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

## 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, classifyiers 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, SGD (we will probably come back to this later)
- 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? Can we understand the computational structure of trained models?

### 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:** [**?**, §1.1-1.4]
- 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:**[**?**, §1.6-1.7], omitting WAIC.
- 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 ([**?**, Rmk 10 in §1.7]. One could discuss [**?**, 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. 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. (NB: there is also a version of Bernstein-Von Mises for regular non-realizable models in [**?**])
- 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) [**?**, §4.2 Theorem 4]

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

- 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.

**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 integrability of K.
- 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, 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!

**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 $\xi_n$ 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 $\xi_n$ 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 [**?**, 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 $\xi_n$.
- Show the convergence in law of the partition function as in [**?**, 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" [**?**, §3.3].
- State (imprecisely) the "basic theorem of Bayesian statistics" [**?**, §3.4 Theorem 3].
- Define the renormalized posterior distribution, show that it satisfies the "scaling law" and define the singular fluctuation [**?**, §5.4 and §6.3]
- Explain how to apply the "basic theorem" in the case of SLT, using the scaling law [**?**, §5.4 and §6.3].
- Deduce the main formulas about

**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 [**?**, ] which does not satisfy it and where the free energy formula does not hold; can we understand and generalize it?
- 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" as in [**?**, §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.** Goal: explain

- Introduce the general problem of model selection in Bayesian statistics.
- Introduce local learning coefficients.

- Explain why the free energy formula implies phase transitions in the Bayesian posterior.
- Discuss why this can be understood as a form of "internal model selection."

**Discussion (11:00-12:00): Phase transitions in {SLT, deep learning, catastrophe theory, condensed matter physics} and how to connect it all.**

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

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

- Define the WBIC
-

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

### Thursday 21st: Singular learning practice - estimating learning coefficients and perspectives for interpretability.

**Talk 1 (10:00-11:00) "Quantifying degeneracy in singular models via the learning coefficient ".** Goal: present the paper [**?**] 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).**