Singular learning theory for stochastic gradient descent

Guillaume Corlouer

PIBBSS summer 2023

July 20, 2023

Mentor : *Nicolas Macé*

Motivation

- Developmental interpretability
- SLT: Singular regions of W are attractors of training
- Shift between singular regions happens via phase transitions
- Phase transition: stop training and look for learned capabilities
- Correspondence between the geometry of singular regions and learned programs?

Salient open questions

Many important empirical questions but I focus on theory:

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

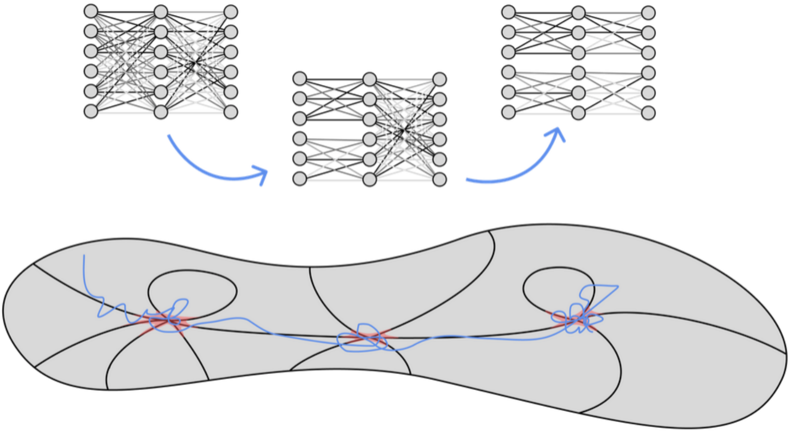


Figure: SGD (blue) shapes the learned architecture. The bayesian posterior narrows around singular regions (red) belonging to different phases.

Setup: Learning as free energy minimisation

- Data $X := \{X_1, ..., X_n\}$ with $X_i \sim q$ i.id., q is true distribution
- 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: $Z_n = \int_W e^{-nK_n(w)}\varphi(w)dw$
- Free energy: $F_n = -\ln Z_n$
- 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
- Selected model is in phase W_α with lowest F_n
- Analogous to statistical physics

Singular models

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

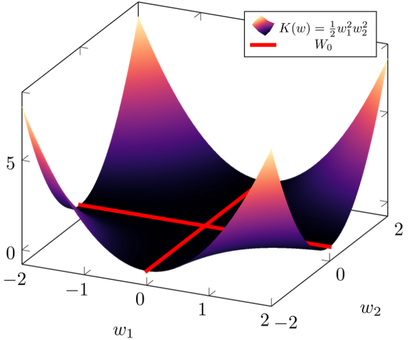
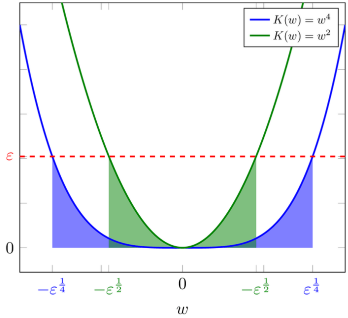


Figure: Example of a 2d singular model. The singular region W_0 is the union of two regions that are flat in the neighborhood their intersection

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^* minimising $K(w)$, consider 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 :

$$F_n = \underbrace{L_n(w^*)}_{\text{Accuracy}}\, n + \underbrace{\lambda}_{\text{Dimension}} \log n + \underbrace{O(\log \log n)}_{\text{Lower order terms}}$$

- 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, a model with slightly lower accuracy but much lower RLCT might be selected

Stochastic gradient descent

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

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

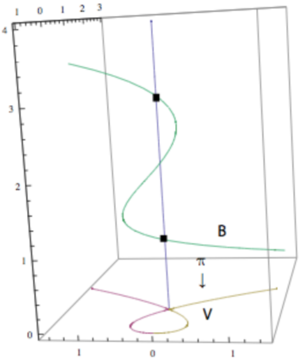
- Typically we have $f = W_b^L \prod_{l=L-1}^{1} (\sigma \circ W_b^l)$;
- The training process search the parameters minimising the empirical loss $L_n^f(w) := \frac{1}{2n} \sum_{i=1}^{n} ||y_i - f(x_i, w)||^2$
- For the likelihood $p((x, y)|w) \propto e^{-nL_n^f(w)}$, this is equivalent to minising $K_n(w)$
- 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** .
- What SLT can tell us about SGD?

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")



First fundamental theorem of SLT

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

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 mean 0 and variance 2.

Intuitively, the formula means that K_n fluctuates around $K(u)$ via ξ_n. In particular, K_n fluctuates around the pre-image of the critical point w^* in the resolution via ξ_n.
We can learn much more about $\xi(u)$ which is a Gaussian processes.
Gaussian process are much more tractable!

The probability of excursion of ξ

Intuition: Understanding how the geometry of $\mathcal{M}$ affects ξ and $\nabla\xi$ near singularities is a promising step 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 d-dimensional resolution $\mathcal{M}$ via:

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

Where q is a polynomial of degree $d-2$, and $Vol_g(\mathcal{M})$ is the volume of the resolution via some metric g induced by ξ - a corrected Fisher metric, 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

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?
- Link with jet schemes? Exceptional divisor?
- Explore connections with the replica methods and Parisi potential?
- Look for extension of SGD as approximate Bayesian inference for singular models

References I

Robert J Adler. *The geometry of random fields*. SIAM, 2010.

James Clift, Daniel Murfet, and James Wallbridge. Geometry of program synthesis. *arXiv preprint arXiv:2103.16080*, 2021.

Jesse Hoogland et. al. Toward developental interpretability, 2023. URL `https://www.lesswrong.com/s/SfFQE8DXbgkjk62JK/p/TjaeCWvLZtEDAS5Ex`.

Carroll Liam. Distilling singular learning theory, 2023. URL `https://www.lesswrong.com/s/czrXjvCLsqGepybHC`.

Sumio Watanabe. *Algebraic geometry and statistical learning theory*, volume 25. Cambridge university press, 2009.

Susan Wei, Daniel Murfet, Mingming Gong, Hui Li, Jesse Gell-Redman, and Thomas Quella. Deep learning is singular, and that's good. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.