup: Learning is free energy minimisation

- Data $X := \{X_1, ..., X_n\}$ with $X_i \sim q$ i.id., q is true distribution
- Model given by $p(X|w)$ such that $\exists w_0,\ q(X) = p(X|w_0)$
- Find w^* such that $K(w) := KL[q||p(X|w)]$ is minimised
- Empirical KL: $K_n(w) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{q(X_i)}{p(X_i|w)}\right)$
- Bayesian posterior: $p(X|w) = \frac{1}{Z}e^{-nK_n(w)}\varphi(w)$
- Partition function: $Z = \int_W e^{-nK_n(w)}\varphi(w)dw$
- Free energy: $F_n = -\ln Z$
- Restricted free energy: $F_n(W_\alpha) = -\ln \int_{W_\alpha} e^{-nK_n(w)}\varphi(w)dw$
- Learning $\iff$ Minimising free energy $\iff$ Internal model selection (ref)
- Selected model is in phase W_α with lowest F_n
- Analogous to statistical physics

Singular models

- At a critical point w^* minimising $K(w)$, a model is singular if $\det(\nabla_w^2 K(w^*)) = 0$
- Intuitively there are flat directions around w^*

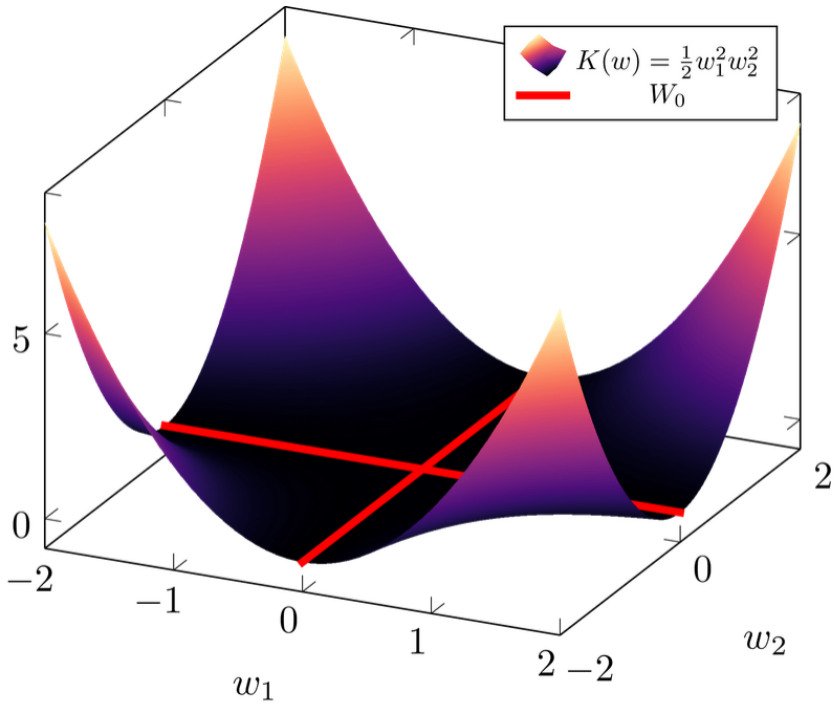
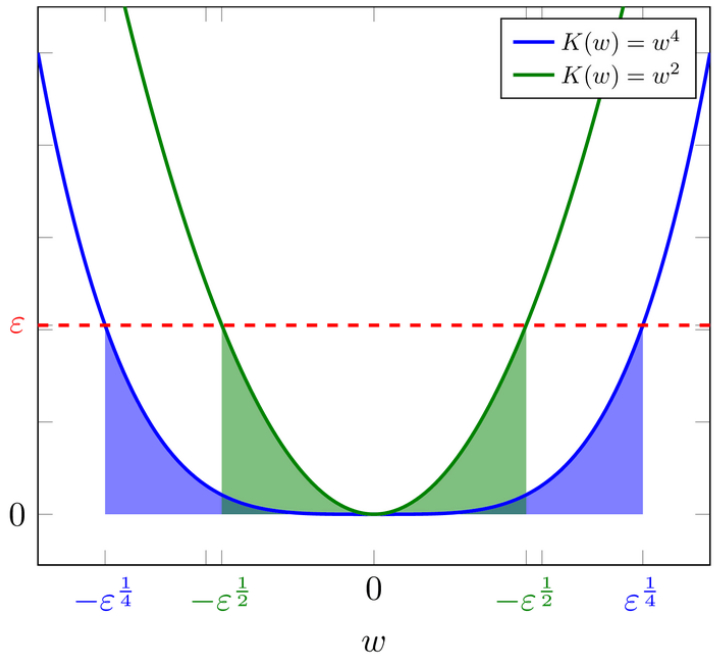


Figure: Example of a 2d singular model, singular region W_0 corresponds to two flat directions in the neighborhood of a critical point at the intersection (taken from ref)

The real log canincal threshold (RLCT) measures effective dimensionality

- Effective dimension is given by a rational number λ (RLCT) which depends on the geometry of the singularity
- Around a critical point w^*, take a ball of radius ϵ. One can show that $V(\epsilon) \propto \epsilon^{\lambda}$



The free energy formula

Free energy is a trade-off between model accuracy and "complexity"

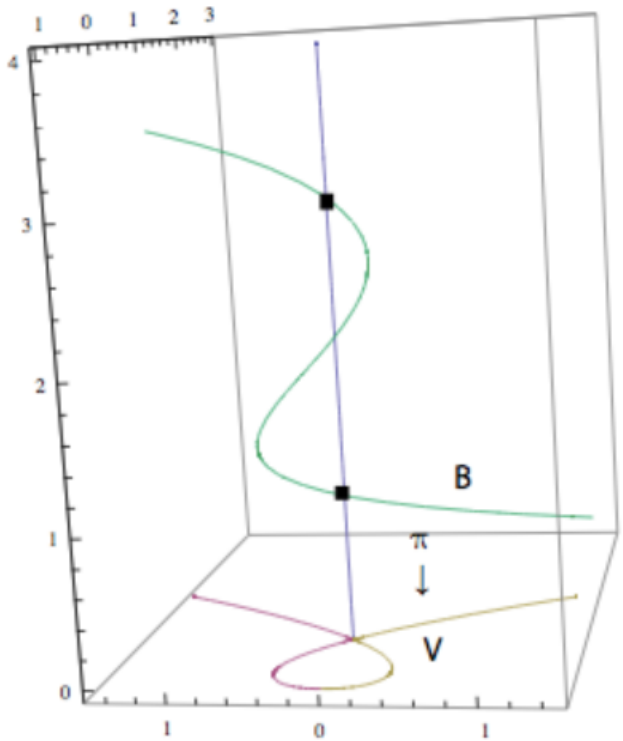
In large sample (ref):

$$F_n = \underbrace{nL_n(w^*)}_{\text{Accuracy}} + \underbrace{\lambda \log n}_{\text{Effective dimension}} + \underbrace{O(\log \log n)}_{\text{Lower order terms}}$$

- For regular models, we recover the Bayesian information criterion (BIC) familiar to statisticians ($\lambda = \frac{d}{2}$)
- For two different phases W_1 and W_2 with same accuracy, the posterior will converge toward the phase with lower RLCT.
- During training, selected model can be different from the truth if they have slightly lower accuracy but much lower RLCT

Resolutions of singularity

- Problem: Doing maths on singular space can be too difficult
- For a singular space W_0 find a map $g : \mathcal{M} \to W_0$
- $\mathcal{M}$ is a higher dimensional space than W_0 which is smooth and compact ("nice")



Stochastic gradient descent

- A deep neural network (DNN) is a function:

$$f : \mathcal{X} \times W \to \mathcal{Y}$$
$$(x, w) \mapsto f(x, w)$$

- Typically we have $f = W_b^L \prod_{l=L-1}^{1} (\sigma \circ W_b^l)$;
- The training process search the parameters minimising the empirical loss $L_n^f(w) := \frac{1}{2n} \sum_{i=1}^{n} ||y_i - f(x_i, w)||^2$
- For the likelihood $p((x, y)|w) \propto e^{-nL_n^f(w)}$, this is the same as minising $K_n(w)$
- A DNN update its weights via SGD. For a random sample of indices $b(t) \subset \{1, ..., n\}$, at time t:

$$w_{t+1} = w_t - \eta \nabla_{w_t} K_{b(t)}(w_t)$$

- Important point: **Deep learning models are singular** (ref).
- What SLT can tell us about SGD?

First fundamental theorem of SLT

Let $W_\epsilon := \{w \in W | K(w) \leq \epsilon\}$,under some mild assumptions (ref):

There exist a resolution of singularities $g : \mathcal{M} \to W_\epsilon$ such that for every local coordinate $u \in U \subset \mathcal{M}$:

$$K_n(g(u)) = u^{2k} - \frac{1}{\sqrt{n}} u^k \xi_n(u)$$

$$K(g(u)) = u^{2k}$$

Where $\xi_n(u)$ is an empirical process that converges in distribution to a gaussian process $\xi(u)$ with 0 mean and variance 2.

Intuitively, the formula means that K_n fluctuates around $K(u)$ via ξ_n. In particular, K_n fluctuates around the preimage of the truth in the resolution via ξ_n.
We can learn much more about $\xi(u)$ which is a Gaussian processes. Gaussian process are great!

The probability of excursion of ξ

Intuition: Understanding how the geometry ξ near singularities is important to understand how SLT and SGD can be related.

The geometry of random fields

Applying key results from (ref), the probability of excursion $P[|sup_{\mathcal{M}}\xi| \geq t]$ of the Gaussian field ξ is constrained by the geometry of the resolution $\mathcal{M}$ of dimension d via:

$$P[|sup_{\mathcal{M}^{\mathcal{R}}}\xi| \geq t] = \frac{(2\pi)^{-d}}{2} e^{-t^2/4} \left[t^{d-1} Vol_g(\mathcal{M}) + q(d-2) \right] + O(e^{-\alpha t^2/4})$$

Where q is a polynomial of degree $d-2$, and $Vol_g(\mathcal{M})$ is the "volume" of the resolution, α is a large constant that depends on the geometry of the resolution.

The geometry of the resolution constrains the probability of excursion via its dimension and volume, which upper bounds the difficulty of SGD to escape the singularity.

Next steps

- The previous formula needs clarification. In particular how does it relates to RLCT?
- Gain a deeper understanding of SLT and the previous formula
- Compute the probability of excursion for simple 2D singular models
- Clarify its meaning for SGD on toy neural networks
- Does it yields more interesting invariants that are easier to compute (using the Gaussianity of ξ?)
- How does the phase transition picture change with SGD?
- Jet schemes? Exceptional divisor?
- Explore connections with the replica methods and Parisi potential?