MACHINE LEARNING, SINGULARITY THEORY AND PHASE TRANSITIONS
UNIVERSITY OF AMSTERDAM, 18-21 SEPTEMBER 2023

Abstract

This is the program of an informal workshop organised at the University of Amsterdam around *Singular learning theory* and its burgeoning applications to the training dynamics of deep neural networks. This is a topic at the interface of machine learning, statistics and singularity theory in real-analytic geometry, so we expect a diverse audience, and the talks of the first day will provide some background from each discipline. Singular learning theory is an application of the tools of singularity theory to Bayesian statistics, and is a promising framework to understand some of the mysteries of modern deep learning. Our goal will be to understand the recent paper Quantifying degeneracy in singular models via the learning coefficient [4], both at a theoretical and practical level, and discuss work-in-progress on other deep learning systems.

Practical details

- The workshop will take place from Monday 18th to Thursday 21st, from 10:00 to 17:00.
- The rooms are still in the process of being reserved and will be announced here shortly.
- If you are not a staff member of the University, you will not be able to access the building. Please send me an email and I will send you my mobile phone number so that we can get you in.

Overview of the program

- Monday 18st: Learning from data, singularity theory
  - Talk 1 (10:00-11:00): Machine learning, deep learning, mysteries
  - Talk 2 (11:00-12:00): Bayesian statistics and statistical learning theory
  - Talk 3 (13:00-14:00): Regular and singular models
  - Q& A, discussion, break (14:00-15:00)
  - Talk 4 (15:00-16:00) Analytic geometry and singularities
  - Talk 5 (16:00-17:00) Measuring degeneracy with the real log-canonical threshold
- Tuesday 19st: Singular learning theory - core results
  - Q& A and discussion (10:00-11:00)
  - Talk 1 (11:00-12:00) Main theorem I: standard form of the log-likelihood function
  - Talk 2 (13:00-14:00) Main theorem II: free energy formula
  - Talk 3 (14:15-15:15) Main theorem III: consequences for generalization errors, singular fluctuation
  - Q& A, discussion, work in small groups (15:15-17:00)
- Wednesday 20th : Singular learning theory - consequences for phase transitions
  - Talk 1 (10:00-11:00) internal model selection and Bayesian phase transitions
  - Discussion (11:00-12:00) Larger picture of phase transitions in deep learning.
  - Talk 2 (13:00-14:00) Computing RLCTs with algebraic geometry
  - Talk 3 (14:15-15:15) Widely applicable Bayesian Information Criterion (WBIC)
  - Q& A, discussion, work in small groups (15:15-17:00)
- Thursday 21st: Singular learning practice - estimating learning coefficients and applications to phase transitions.
  - Talk 1 (10:00-11:00) "Quantifying degeneracy in singular models via the learning coefficient "
  - Talk 2 (11:00-12:00) Computing $\widehat{\lambda}$ in practice and applications
  - Talk 3 (13:00-14:00) DIY $\widehat{\lambda}$: libraries, internals, APIs
  - Q& A, discussion, work in small groups: (14:15-17:00)

Monday 18th: Learning from data, singularity theory

Talk 1 (10:00-11:00) Machine learning, deep learning, mysteries.

- Discuss the basic framework of (supervised) machine learning ML: parametrized models, classifiers with loss functions, train and test datasets, etc.
- Introduce simple feed-forward neural networks, discuss tanh and ReLU activation functions.
- Briefly discuss how they are trained in practice: first-order optimization, stochastic gradient descent (SGD) - we will come back to this later on.
- Give an impressionistic overview of how successful deep learning has been in many domains.
- Explain some of the basic mysteries: why are such models trainable? Why do the trained models generalize so well and tend not to overfit? Can we understand the computational structure of trained models and how it forms during training?

Talk 2 (11:00-12:00) Bayesian statistics and statistical learning theory.

- Introduce the framework of Bayesian statistics as used by Watanabe. Explain how to work with a supervised learning ML model such as a DNN in this framework, with the mean square error. Discuss briefly the fact that this is very different from ML practice **Refs:** [1, §1.1-1.4]
- Introduce the basic quantities of statistical learning theory (which let us track how well the model is learning) and their relationships: generalization,training and cross-validation loss (and their averages), Kullback-Leibler divergence (empirical and averaged), marginal likelihood/partition function, free energy. Point out the difference between losses and errors. **Refs:**[1, §1.6-1.7], omitting WAIC.
- Show in particular that, in the setting of function approximation with mean square error and Gaussian noise, the function K is nothing but the L^2-distance between the function predicted by the model and the true function (where L^2 is taken with respect to the distribution on inputs). This example is useful to keep in mind throughout the week, as functions are more concrete objects than probability distributions.
- Explain that one goal of statistical learning theory is to understand the behaviour of the various errors in the large n limit, and how the free energy is particularly important ([1, Rmk 10 in §1.7]. One could discuss [1, Example 1.9.4] of a simple Gaussian model where everything can be computed explicitly.

Talk 3 (13:00-14:00) Regular and singular models.

- Introduce the set of optimal parameters W_0; explain why we expect the posterior distribution to concentrate along W_0 and so why the "geometry" of W_0, of K_n and K (in a so far imprecise sense) should be relevant to statistical learning theory.
- Discuss identifiable and non-identifiable models, noting in particular that W_0 is a point for identifiable models. Mention Doob's theorem. Explain that DNNs are highly non-identifiable.
- Introduce the Fisher information matrix and its link with K. Define regular and singular models.
- State (semi-rigorously) the Bernstein-Von Mises theorem, which shows that for regular realizable models the posterior density converges in L^1-norm to a Gaussian distribution centered at the optimal parameter. Mention that the starting point is the Laplace approximation for integral, and sketch the proof in 1d.
- Explain (maybe showing plots!) that this is very false for singular models.
- State the asymptotic free energy formula for regular models (perhaps only in the realizable case to keep things simple) [1, §4.2 Theorem 4]

Q& A, discussion, break (14:00-15:00). Possible discussion topics.

- What are standard approaches in statistical learning theory? Why do they fall short for deep learning?

- SGD vs Bayesian inference, what is known rigorously (e.g. in regular convex models)? empirically?

Talk 4 (15:00-16:00) Analytic geometry and singularities.

- Recall definition of smooth and real-analytic functions. Explain why real-analytic functions are better suited to "do geometry" (basically, their sets of solutions are reasonable geometric objects, "generically" manifolds, unlike sets of solutions of smooth functions). Explain why K is a real-analytic function for a tanh DNN.
- Discuss smooth and singular points of a real-analytic set (and the implicit function theorem), and critical points of a real-analytic function.
- Define Morse and Morse-Bott functions and state the Morse-Bott lemma. Connect this to regular/"minimally singular" statistical models.
- Define normal crossings functions and state embedded resolution of singularities for real-analytic functions, in the form used by Watanabe. Explain how this looks like in particular for a positive function like K.
- Briefly mention that, in practice, singularity theory for real-analytic functions can often be reduced to singularity theory for polynomials, and so to real algebraic geometry, which is convenient for some computations; note however that K itself is almost never a polynomial.
- If there is time (!), give example of ADE singularities and their resolutions.

Talk 5 (16:00-17:00) Measuring degeneracy with the real log-canonical threshold.

- Define the real log-canonical threshold (RLCT) or learning coefficient in terms of volume asymptotics, and closely related in terms of integrability.
- Introduce the "density of states" (DOS) function/distribution.
- Explain roughly, using the discussion of the regular case, why understanding the asymptotic behaviour of the DOS close to the optimal parameters is a key step to understanding the asymptotic behaviour of the partition function.
- Explain why the asymptotics of the DOS is closely related to the asymptotics of the volume function, and so the RLCT as defined above. Discuss also the log term and the multiplicity.
- Introduce the zeta function and give the characterization of the RLCT in terms of poles (mentioning the role of the Mellin transform)
- Explain that the zeta function can be computed on a resolution of singularities and give the resulting formula for the RLCT.
- Discuss basic properties of the RLCT (semi-continuity?), and give the formula in the minimally singular case (K Morse-Bott).
- If there is time (!), give example of ADE singularities as a case where we can compute the RLCT and they reflect the "complexity" of the singularity.

Tuesday 18th: Singular learning theory - the core results

Q & A and discussion (10:00-11:00). Digesting the material from Monday, before going into the main results! Possible points that deserve clarification:

- Singularities are "measure-zero" objects, so why do we expect them to have any role in a probabilistic/statistical setting? Relatedly, machine learning systems are implemented on a computer, so with finite floating-point precision, so singularities should be "blurred out"; is this a reason to neglect them?

Talk 1 (11:00-12:00) Main theorem I: standard form of the log-likelihood function. Note: for this talk, assume $\beta = 1$ (ordinary Bayesian inference) to keep things simple?

- Introduce the "relatively finite variance" condition, explain that it holds for realizable models.

- Show that relatively finite variance implies essential uniqueness of the model along W_0, and that it implies that "$f(x,w)$ is divisible by $\sqrt{K(w)-K_0}$" in a neighbourhood of the set W_0 of optimal parameters.
- Use this to define the statistical process ξ_n on the resolution, give formulas for its expectation and variance. Draw analogy with the central limit theorem.
- Explain roughly what empirical process theory is about and how to deduce that ξ_n converges to a Gaussian process (under some technical assumptions which we don't need to discuss).

Talk 2: (13:00-14:00) Main theorem II: Free energy formula. Note: again, assume $\beta = 1$ (ordinary Bayesian inference) to keep things simple?

- Explain the simple upper bound for the expectation of the free energy [2, Theorem 6.4], which does not use
- Explain the decomposition of the partition function in two parts, an essential part (around W_0) and a non-essential part (away from W_0.)
- Explain how to bound the non-essential part.
- Explain how resolution of singularities combined with the standard form of K_n gives a formula for the essential part in terms of the stochastic process ξ_n.
- Show the convergence in law of the partition function as in [2, Theorem 6.7].
- Deduce the free energy formula, both empirically and in expectation.

Talk 3 (14:15-15:15): Main theorem III : consequences for generalization errors, singular fluctuation.

- Discuss tempered Bayesian inference (adding *beta*!) and the analogy with thermodynamics.
- Recall definitions of the various errors in Bayesian learning theory, and add in the Gibbs ones.
- Explain how to unify some of these errors with the "functional cumulant generating function" [1, §3.3].
- State (imprecisely) the "basic theorem of Bayesian statistics" [1, §3.4 Theorem 3].
- Define the renormalized posterior distribution, show that it satisfies the "scaling law" and define the singular fluctuation [1, §5.4 and §6.3]
- Explain how to apply the "basic theorem" in the case of SLT, using the scaling law, to obtain various asymptotic formulas [1, §5.4 and §6.3].

Q& A, discussion, work in small groups: (15:15-17:00). Possible discussion topics:

- How realistic is the condition of relatively finite variance? There is at least one example in [1, §3.4 Example 19] which does not satisfy it and where the free energy formula does not hold; can we understand and generalize it?
- ReLU activations, which are popular in modern deep-learning architectures are only piecewise-analytic, so the theory above cannot be literally applied to it. What can be said nonetheless?
-
- Can we understand qualitatively the next order terms in the free energy formula? Do we expect large multiplicities to play a role? The prior only comes in the constant term, which is very complicated; can be say something about it, perhaps by analogy with the simple formula in the regular case?
- What does the singular fluctuation mean, really? And how should it manifest in practice?
- Can we interpret the description of the "renormalized posterior distribution" and its scaling law as in [2, §6.3] as a "singular Bernstein-Von Mises theorem"? Can we recover a form of the classical BVM theorem by applying this to regular models?
- Does SLT say something about generalization out-of-distribution?
- We haven't discussed what Watanabe proves about maximum likelihood estimate in the singular context, but it is interesting and someone could try to summarize it.

Wednesday 20th: Singular learning theory - consequences for phase transitions

Talk 1 (10:00-11:00): internal model selection and phase transitions in the Bayesian posterior. Besides Watanabe's book, a useful reference for this is the talk of Dan Murfet .

- Introduce the general problem of model selection in Bayesian statistics [1, §8.1]
- Introduce local learning coefficients [4]
- Explain why the free energy formula implies phase transitions in the Bayesian posterior and discuss why this can be understood as a form of "internal model selection."[2, §7.6] [1, §9.4]

Discussion (11:00-12:00): Phase transitions in {SLT, deep learning, catastrophe theory, condensed matter physics} and how to connect it all. Phase transitions in the Bayesian framework of SLT are mathematically quantifiable phenomena, but we would like to use them to study the phase transitions which are observed empirically during SGD training in deep learning. To help translate between those two set-ups, ideas coming from the study of phase transitions in other contexts should prove useful. Since this is very much an unfinished story, a discussion format seems appropriate.

Talk 2 (13:00-14:00): Computing RLCTs with algebraic geometry.

- Explain why, in some cases, it is possible to reduce RLCT computations to the case of polynomials.
- Define blow-ups and explain that resolution of singularities can always be achieved by a sequence of blow-ups of smooth subvarieties.
- Choose a paper of Watanabe to illustrate the computation procedure.
- Could mention Newton polygons, toric methods, etc.

Talk 3 (14:15-15:15): WBIC paper. Based on the paper [3], which is an important theoretical foundation for the paper [4] which we look at on Thursday.

- Define the classical Bayesian information criterion and explain why it is appropriate for model selection for regular models.
- Define the Widely applicable Bayesian Information Criterion (WBIC). Explain why this is easier to estimate in practice than the actual free energy.
- State the main theorem [3, Theorem 4]. The proof is based on the same fundamental ingredients as the free energy formula; the key "new" computations are in [3, §5.7] and involve the Gamma function in some interesting way.
- Explain how to use WBIC to estimate the global RLCT, *provided* you can sample from the tempered posterior.

Q& A, discussion, work in small groups: (15:15-17:00). Many possible topics here.

- The function K (and hence the learning coefficient and other quantities relevant to statistical learning) depends in a complicated way of the interplay between the model and the true distribution. Can we quantify this interplay in some cases? For instance, if the true distribution has a certain symmetry property *and* the model contain a large subset of weights for which the model has this symmetry property, what does it imply about the structure of K? Can we deduce something about (local) learning coefficients?
- Another interesting notion introduced in [3, §3] is the *parity* of the model, which has to do with the analytic continuation of $\sqrt{K(w)}$ and explains some sign issues in the core SLT theory. The result [3, Corollary 1] shows that WBIC for models with odd parity should be a better estimator than for models with even parity; do we observe this in practice?
- Catastrophe theory!

Thursday 21st: Singular learning practice - estimating learning coefficients and applications to phase transitions

Talk 1 (10:00-11:00) "Quantifying degeneracy in singular models via the learning coefficient ". Goal: present the paper [4] by Lau-Murfet-Wei.

Talk 2 (11:00-12:00) Computing $\widehat{\lambda}$ in practice and applications.

Talk 3 (13:00-14:00) DIY $\widehat{\lambda}$: libraries, internals, APIs.

Q& A, discussion, work in small groups: (14:15-17:00).

References

[1] S. Watanabe, "Mathematical Theory of Bayesian Statistics"
[2] S. Watanabe, "Algebraic Geometry and Statistical Learning Theory"
[3] S. Watanabe, "A Widely applicable Bayesian Information Criterion"
[4] E. Lau, D. Murfet, S. Wei, "Quantifying degeneracy in singular models via the learning coefficient", `https://arxiv.org/abs/2308.12108`