Singular learning theory: from Bayesian statistics to Machine Learning

Simon Pepin Lehalleur

April 2, 2024

Introduction

- Singular learning theory (SLT) is an application of singularity theory to Bayesian statistics developed by Sumio Watanabe.

- Recent applications of singular learning theory to machine learning, specifically to developmental interpretability of deep neural networks, due to Daniel Murfet and his collaborators (Liam Carroll, Zhongtian Chen, Matthew Farrugia-Roberts, Zach Furman, Jesse Hoogland, Edmund Lau, Jack Mendel, Stan van Wingerden, George Wang, Susan Wei)

Key result of singular learning theory

Rough form

The asymptotic performance of parametric statistical models is controlled by the singularities of a function associated to the model and the data-generating process.

Precise form

The asymptotic Bayesian generalization performance of

- real-analytic parametric statistical models for i.i.d data is controlled by
- the real log-canonical threshold of
- the relative entropy between the data-generating distribution and the model.

Plan

I: **Singularity theory and real log-canonical thresholds**

II: **Classical Bayesian statistics**

III: **Statistical learning theory of regular and singular models**

IV: **SLT from theory to practice**

V: **Machine learning and Developmental Interpretability**

References on real log-canonical thresholds

- [Atiyah] M. Atiyah, **“Resolution of Singularities and Division of Distributions”**
- [AGZV2] V. Arnold, S. Gusein-Zade, A. Varchenko, **“Singularities of differentiable maps volume II”**
- [Saito] M. Saito, **“Real log-canonical thresholds”**

References on Singular learning theory in Bayesian statistics

- [Watanabe09] S. Watanabe, **"Algebraic geometry and statistical learning theory"** (Cambridge University Press!)
- [Watanabe18] S. Watanabe, **"Mathematical theory of Bayesian statistics"**
- [WBIC] S. Watanabe, **"A widely applicable Bayesian information criterion"**
- [Watanabe_survey] S. Watanabe, **"Recent advances in algebraic geometry and Bayesian statistics"**
- [Lin_thesis] S. Lin, **"Algebraic Methods for Evaluating Integrals in Bayesian Statistics"**
- [Aoyagi-DLN] M. Aoyagi, **"Consideration on the learning efficiency of multiple-layered neural networks with linear units"**
- [Nagayasu-Watanabe] S. Nagayasu, S. Watanabe, **"Asymptotic behavior of free energy when optimal probability distribution is not unique"**

References on Developmental Interpretability

- [DL-sing] D. Murfet et al, **“Deep learning is singular, and that’s good”**
- [LLC] E. Lau, D. Murfet, S. Wei, **“Quantifying degeneracy in singular models via the learning coefficient”**
- [LLC-scale] Z. Furman, E. Lau, **“ Estimating the Local Learning Coefficient at Scale”**
- [TMS] Z. Chen, E. Lau, J. Mendel, S. Wei, D. Murfet, **“Dynamical versus Bayesian Phase Transitions in a Toy Model of Superposition ”**
- [ICL] J. Hoogland, G. Wang, M. Farrugia-Roberts, L. Carroll, S. Wei, D. Murfet, **“ The Developmental Landscape of In-Context Learning”**

Singularity theory and real log-canonical thresholds

Overview

- Real-analytic functions and their singularities
- Two key examples
- Real-log canonical threshold and its geometric interpretations
- More examples
- Wider context in singularity theory

Set-up

- $W \subseteq \mathbb{R}^d$ compact subset with non-empty interior $\mathring{W} \neq \emptyset$.

  Assume W semi-analytic, i.e. given by real-analytic inequalities.
- $F : W \to \mathbb{R}$ real-analytic function.

  i.e. F restriction of a real-analytic function on an open neighbourhood of W.

Remarks

- *In our applications $F \geq 0$ and we concentrate on this case.*
- *One should be treat carefully what happens at boundary ∂W; in this talk we will mostly pretend "$W = \mathring{W}$".*
- *Important special case: $F \in \mathbb{R}[w_1, \dots, w_d]$ polynomial. Pros:*
  - * *Explicit computations with computer algebra systems.*
  - * *Can bring in tools from (real) algebraic geometry*
  - * *Can sometimes reduce real-analytic situations to algebraic ones*

However statistical applications involve non-polynomial functions!

Critical points and singularities

- A point $w \in W$ is a zero of F if $F(w) = 0$. We define

$$W_0 := F^{-1}(0) = \{w \in W \mid F(w) = 0\}.$$

- A point $w \in W$ is a critical point of F if

$$\nabla F(w) = 0 \iff \forall i,\ \frac{\partial F}{\partial w_i}(w) = 0.$$

- A point $w \in W$ is a singularity of F if it is both a zero and a critical point.

Example

- Local minima and maxima of F are critical points.
- If $F \geq 0$, then every $w \in W_0$ is a (global) minimum and hence a singularity of F.

First key example: sum of squares

Take $W = B(0, R)$ closed ball of radius R and

$$F(w) = w_1^2 + \ldots + w_d^2.$$

Note that $F \geq 0$, we have $W_0 = \{0\}$ and F has a unique singularity at 0.

(Vol) Consider the sublevel set:

$$B_F(\epsilon) := \{w \in W | \ |F(w)| \leq \epsilon\}.$$

Then in our case $B_F(\epsilon) = B(0, \sqrt{\epsilon})$ so $\mathrm{Vol}\ B_F(\epsilon) = \frac{\pi^d}{\Gamma(\frac{d}{2}+1)} \epsilon^{\frac{d}{2}}$.

Volume scaling of sublevel sets

$$\mathrm{Vol}\ B_F(\epsilon) \underset{\epsilon \to 0}{\sim} C\ \epsilon^{\frac{d}{2}} \quad \text{(for some } C > 0\text{)}$$

Moreover, the volume concentrates around W_0 as $\epsilon \to 0$.

First key example: sum of squares

(Zeta) Define for $s \in \mathbb{C}$ with $\mathrm{Re}(s) >> 0$ the zeta function:

$$\zeta_F(s) := \int_W |F(w)|^s dw.$$

In our case we have by integrating over spheres:

$$\zeta_F(s) = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})} \int_{r=0}^{R} r^{2s+d-1} dr$$

hence

Meromorphic continuation and poles of ζ_F

- $\zeta_F(s)$ has meromorphic continuation to $\mathbb{C}$ with poles in $\mathbb{R}_{<0}$
- the largest pole of $\zeta_F(s)$ is $-\frac{d}{2}$ and has order 1

First key example: sum of squares

(Laplace) We have

$$\int_W e^{-nF(w)} dw \underset{n\to\infty}{\sim} \int_{\mathbb{R}^d} e^{-nF(w)} dw$$

and

$$\int_{\mathbb{R}^d} e^{-nF(w)} dw \overset{\text{Fubini}}{=} \left(\int_{\mathbb{R}} e^{-nw^2} dw\right)^d \overset{\text{Gaussian}}{=} \left(\sqrt{\frac{\pi}{n}}\right)^d .$$

Hence

Asymptotics of Laplace-type integral

$$Z_F(n) := \int_W e^{-n|F(w)|} dw \underset{n\to\infty}{\sim} C\ n^{-d/2}.$$

Moreover, the integral concentrates around W_0 as $n \to \infty$.

Second key example: monomial

Let $W = [-1,1]^d$. Fix $k_1, ..., k_d \in \mathbb{N}$ and let F be the monomial function

$$F(w) = w_1^{k_1} \dots w_d^{k_d}$$

$F \geq 0$ iff all k_i are even. We have $W_0 = [-1,1]^d \cap \bigcup_{i=1}^d \{w_i = 0\}$. The singularities of F are all of $W_0 \Rightarrow$ non-isolated singularities.

(Vol) Exercise:

Volume scaling of sublevel sets

$$\mathrm{Vol}\ B_F(\epsilon) \underset{\epsilon \to 0}{\sim} C\ \epsilon^{\lambda}\ (-\log(\epsilon))^{m-1}$$

with

$$\lambda = \min_i \left\{ \frac{1}{k_i} \right\} \text{ and } m = \# \left\{ i \mid \frac{1}{k_i} = \lambda \right\}.$$

Moreover, the volume concentrates around the subset $W_0^{\mathrm{deg}} := [-1,1]^d \cap \bigcup_{\frac{1}{k_i} = \lambda} \{w_i = 0\} \subseteq W_0$ as $\epsilon \to 0$.

Second key example: monomial

(Zeta) We compute the zeta function for a monomial:

$$\zeta_F(s) \overset{\text{Symmetry}}{=} 2^d \int_{[0,1]^d} |F(w)|^s dw \overset{\text{Fubini}}{=} 2^d \prod_{i=1}^{d} \int_{w_i=0}^{1} |w_i|^{k_i s} dw_i$$

Each factor converges iff $\mathrm{Re}(s) > -\frac{1}{k_i}$, and we see easily that

Meromorphic continuation and poles of ζ_F

- $\zeta_F(s)$ has meromorphic continuation to $\mathbb{C}$ with poles in $\mathbb{R}_{<0}$.
- the largest pole of $\zeta_F(s)$ is $-\lambda$ and has order m.

Second key example: monomial

(Laplace) Harder exercise:

Asymptotics of Laplace-type integral

When F is a monomial, we have

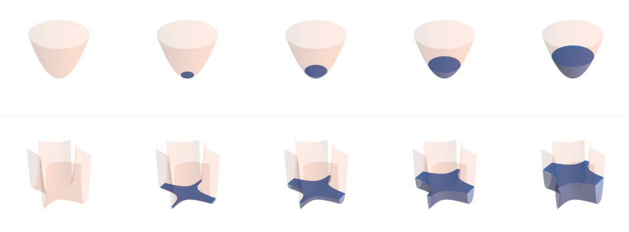
$$Z_F(n) = \int_W e^{-n|F(w)|} dw \underset{n \to \infty}{\sim} C \; n^{-\lambda} (\log(n))^{m-1}.$$

Moreover, the integral concentrates around the subset $W_0^{\mathrm{deg}} = [-1,1]^d \cap \bigcup_{\frac{1}{k_i} = \lambda} \{w_i = 0\} \subseteq W_0$ as $n \to \infty$.

Lessons from key examples

For those two real-analytic functions, there are two geometric invariants $\lambda \in \mathbb{Q}_{\geq 0}$ and $m \in \mathbb{N}$ of the singularities in W_0 which seem to control

- **(Vol)** the asymptotic volume of $B_F(\epsilon)$ as $\epsilon \to 0$.



(Image credit: Jesse Hoogland)

- **(Zeta)** the behaviour of the largest pole of the meromorphic zeta function $\zeta_F(s)$, (and in particular the integrability of $|F|^s$ on W)

- **(Laplace)** the asymptotic of the Laplace-type integral $Z_F(n)$ as $n \to \infty$.

Real log-canonical threshold

Definition

The real log-canonical threshold (rlct) of F is

$$\mathrm{rlct}(F) := \sup\{s \in \mathbb{R}_{\geq 0} \mid |F|^{-s} \textit{ is integrable}\} \in \mathbb{R}_{\geq 0} \cup \{\infty\}.$$

Alternative terminology: $\mathrm{rlct}(F)$ "critical integrability index".

Lemma

We have $\mathrm{rlct}(F) < \infty \Leftrightarrow W_0 \neq \emptyset$ and if $W_0 \neq \emptyset$, then $\mathrm{rlct}(F) \leq \frac{d}{2}$.

We assume $W_0 \neq \emptyset$ from now on. For our two key examples:

- $\mathrm{rlct}(w_1^2 + \ldots + w_d^2) = \frac{d}{2}$ (maximal possible)
- $\mathrm{rlct}(w_1^{k_1} \ldots w_d^{k_d}) = \min_i \left\{\frac{1}{k_i}\right\}$

Equivalent characterizations of rlct

Theorem (Arnold-Gusein-Zade-Varschenko)

Write $\lambda = \mathrm{rlct}(F)$. There exists $m = \mathrm{rlcm}(F) \in \mathbb{N}$ (real log-canonical multiplicity) such that

- **(Vol)** *The volume of sub-level sets of F satisfies*

$$\mathrm{Vol}\ B_F(\epsilon) \underset{\epsilon \to 0}{\sim} C\ \epsilon^{\lambda}\ (-\log(\epsilon))^{m-1}$$

- **(Zeta)** *The zeta function $\zeta_F(s)$ has meromorphic continuation to $\mathbb{C}$ with poles in $\mathbb{R}_{<0}$. The largest pole of $\zeta_F(s)$ is $-\lambda$ and has order m.*
- **(Laplace)** *The Laplace-type integral of F satisfies*

$$Z_F(n) = \int_W e^{-n|F(w)|} dw \underset{n \to \infty}{\sim} C\ n^{-\lambda} (\log(n))^{m-1}.$$

Local rlct and concentration of measure

Definition

Let $w \in W_0$. The local real log-canonical threshold of F at w is

$$\begin{aligned} \mathrm{rlct}_w(F) &:= \sup\{s \in \mathbb{R}_{\geq 0} \mid |F|^{-s} \text{ is locally integrable at } w\} \\ &= \mathrm{rlct}(F_{|U_w}) \end{aligned}$$

for U_w small enough open neighbourhood of w.

Proposition

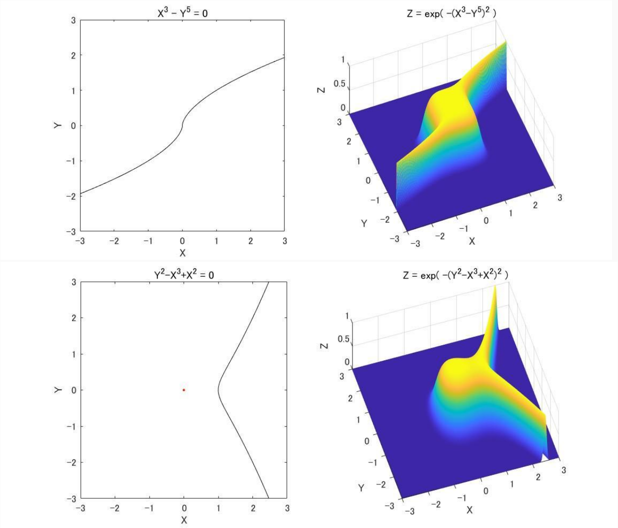
- *The global rlct is determined by the local ones:*

$$\mathrm{rlct}(F) = \min_{w \in W_0} \mathrm{rlct}_w(F)$$

- *The integrals* **(Vol)** *and* **(Laplace)** *concentrate asymptotically around the subset of "most degenerate" singularities:*

$$W_0^{\mathrm{deg}} := \{w \in W_0 \mid \mathrm{rlct}_w(F) = \mathrm{rlct}(F)\}.$$

Examples of concentration of Laplace integrals



(Image credit: S. Watanabe)

Smooth points and non-degenerate singularities

Smooth points: If $w \in W_0$ is not a singularity of F, then

$$\mathrm{rlct}_w(F) = 1 \text{ and } \mathrm{rlcm}_w(F) = 1.$$

Note that since $W_0 \neq \emptyset$ this never happens if $F \geq 0$.

Non-degenerate singularities/ Morse singularities:

Assume $F \geq 0$ and $w_0 \in W_0$ is such that the Hessian $\mathrm{Hess}_{w_0}(F)$ is positive-definite ($\Rightarrow w_0$ is an isolated point of W_0).

Classical Laplace approximation:

$$\int_{U_{w_0}} e^{-nF(w)} dw \underset{n\to\infty}{\sim} \sqrt{\frac{(2\pi)^d}{\det(\mathrm{Hess}_{w_0}(F))}}\; n^{-\frac{d}{2}}$$

hence $\mathrm{rlct}_{w_0}(F) = \frac{d}{2}$ (maximal possible) and $\mathrm{rlcm}_{w_0}(F) = 1$.

NB: One of very few cases with a simple formula for the constant C.

Key geometric input: Hironaka's resolution of singularities

By a fundamental theorem of Hironaka, there exists a real-analytic log-resolution of F: a proper real-analytic map $\pi : \widetilde{W} \to W$ from a real-analytic manifold $\widetilde{W}$ such that

- π induces a diffeomorphism

$$\widetilde{W} \setminus \pi^{-1}(W_0) \underset{\sim}{\xrightarrow{\pi}} W \setminus W_0.$$

- Locally on $\widetilde{W}$, the function $F \circ \pi$ is monomial:

$$F \circ \pi(\widetilde{w}) = G(\tilde{w})\tilde{w}_1^{k_1} \dots \tilde{w}_d^{k_d}$$

  with G non-vanishing and $k_i \in \mathbb{N}$.
- Locally on $\widetilde{W}$, the Jacobian determinant $|\mathrm{Jac}(\pi)|$ is monomial:

$$|\mathrm{Jac}(\pi)| = G'(\tilde{w})\tilde{w}_1^{h_1} \dots \tilde{w}_d^{h_d}$$

  with G' non-vanishing and $k_i \in \mathbb{N}$.

Real log-canonical threshold via resolution

Theorem (Arnold-Gusein-Zade-Varschenko)

(Res) *With the notation above, we have*

$$\operatorname{rlct}(F) = \min_{\textit{charts}} \min_{i} \left\{ \frac{h_i + 1}{k_i} \right\}$$

and

$$\operatorname{rlcm}(F) = \max_{\textit{charts}} \left(\# \left\{ i \mid \frac{h_i + 1}{k_i} = \operatorname{rlct}(F) \right\} \right)$$

Corollary

$\operatorname{rlct}(F) \in \mathbb{Q}$ *(!)*

Remarks

- *When* $F \in \mathbb{R}[w_1, \ldots, w_d]$*, a resolution can sometimes be computed explicitly* $\Rightarrow$ *formulas for* $\operatorname{rlct}(F)$*.*
- *The resolution* π *is non-canonical while* $\operatorname{rlct}(F)$ *is independent of the choice of* π*.*

Sketch of proof of theorems

Relevant for us because the main proofs of SLT build on this strategy:

- Fix a resolution $\pi : \widetilde{W} \to W$ and *define* (λ, m) as in **(Res)**.
- By base change formula we have (at least locally, choosing local coordinates):

$$\zeta_F(s) = \int_{\widetilde{W}} |F \circ \pi(\tilde{w})|^s \cdot |\mathrm{Jac}(\pi)(\tilde{w})| d\tilde{w}$$

- Locally, $F \circ \pi$ and $|\mathrm{Jac}|$ are monomial. Computation of ζ_F in monomial case + partition of unity argument ⇒ **(Zeta)**.
- Structure of poles of $\zeta_F(s)$ + inverse Mellin transform ⇒ asymptotic expansion of the "density of states"/Gelfand-Leray differential form.
- Integrate the density of states ⇒ **(Vol)**
- Take Laplace transform of the density of states ⇒ **(Laplace)**.

Some more examples

- **Morse-Bott singularities:** Assume that W_0 is a submanifold of dimension $d' \leq d$ and that $\mathrm{Ker}(\mathrm{Hess}_w(F)) = T_w W_0$ for all $w \in W_0$. Then
$$\mathrm{rlct}(F) = \frac{d'}{2} \text{ and } \mathrm{rlcm}(F) = 1.$$
- $F \in \mathbb{R}[w_1, \ldots, w_d]$ homogeneous of degree N with an isolated singularity at 0 and $F \geq 0$. Then
$$\mathrm{rlct}(F) = \frac{d}{N} \text{ and } \mathrm{rlcm}(F) = 1.$$
- $F(x, y) = x^2 y^2 (x + y)^2$. Then
$$\mathrm{rlct}(F) = \frac{1}{3} \text{ and } \mathrm{rlcm}(F) = 3.$$
- $F(x, y) = x^2 - y^3$. Then
$$\mathrm{rlct}(F) = \frac{5}{6} \text{ and } \mathrm{rlcm}(F) = 1.$$

Wider context in singularity theory

- Parallel invariant in complex-analytic and algebraic geometry, the log-canonical threshold $\mathrm{lct}(F)$.
- In fact, if F is complex-analytic, we can also consider the associated real-analytic function $|F|^2$ and then $\mathrm{lct}(F) = \mathrm{rlct}(|F|^2)$.
- The lct has been intensively studied and is connected to central topics in singularity theory and birational geometry:
  - * Canonical singularities (hence the name)
  - * Multiplier ideals
  - * $\mathcal{D}$-modules, Bernstein-Sato polynomial and V-filtration
  - * Spectrum of the monodromy and mixed Hodge structure on vanishing cycles
  - * Irreducible components of jet schemes of F
  - * Poles of motivic and p-adic local zeta functions of F
  - * Singularities in positive characteristics and Frob-pure threshold.
- Q: Which aspects of this story extend to the rlct? Which are relevant to statistics?

Summary of real-log canonical thresholds

- We have associated to a real-analytic function $F : W \to \mathbb{R}$ two invariants

  $$\mathrm{rlct}(F) \in \mathbb{Q}_{\geq 0} \text{ and } \mathrm{rlcm}(F) \in \mathbb{N}$$

  which control the analytic behaviour of F around $W_0 = F^{-1}(0)$:

  - * Volume scaling of sublevel sets
  - * Poles of zeta function
  - * Laplace-type integral (most relevant for SLT)

- These quantities are local, we have

  $$\mathrm{rlct}(F) = \min_{w \in W_0} \mathrm{rlct}_w(F)$$

  and the integrals in **(Vol)** and **(Laplace)** concentrate asymptotically around the most degenerate singularities W_0^{deg} of F.

Classical Bayesian statistics

Overview

Warning: I am not a statistician, so interpret everything from now on as a noisy sample of some underlying true mathematical facts...

- **What is statistics about?**
- **Set-up**
- **Regular and singular models**
- **Bayes comes in: prior and posterior distributions, Partition function, free energy**
- **Predictive distribution, Bayesian generalization error**

~~What is statistics about?~~ What is statistics about in this talk?

- In statistics, we collect/are given data $D_n = \{x_1, \ldots, x_n\}$ produced by some unknown, noisy data-generating process.

- We want to infer information about the data-generating process from D_n.

- To formalize this mathematically, we treat D_n as samples from a data-generating probability distribution $q(x)$.

- We then try to approximate $q(x)$ by members of a well-chosen parametric statistical model $\{p(x|w)\}_{w \in W}$, i.e. a family of probability distributions parametrized by $w \in W \subseteq \mathbb{R}^d$.

- In Bayesian statistics, we also give ourselves a prior distribution $\phi(w)$ on W which reflects our a priori belief in how well $p(x|w)$ approximates $q(x)$.

Cast of characters

Bayesian statistical model, fully specified:

- $q(x)$ data-generating distribution
- $D_n = \{x_1, \ldots, x_n\}$ dataset sampled from $q(x)$
- $\{p(x|w) \mid w \in W \subset \mathbb{R}^d\}$ parametric statistical model
- $\phi(w)$ prior distribution

(Some) derived quantities:

- Relative entropy $K(w) = K(q(x) \| p(x|w))$
- Empirical relative entropy $K_n(w)$
- Posterior distribution $p(w|D_n)$
- Normalized partition function $\overline{Z}_n$ and normalized free energy $\overline{F}_n$
- Predictive distribution $p(x|D_n)$
- Bayesian generalization error $B_n = K(q(x) \| p(x|D_n)$

Set-up

- Dataset $D_n = \{x_1, \dots, x_n\}$ with each $x_i \in X = \mathbb{R}^N$.
- We assume that the x_i are sampled independently from the same data-generating distribution

$$q(x) \in \mathcal{P}(X) := \left\{ p \in L^1(X, m) \mid p \geq 0, \int p dm = 1. \right\}$$

- We are given $W \subset \mathbb{R}^d$ as in the previous section and a parametric statistical model $W \to \mathcal{P}(X), w \mapsto p(-|w)$.

Key assumption of SLT

The distribution $p(-|w) \in \mathcal{P}(X)$ is real-analytic in w.
(i.e., can be locally written as convergent power series in w with coefficients in $L^1(X, m)$)

This holds for most statistical/machine learning models.

Regression models / Supervised learning

- Suppose our data consists of pairs $(x_i, y_i) \in X \times Y = \mathbb{R}^N \times \mathbb{R}^M$ and we expect a functional relationship

  $$y_i = f(x_i) + \text{ Gaussian noise } .$$

  Assume moreover that we have many other examples of x_j (without the corresponding y_j) and we want to predict y_j given x_j.
- We interpret this as $q(x, y) = q(y|x)p(x)$ distribution on $X \times Y$ with $p(x)$ known empirically (from the inputs of our regression) and:

  $$q(y|x) \propto \exp\left(-\frac{\|y - f(x)\|^2}{\sigma^2}\right)$$

  for an unknown function $f : X \to Y$. We then put

  $$p(y|x, w) := p(y|x, w)p(x)$$

  with

  $$p(y|x, w) \propto \exp\left(-\frac{\|y - f_w(x)\|^2}{\sigma^2}\right)$$

  for some parametrized real-analytic function $f_w : W \times X \to Y$.

Example: neural networks as regression models

- The function $f_w : W \times \mathbb{R}^N \to \mathbb{R}^M$ in regression models can be as simple or as complicated as appropriate for the statistical problem at hand. A classical choice is linear regression where each map f_w is affine and $W \subseteq \mathrm{Mat}_{M,N}(\mathbb{R}) \times \mathbb{R}^M$.
- Popular choice these days: f_w function computed by a neural network with weights $w \in W$. We describe the simplest architecture: feed-forward fully connected networks (or multi-layer perceptrons) with no biases.
- Fix a depth L and dimensions $N_1 = N,\ N_2, \ldots,\ N_{L+1} = M$. Fix an activation function $\alpha : \mathbb{R} \to \mathbb{R}$. We put $W = \prod_{i=1}^{L} \mathrm{Mat}_{N_i,N_{i+1}}(\mathbb{R})$. For $w = (A_1, \ldots, A_L)$, we define

$$f_w(x) := A_L \cdot \alpha(A_{L-1} \cdot \alpha(\ldots A_1 \cdot x) \ldots)$$

where α is applied to a vector coordinate-wise.

Relative entropy

We want to measure the “distance” between our model and the data-generating distribution. We use an information-theoretic notion whose statistical meaning will become clear later:

Definition

Let $p(x), p'(x) \in \mathcal{P}(X)$ with same support. Their relative entropy (or Kullback-Leibler divergence) is

$$K(p'(x) \| p(x)) := \int_X p'(x) log \frac{p'(x)}{p(x)} dx \in \mathbb{R} \cup \{\infty\}.$$

Warning: $K(-\|-)$ is not symmetric in p and p' and does not satisfy a triangle inequality.

Lemma (Gibbs’ inequality (corollary of Jensen))

We have $K(p'\|p) \geq 0$ and $K(p'\|p) = 0 \Leftrightarrow p \overset{a.e.}{=} p'$.

Relative entropy

From now on we assume $q(x)$ and $p(x|w)$ have the same support (e.g. all of X) for every $w \in W$ and define our main player, the relative entropy between the model and the true distribution:

$$K(w) := K(q(x)\|p(x|w)) = \int_X q(x) \log \frac{q(x)}{p(x|w)} dx.$$

Under mild convergence assumptions:

$$p(-|w) \text{ real-analytic on W} \Rightarrow K(w) \text{ real-analytic on W}.$$

Example

For a regression model as above,

$$K(w) = \frac{1}{2} \int_{\mathbb{R}^N} \|f(x) - f_w(x)\|^2 p(x) dx = \frac{1}{2} \|f - f_w\|^2_{L^2(\mathbb{R}^N, \mathbb{R}^M, p(x))}$$

measures the mean square error between f_w and f.

Empirical relative entropy

So far we have not used our dataset D_n at all!

Definition

The empirical relative entropy $K_n(w)$ is defined as

$$K_n(w) := \frac{1}{n}\sum_{i=1}^{n} \log(q(x_i)/p(x_i|w)).$$

$K_n(w)$ measures "how unlikely it is that D_n was sampled from $p(x|w)$ instead of the true distribution $q(x)$."

Lemma

- *We have $\mathbb{E}_q[K_n(w)] = K(w)$.*
- **(Law of large numbers)** *Given $w \in W$, we have $K_n(w) \underset{n\to\infty}{\to} K(w)$.*

Key idea of Watanabe: use geometry and singularity theory to control the fluctuations of K_n around its expectation K.

Realizable and unrealizable models

Definition

A pair $(q(x), p(x|w))$ is realizable if there exists $w \in W$ such that $q(x) \overset{a.e}{=} p(x|w)$ and unrealizable otherwise.

Define

$$W_0 = K^{-1}(0) = \{w \in W \mid q(x) \overset{a.e}{=} p(x|w)\}$$

so that $(q(x), p(x|w))$ is realizable if and only if $W_0 \neq \emptyset$.

Realizability

For simplicity, in the rest of the talk, we assume $W_0 \neq \emptyset$.

SLT for unrealizable models will have to wait for tomorrow.

Regular and singular models

Definition

A (realizable) pair $(q(x), p(x|w))$ *is a regular model if*

- $W_0 = \{w_0\}$ *consists of a single point*
- $\mathrm{Hess}_{w_0}(K)$ *is positive definite*

In other words, K *admits a unique non-degenerate singularity.*

A model which is not regular is called singular.

Example

Any model which admits (discrete or continuous) symmetries is singular. Modern machine learning models like neural networks are almost always singular.

Prior and posterior distributions

- Up until now, nothing particularly “Bayesian” about this story.
- In Bayesian statistics, we start with a prior distribution $\phi(w) \in \mathcal{P}(W)$ encoding our initial belief about the accuracy of the model for different parameters w.
- We then want to find the posterior distribution $p(w|D_n)$ of our “updated belief after seeing the data”. By Bayes rule:

$$\begin{aligned} p(w|D_n) &\overset{Bayes}{=} \frac{p(D_n|w)\phi(w)}{p(D_n)} \\ &\overset{D_n\ i.i.d}{=} \frac{\left(\prod_{i=1}^n p(x_i|w)\right)\phi(w)}{\int_W \left(\prod_{i=1}^n p(x_i|w)\right)\phi(w)dw} \end{aligned}$$

Normalized partition function and free energy

Definition

The normalized partition function (or normalized marginal likelihood) is

$$\overline{Z}_n = \int_W \phi(w) \prod_{i=1}^{n} \frac{p(x_i|w)}{q(x_i)} dw = \int_W \exp(-nK_n(w))\phi(w) dw$$

The normalized free energy is

$$\overline{F}_n = -\log \overline{Z}_n.$$

We can rewrite the posterior distribution in terms of $\overline{Z}_n$ as

$$p(w|D_n) = \frac{1}{\overline{Z}_n} \exp(-nK_n(w))\phi(w)$$

The role of Laplace-type integrals is finally becoming apparent! (the factor $\phi(w)$ does not change asymptotics as long as $\phi_{|W_0} > 0$)

Statistical interpretation of $\overline{F}_n$

Write

$$q(D_n) = \prod_{i=1}^{n} q(x_i)$$

and

$$p(D_n) = \int_W \phi(w) \prod_{i=1}^{n} p(x_i|w) dw.$$

Both expressions define probability distributions $q(D_n), p(D_n) \in \mathcal{P}(X^n)$ on the product space X^n of datasets.

Lemma

$$\mathbb{E}_q[\overline{F}_n] = K(q(D_n) \| p(D_n)).$$

In other words, in expectation, $\overline{F}_n$ measures how much the distribution of possible datasets D_n predicted by averaging over the model differs from the true distribution of D_n according to q.

Predictive distribution and generalization error

In the Bayesian framework, we use the posterior distribution to make predictions by averaging over the model.

Definition

The predictive distribution $p(x|D_n) \in \mathcal{P}(X)$ *is defined as*

$$p(x|D_n) := \int_W p(x|w)p(w|D_n)dw = \frac{1}{\overline{Z}_n}\int_W p(x|w)exp(-nK_n(w))\phi(w)dw$$

I.e. "we believe that a new datapoint $x_{n+1} \sim q(x)$ *will come up with probability* $p(x_{n+1}|D_n)$*".*

How accurate is this procedure? Again we turn to relative entropy:

Definition

The Bayesian generalization error B_n *is the relative entropy*

$$B_n := K(q(x)\|p(x|D_n))$$

between data-generating and predictive distributions.

Normalized free energy vs Bayesian generalization

$\overline{F}_n$ and B_n are both measures of the quality of the model, but measure slightly different things. They are connected by

Lemma (B_n "discrete derivative" of $\overline{F}_n$ on average)

$$\mathbb{E}_q[B_n] = \mathbb{E}_q[\overline{F}_{n+1}] - \mathbb{E}_q[\overline{F}_n].$$

Both measures can be used for Bayesian model selection, i.e. comparing two different models on the same data.

Summary of Bayesian statistics

Full statistical model:

- $q(x)$ data-generating distribution
- $D_n = \{x_1, \ldots, x_n\}$ dataset sampled from $q(x)$
- $\{p(x|w) \mid w \in W\}$ parametric statistical model
- $\phi(w)$ prior distribution

Derived quantities:

- Relative entropy $K(w) = K(q(x) \| p(x|w))$
- Empirical relative entropy $K_n(w)$
- Posterior distribution $p(w|D_n) = \frac{1}{\overline{Z}_n} \exp(-nK_n(w))\phi(w)dw$
- Normalized partition function $\overline{Z}_n = \int_W \exp(-nK_n(w))\phi(w)dw$ and free energy $\overline{F}_n = -\log(\overline{Z}_n)$
- Predictive distribution $p(x|D_n) = \int_W p(x|w)p(w|D_n)dw$
- Bayesian generalization error $B_n = K(q(x) \| p(x|D_n))$

Statistical learning theory of regular and singular models

Overview

- **Stastical learning theory**
- **Posterior concentration**
- **Regular models**
- **Singular learning theory**

Statistical learning theory

Heuristically, we expect as the number n of data samples goes to ∞:

- The model will predict better and better.
  $\Rightarrow$ B_n will decrease.
- The model will become more confident.
  $\Rightarrow$ $p(w|D_n)$ will concentrate around W_0.
  $\Rightarrow$ Asymptotic behaviour of $\overline{Z}_n$ and $\overline{F}_n$

The two phenomena are connected by

$$\mathbb{E}_q[B_n] = \mathbb{E}_q[\overline{F}_{n+1}] - \mathbb{E}_q[\overline{F}_n].$$

The goal of (Bayesian) statistical learning theory is to quantify this.

Posterior consistency and large deviations principle

Beautiful (and classical) connection between Bayesian statistics and information theory which clarifies the statistical role of K.

Theorem

(Under mild technical assumptions, doesn't require analytic)
The posterior distribution $p(w|D_n)$ concentrates exponentially quickly in n with rate function the relative entropy $K(w)$: for any measurable $U \subset W$, we have very roughly:

$$\int_U p(w|D_n)dw \approx C \ \exp(-n \min_{u \in U} K(u))$$

or more precisely (but still slightly inaccurately):

$$-\frac{1}{n} \log \int_U p(w|D_n)dw \underset{n \to \infty}{\to} \min_{u \in U} K(u).$$

Q: Can we refine the asymptotic around W_0?

Statistical learning theory of regular models

Theorem

(Under mild technical assumptions, doesn't require analytic)

Assume that $(q(x), p(x|w))$ *is regular with* $W_0 = \{w_0\}$ *and* $\phi(w_0) > 0$.

- **Free energy formula:**

$$\mathbb{E}_q[\overline{F}_n] \underset{n\to\infty}{=} \frac{d}{2}\log(n) - \log(C) + o(1)$$

with

$$C = \frac{(2\pi)^{\frac{d}{2}} \exp(\frac{d}{2})\phi(w_0)}{\sqrt{\det(\mathrm{Hess}_{w_0}(K))}}$$

- **Asymptotic Bayesian generalization:**

$$\mathbb{E}_q[B_n] \underset{n\to\infty}{=} \frac{d}{2n} + o(1/n).$$

Asymptotic theory of regular models

More precise result about the concentration of the posterior distribution itself:

Theorem (Bernstein-Von Mises)

(Under some technical assumptions) Assume that $(q(x), p(x|w))$ is regular with $W_0 = \{w_0\}$. Then the posterior distribution $p(w|D_n)$ is asymptotically normal around the maximum likelihood estimator:

If we fix for each n any minimum w_n^* of $K_n(w)$ (MLE), then for any measurable $U \subseteq W$, we have

$$\int_U |p(w|D_n) - \mathcal{N}(w_n^*, n^{-1}\mathrm{Hess}_{w_0}(K)^{-1})| dw \underset{n\to\infty}{\to} 0.$$

with $\mathcal{N}(w^*, I)$ the normal distribution on $\mathbb{R}^d$ with mean w^* and covariance matrix I.

Singular learning theory

Theorem (Watanabe)

(Under some technical assumptions) Assume that $(q(x), p(x|w))$ is a realizable real-analytic model and that $\phi_{|W_0} > 0$. Write $\lambda = \mathrm{rlct}(K)$ and $m = \mathrm{rlcm}(K)$. Then the posterior distribution concentrates around W_0^{deg} and:

- **Free energy formula:**

$$\mathbb{E}_q[\overline{F}_n] \underset{n\to\infty}{=} \lambda \log n - (m-1) \log\log n + O(1)$$

- **Asymptotic Bayesian generalization:**

$$\mathbb{E}_q[B_n] \underset{n\to\infty}{=} \frac{\lambda}{n} + o(1/n).$$

Watanabe also describes the constant C and the asymptotic shape of the posterior distribution but it is much more complicated to state than in the regular case (no asymptotic normality!) and depends on a resolution.

Sketch of proof of Watanabe's theorem

- Fix a log-resolution $\pi : \widetilde{W} \to W$. Using a fancy functional version of the central limit theorem (empirical process theory), Watanabe shows that

$$\xi_n(\tilde{w}) := \frac{1}{\sqrt{n}} \frac{K(\pi(\tilde{w})) - K_n(\pi(\tilde{w}))}{\sqrt{K(\pi(\tilde{w}))}}$$

  is well-defined and converges when $n \to \infty$ to a Gaussian process $\xi(\tilde{w})$ on $\tilde{w}$ (small lie here). We thus have

$$K_n(\pi(\tilde{w})) = K(\pi(\tilde{w})) - \frac{1}{\sqrt{n}} \sqrt{K(\pi(\tilde{w}))} \xi_n(\tilde{w})$$

  with some probabilistic control over the fluctuations ξ_n as $n \to \infty$.
- Then Watanabe plugs this formula into the proof by Arnold-Gusein-Zade-Varschenko of asymptotics of Laplace integrals (zeta function, inverse Mellin transform, density of states...), with a lot of additional work to deal with the fluctuations.

First statistical consequences

$$\mathbb{E}_g[B_n] = \frac{\lambda}{n} + o(\frac{1}{n}) \quad \text{with} \quad \lambda = \mathrm{rlct}(K)$$

- Precise formulation of the slogan we started the talk with.
  Watanabe also calls λ the learning coefficient of the model.

- By comparing with the regular case, one way we can think about λ is as (half) an effective parameter count/complexity measure.

- Since $\lambda \leq \frac{d}{2}$, singular models generalize better!

- In the regular case $\mathbb{E}_g[B_n] = \frac{d}{2n} + o(\frac{1}{n})$
  $\Rightarrow$ Regular models with more parameters generalize worse.
  $\Rightarrow$ Overparametrization hurts generalization.

- This is not (necessarily) true for singular models! Important because empirically overparametrized ML models can generalize very well.

Summary of statistical learning theory

- The Bayesian posterior distribution of a (realizable) model "always" concentrates around W_0 exponentially fast with rate given by the relative entropy $K(w) = K(q(x)\|p(x|w))$.
- For regular models, the posterior is asymptotically normal around the unique minimum and we have

  $$\mathbb{E}_q[B_n] \underset{n\to\infty}{=} \frac{d}{2n} + o(\frac{1}{n}).$$

  and

  $$\mathbb{E}_q[\overline{F}_n] \underset{n\to\infty}{=} \frac{d}{2}\log n - \log(C) + o(1)$$

- For general singular models, the posterior concentrates around the most degenerate singularities W_0^{deg} and we have

  $$\mathbb{E}_q[B_n] \underset{n\to\infty}{=} \frac{\lambda}{n} + o(1/n).$$

  and

  $$\mathbb{E}_q[\overline{F}_n] \underset{n\to\infty}{=} \lambda \log n - (m-1)\log\log n + O(1).$$

SLT from theory to practice

Applications of SLT in Bayesian statistics

In the Bayesian context, SLT has been applied and extended by Watanabe and his collaborators and students in many directions:

- Other more practical asymptotic results in statistical learning theory (e.g. maximum likelihood estimation, Gibbs error, cross-validation error)
- Information criteria and Bayesian model selection
- Analysis of certain MCMC algorithms
- ...

See [Watanabe09], [Watanabe18], [Watanabe_survey] and the many references therein. I will mention a few of these works later but there are many more!

A few more definitions

In our theoretical analysis we used liberally the data-generating distribution $q(x)$. Since we do not have access to it in practice (that's the whole point!), we also need to study unnormalized quantities which only depend on the known: data, model, prior.

- The negative log-likelihood $L_n(w)$ is

$$L_n(w) = -\frac{1}{n}\sum_{i=1}^{n} \log p(x_i|w).$$

- The partition function Z_n is

$$Z_n = \int_W \phi(w) \prod_{i=1}^{n} p(x_i|w) dw = \int_W \exp(-nL_n(w))\phi(w) dw$$

- the free energy F_n is

$$F_n = -\log(Z_n)$$

Free energy formula, final version

Watanabe's approach actually provides an asymptotic formula for the random variable F_n and not just its expectation:

Theorem (Watanabe)

(Under some technical assumptions) Assume that $(q(x), p(x|w))$ is a realizable real-analytic model and that $\phi_{|W_0} > 0$. Let $w_0 \in W_0$. Write $\lambda = \mathrm{rlct}(K)$ and $m = \mathrm{rlcm}(K)$.

$$F_n \underset{n\to\infty}{=} nL_n(w_0) + \lambda \log n - (m-1)\log\log n + O_p(1).$$

Statistical intrepretation

Model selection by minimizing the free energy F_n involves trade-off between

- Model accuracy on the data $nL_n(w_0)$
- Model complexity/degeneracy λ

SLT for unrealizable models

- Assume that $(q(x), p(x|w))$ is not realizable. We do the simplest possible correction and just shift the function to take the value 0:

$$K(w) := K(q(x)\|p(x|w)) - \min_{v} K(q(x)\|p(x|v))$$

  and try to proceed as in the realizable case.

- The asymptotic formulas for the free energy and the generalization error do not hold in general! $\Rightarrow$ we need additional assumptions.

- Watanabe shows in [Watanabe18] how to extend SLT under the assumption of relatively finite variance, which I won't discuss here.

- [Nagayasu-Watanabe] explores SLT beyond relatively finite variance; still a lot to do!

Non-analytic models

- Q: SLT as stated above applies to real-analytic models; what happens beyond this?
- Non-analytic models come up in practice in (at least) two important ways: neural networks with non-analytic activation functions such as rectified linear units (ReLU), and mixture models.
- By definition $\alpha_{ReLU}(x) := \max(x, 0)$ so ReLU neural networks parametrize piecewise linear functions (!) and their relative entropy function K is only piecewise analytic.
- **Expensive potential solution:** redevelop the whole theory in the context of sub-analytic geometry. Technically daunting, but the basic ingredients seem to be available?
- **Practical solution:** table the issue and pretend everything is (close enough to) real-analytic; seems to work well empirically!

Asymptotic versus finite *n*

- Q: SLT gives asymptotic formulas; what information do we get for practical dataset sizes *n*?
- Common issue in mathematical statistics! Modern trend seems for this reason to lean towards non-asymptotic results and techniques.
- SLT corrects the finite *n* behaviour of the classical theory: if you take a regular model which is "close to a singular model" and apply the regular asymptotic theory, the results are only meaningful for very large *n*, while SLT applied to the nearby singular model gives meaningful results for much smaller *n* (cf. [Watanabe18, §1.5] for examples and further discussion)
- **Theoretical solution:** Watanabe's proofs actually produce bounds for every *n*. As far as I know they haven't been applied so far.
- **Empirical solution:** Evaluate SLT's claims on small models and show that they work for reasonable *n*. (cf. [TMS] for a thourough examples).

What can we compute? Learning coefficient

- Thanks to Watanabe and his collaborators, the learning coefficient $\lambda = \mathrm{rlct}(K)$ is known in a number of “simple” cases:
  - * Various mixture models (normal, Poisson, multinomial...)
  - * Hidden Markov models
  - * latent Dirichlet allocation
  - * Deep linear networks (tour de force by M. Aoyagi [Aoyagi-DLN])
  - * $\ldots$
- **Fundamental observation:** λ depends on the (unknown) data-generating distribution $q(x)$!
- From a theoretical point of view this is a feature: the same model generalizes differently for different $q(x)$, and SLT tells us how.
- From a practical point of view, this means that we can almost never know the exact value of λ.

What can we compute? Free energy and posterior sampling

- In Bayesian statistics, F_n (or Z_n) is known to be difficult to estimate.
- Unfortunate, because we only know the posterior distribution *up to* the partition function:

$$p(w|D_n) = \frac{1}{Z_n} \exp(-nL_n(w))\phi(w)$$

- Markov Chain Monte Carlo (MCMC) algorithms construct approximate samples from a probability distribution which is only known up to a constant.
  ⇒ we can sample from the posterior distribution.
  ⇒ MCMC workhorse of pratical Bayesian inference.
- MCMC does not scale up well to very large models. Various approximations are used, e.g. Stochastic gradient Langevin dynamics (SGLD). No time to discuss here, but fundamental to applications below.

Widely applicable Bayesian information criterion

Can we use MCMC/SGLD sampling to define a computationally tractable estimator of F_n? Watanabe [WBIC] to the rescue!

Definition

$$WBIC_n := \mathbb{E}_w^{\frac{1}{\log(n)}}[nL_n(w)]$$

where the notation $\mathbb{E}_w^\beta$ indicates expectation with respect to the tempered Bayesian posterior at inverse temperature $\beta > 0$:

$$p^\beta(w|D_n) = \frac{\prod_{i=1}^n p(x_i|w)^\beta \phi(w)}{\int_W \prod_{i=1}^n p(x_i|w)^\beta \phi(w) dw}$$

- Samples from the tempered posterior distribution can also be obtained by MCMC/SGLD sampling. Samples can be used to estimate expectations ⇒ $WBIC_n$ is computationally tractable.
- The terminology "Bayesian Information criterion" comes from context of model selection.

Estimating the free energy and the rlct via WBIC

Theorem (Watanabe)

Under technical assumptions + relative finite variance

$$F_n = WBIC_n + O_p(\sqrt{\log n})$$

and hence by combining with the free energy formula

$$WBIC_n = nL_n(w_0) + \lambda \log n + O_p(\sqrt{\log n}).$$

Corollary

$$\hat{\lambda} := \frac{WBIC_n - nL_n(w_n^*)}{\log n}$$

(with $\hat{w}_n^*$ argmin of L_n over samples) is an estimator of the learning coefficient λ.

We can estimate real log-canonical thresholds from data! but not rlcm ;-(

Local Bayesian statistics

- We cannot directly apply the tools of SLT, even with WBIC, to a large model such as a deep neural network with millions of parameters; computationally intractable!

- Moreover, it is not even clear that we really want to do this, since in deep learning we want to understand the training process of optimization algorithms such as stochastic gradient descent which only explore a *very* small part of W.

- **Solution:** apply SLT to the models obtained by "localizing" the large model to small parts of W!

- Interesting story of free energy minimization and phase transitions in the Bayesian posterior which I have to omit (see [TMS])

Local learning coefficient estimator

In [LLC] these ideas are synthesized into

Definition (Lau,Murfet,Wei)

- *The tempered local posterior* $p^{\beta}(w|D_n, w^*, \gamma)$ *at* $w^* \in W$ *with confinement strength* $\gamma > 0$ *and inverse temperature* β *is*

$$p^{\beta}(w|D_n, w^*, \gamma) \propto \exp(-\frac{\gamma}{2}\|w - w^*\|^2)\exp(-n\beta L_n(w))\phi(w)$$

(multiplying the tempered posterior with a confining Gaussian prior)

- *The local learning coefficient estimator* $\hat{\lambda}(w^*)$ *is*

$$\hat{\lambda}(w^*) = \frac{\mathbb{E}^{\frac{1}{\log(n)}}_{w|w^*,\gamma}(nL_n(w)) - nL_n(\hat{w}^*_n)}{\log n}$$

with the expectation taken with respect to the local tempered posterior $p^{\beta}(w|D_n, w^*, \gamma)$ *and* $\hat{w}^*_n$ *argmin of* $L_n(w)$ *over samples.*

LLC estimation and local rlcts

- The confinement Gaussian prior ensures that the local posterior is tightly concentrated around w^*; tuning its value is important for numerical stability but does not affect (too much) the results.
- In a large model, sampling from the local (tempered) posterior distribution is much much easier than sampling from the full posterior distribution! We only need to sample the neighbourhood of $w^* \in W$, which is computationally tractable (especially with SGLD).
- When $w^* \in W_0$, Watanbe's WBIC theorem implies that $\hat{\lambda}(w^*)$ is an estimator of the *local* real log-canonical threshold $\mathrm{rlct}_{w^*}(K)$ from data!
- Mystery: $\hat{\lambda}(w^*)$ "behaves well" away from W_0, and seems to give a consistent local complexity measure everywhere!
- Q: What is $\hat{\lambda}(w^*)$ measuring geometrically in general?

Also... it works!

- For deep linear networks, we know the true value of λ by the theorem of M. Aoyagi [Aoyagi-DLN].
- In [LLC-scale], Z. Furman and E. Lau compared $\hat{\lambda}(w^*)$ for $w^* \in W_0$ with this theoretical baseline:

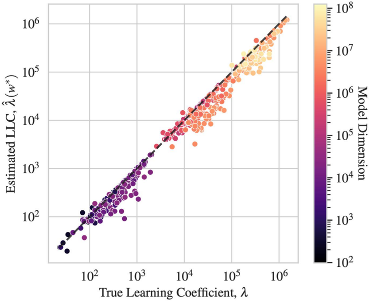


Figure 1. Estimated local learning coefficient against true learning coefficient; model dimension shown in color. The estimated LLC accurately measures the learning coefficient λ up to 100 million parameters in deep linear networks, as compared to known theoretical values (dashed line). See Figure 10 for linear-scale plots.

Machine learning and Developmental interpretability

Maximum likelihood estimation

- Note that we have $K_n(w) = L_n(w) - S_n$ with

$$S_n = -\frac{1}{n}\sum_{i=1}^{n} \log(q(x_i))$$

empirical entropy of the data independent of w, hence

- Minimizing $K_n(w) \Leftrightarrow$ minimizing $L_n(w)$, i.e. maximum likelihood estimation (MLE).
- For a regression model with dataset $D_n = \{(x_i, y_i)\}_{i=1}^{n}$, we have

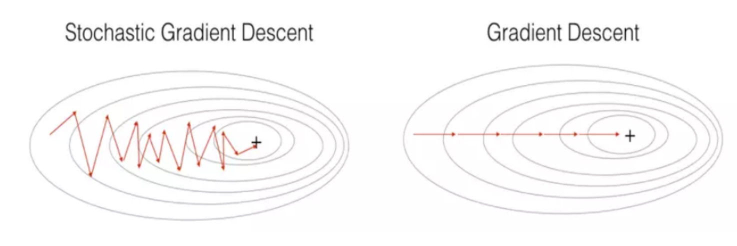
$$L_n(w) = \frac{1}{n}\sum_{i=1}^{n} \|y_i - f_w(x_i)\|^2 + cst$$

hence MLE $\Leftrightarrow$ minimizing mean square error loss function.

Stochastic gradient descent

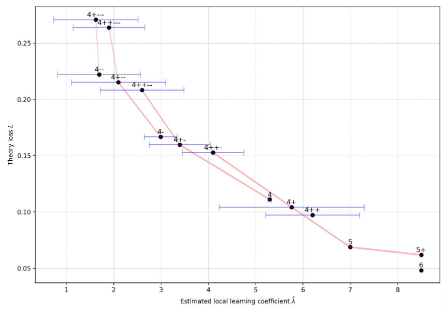
- In machine learning (esp. deep learning), minimizing the loss function is typically done by stochastic first order optimization algorithms like stochastic gradient descent (SGD).

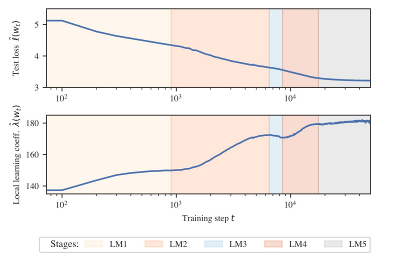
- Stochastic gradient descent *noisy, discretized gradient flow*:
  - * Initialize your model at some $w_{(0)} \in W$.
  - * Compute the gradient $\vec{v}_{(0)}$ at $w_{(0)}$ of minibatch loss function $\frac{1}{b}\sum_{i \in B} \|y_i - f_w(x_i)\|^2$ for a random minibatch $B \subset D_n$ of fixed size b.
  - * Set $w_{(1)} := w_{(0)} + \eta \vec{v}^{(0)}$ (with η fixed step-size and repeat (until...?)



Developmental Interpretability

- Goal: understanding the development of computational stucture during the training (by SGD) of a neural network, by tracking the changes in the local geometry/singularity theory of the loss function.
- The local learning coefficient estimator $\hat{\lambda}$ is a first tool for that purpose.
- Papers [TMS] ("toy model", very well understood from SLT picture) [ICL] (language model, towards more realistic applications)


Theory loss L
Estimated local learning coefficient $\hat{\lambda}$


Test loss $\hat{\ell}(w_t)$
10^2
10^3
10^4
Local learning coeff. $\hat{\lambda}(w_t)$
10^2
10^3
10^4
Training step t
Stages:
LM1
LM2
LM3
LM4
LM5