Singular Learning Theory: Generalization in Deep Networks

Nahom Seyoum

2025-05-02

Contents

Introduction

Deep learning's impressive empirical success(broadly conceived) across many domains has exposed an inadequacy in classical statistical learning theory. Traditional frameworks like the Fisher information, Laplace approximations for marginal likelihood, and model selection criteria such as BIC, fundamentally presuppose a well-behaved, locally *regular* relationship between model parameters and the probability distribution they define. They assume the likelihood function is smooth and can be approximated by a quadratic surface around its optima, characterized by a non-singular Hessian or Fisher information matrix.

Feed-forward neural networks, which is what deep learning is primarily predicated upon, routinely violate these assumptions. The mapping from network weights to the input-output function is often highly non-identifiable and redundant. That is to say, different parameter values can produce the exact same function, leading to parameter spaces where the set of optimal solutions forms a complex, non-smooth structure known as an *analytic variety.* At points within this variety corresponding to the true distribution, the Fisher information matrix becomes singular (rank-deficient). This essentially makes standard quadratic approximations invalid and classical complexity measures (think VC, Rademacher complexity, Littlestone Dimension) like the number of parameters (d) misleading.

Singular Learning Theory (SLT), introduced by Sumio Watanabe, attempts to construct a mathematical framework to address this challenge. By using tools from algebraic geometry, SLT replaces classical curvature-based analyses with invariant measures derived from the *singular geometry* of the parameter-to-distribution map.

Sumio Watanabe's paper essentially lays out a theoretical framework, defining λ and proving its role in Bayesian generalization error and model evidence asymptotics through relations like:

$$\mathrm{Bg}(n) + \mathrm{Cv}(n) = \frac{2\lambda}{n} + o(n^{-1}) \quad (*)$$

A decade later, Daniel Murfet et al. empirically validate SLT for deep networks, showing they are strictly singular. They develop pretty creative practical methods to estimate λ from tempered posteriors and demonstrate its correlation with generalization. The key insight is this an interpretation of "flat minimum" and consequent architectural effects.

Finally, I look at a recent(2025) work by Miki Aoyagi where she completes the picture by providing exact, closed-form computations of the learning coefficients (λ, θ) for linear and three-layer ReLU networks. The work essentially confirms that λ decreases with network depth, and this is helpful in connecting singular geometry to phenomena like double descent.

Bayesian learning in singular models and Watanabe's equivalence theorem

A lot of this framework happens in the Bayesian setting. This is not without reason. The setup provides a way to quantify uncertainty over the parameters via the posterior distribution and implicitly handles model complexity through the properties of the marginal likelihood (the denominator in Bayes' theorem).

We can now move on to thinking about learning machines within this framework. A learning machine is formally defined by a family of probability densities $p(x \mid w)$ indexed by a parameter w that ranges in a compact analytic set $W \subset \mathbb{R}^d$. We assume that observed training samples $X_1, \ldots, X_n$ and an independent test point X are drawn from an unknown "true" distribution $q(x)dx$.

Given an inverse temperature $0 < \beta < \infty$, which allows us to interpolate between uniform prior ($\beta \to 0$) and concentration at the maximum likelihood estimate ($\beta \to \infty$), the posterior distribution $\pi_{n,\beta}(dw)$ on the parameter space W is given by:

$$\pi_{n,\beta}(dw) = \frac{\left[\prod_{i=1}^{n} p(X_i \mid w)^{\beta}\right] \varphi(w)\, dw}{\int_W \left[\prod_{i=1}^{n} p(X_i \mid u)^{\beta}\right] \varphi(u)\, du}, \qquad \varphi \text{ is an analytic and positive prior density on } W.$$

In the standard Bayesian setting, $\beta = 1$. The denominator in the posterior formula is the marginal likelihood or model evidence, $p(D_n|\text{model}) = \int_W \prod_{i=1}^n p(X_i|u)^\beta \varphi(u) du$.

To find the Bayes predictive density for a new data point x. we just average the model's prediction $p(x|w)$ over the posterior distribution:

$$p_n^*(x) = \int p(x \mid w)\, \pi_{n,\beta}(dw).$$

The goal of learning, then becomes making $p_n^*(x)$ a good approximation of the true distribution $q(x)$.

Generalisation, training, and functional variance

We quantify performance using the log-loss function $f(x, w) = -\log p(x \mid w)$. The quality of the predictive distribution $p_n^*(x)$ on unseen data is measured by the Bayes generalisation loss (BgL), which is the expected log-loss on a new test point X drawn from the true distribution q, averaged over the posterior $\pi_{n,\beta}$: $\text{BgL}(n) = \mathbb{E}_{X\sim q}[f(X, w)]_{\pi_{n,\beta}}$. I think about this as the average error we expect when using the learned model on new data.

The Bayes training loss (BtL) is the average log-loss on the training data, again averaged over the posterior: $\text{BtL}(n) = \frac{1}{n}\sum_{i=1}^n f(X_i, w)_{\pi_{n,\beta}}$. An intuition for this is that measures how well the model fits the data it was trained on.

The difference between generalization and training loss, BgL(n) - BtL(n), is effectively overfitting.. Classical methods estimate BgL(n) using BtL(n) plus a penalty, somewhat related to the model dimension. However, they assume model regularity. Watanabe addresses this by considering the fluctuation of the loss across the posterior. The amplitude of this fluctuation is measured by the functional variance $V(n)$:

$$V(n) = \sum_{i=1}^n \left\{ \mathbb{E}[f(X_i, w)^2]_{\pi_{n,\beta}} - \left(\mathbb{E}[f(X_i, w)]_{\pi_{n,\beta}}\right)^2 \right\}.$$

$V(n)$ is how much the loss value for each training point varies when w is sampled from the posterior, summed over all training points. It's a measure of the uncertainty in the training loss prediction conferred by the posterior distribution.

Watanabe defines the Widely Applicable Information Criterion (WAIC) as a potential estimator for the generalization loss that works even for singular models:

$$\text{WAIC}(n) = \text{BtL}(n) - \frac{\beta}{n} V(n).$$

This form resembles classical information criteria: a training error term (BtL) plus a complexity penalty term $(\beta/n)V(n)$. The penalty is based on the functional variance $V(n)$, which intuitively measures model flexibility with respect to the data distribution, rather than simply parameter count. Watanabe proves that $\mathbb{E}[\text{WAIC}(n)]$ asymptotically estimates $\mathbb{E}[\text{BgL}(n)]$ up to $o(n^{-1})$ terms, regardless of whether the model is singular or regular.

Cross-validation in the Bayesian posterior

Cross-validation is probably the most popular measure of generalization error. Bayesian leave-one-out cross-validation computes the predictive probability of each training point X_i when the model is trained on the remaining $n - 1$ points $D_n \setminus \{X_i\}$. For each index i, let $\pi_{n,\beta}^{(i)}$ be the posterior constructed from $D_n \setminus \{X_i\}$, and $p_n^{(i)}(x)$ the corresponding predictive density. The leave-one-out cross-validation loss is:

$$\text{CvL}(n) = -\frac{1}{n}\sum_{i=1}^n \log p_n^{(i)}(X_i).$$

Using importance sampling weighted by $p(X_i|w)^{-\beta}$, the predictive probability of X_i given the other data points $D_n \setminus \{X_i\}$ is:

$$p_n^{(i)}(X_i) = \frac{\mathbb{E}_w[p(X_i \mid w)^{1-\beta}]}{\mathbb{E}_w[p(X_i \mid w)^{-\beta}]},$$

where the expectations $\mathbb{E}_w[\cdot]$ are taken with respect to the normalized density proportional to $\prod_{j=1}^n p(X_j|w)^\beta \varphi(w)$. This leads to the form of CvL given in the draft:

$$\mathrm{CvL}(n) = -\frac{1}{n}\sum_{i=1}^n \log \frac{\mathbb{E}_w[p(X_i \mid w)^{1-\beta}]}{\mathbb{E}_w[p(X_i \mid w)^{-\beta}]}.$$

We can compare CvL(n) and WAIC(n) asymptotically and Watanabe uses the functional-cumulant generating function:

$$F(\alpha) = \frac{1}{n}\sum_{i=1}^n \log \mathbb{E}_w\left[e^{-\alpha f(X_i,w)}\right],$$

The neat part is that the logarithm of an expectation $\log \mathbb{E}[e^{-\alpha Y}]$ has a Taylor expansion in α whose coefficients are the cumulants of the random variable Y. For $f(X_i, w)$ under the posterior, let $\kappa_k(i)$ denote its k-th cumulant. Then $\log \mathbb{E}_w[e^{-\alpha f(X_i,w)}] = \sum_{k=1}^\infty \frac{(-\alpha)^k}{k!}\kappa_k(i)$. $F(\alpha) = \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^\infty \frac{(-\alpha)^k}{k!}\kappa_k(i) = \sum_{k=1}^\infty \frac{(-\alpha)^k}{k!}\left(\frac{1}{n}\sum_{i=1}^n \kappa_k(i)\right)$. Let $Y_k(n) = \frac{1}{n}\sum_{i=1}^n \kappa_k(i)$ be the empirical average of the k-th cumulant over the training data. $Y_1(n) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_w[f(X_i, w)] = \mathrm{BtL}(n)$. $Y_2(n) = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}_w[f(X_i, w)] = V(n)/n$. The expansion of $F(\alpha)$ becomes $F(\alpha) = -Y_1(n)\alpha + \frac{Y_2(n)}{2}\alpha^2 - \frac{Y_3(n)}{6}\alpha^3 + O_p(n^{-2})$ as $Y_k(n)$ scales as $O_p(n^{1-k/2})$.

Using these expansions for CvL(n) and WAIC(n), we get the asymptotic forms (truncating to order n^{-2}):

$$\mathrm{WAIC}(n) = -Y_1(n) + \frac{2\beta - 1}{2}Y_2(n) - \frac{1}{6}Y_3(n) + O_p(n^{-2}),$$

$$CvL(n) = -Y_1(n) + \frac{2\beta - 1}{2}Y_2(n) - \frac{3\beta^2 - 3\beta + 1}{6}Y_3(n) + O_p(n^{-2}),$$

The difference becomes

$$CvL(n) - \mathrm{WAIC}(n) = -\frac{\beta(\beta - 1)}{2}Y_3(n) + O_p(n^{-2}).$$

Given $Y_3(n) = O_p(n^{-1/2})$, the difference is $O_p(n^{-3/2})$ for general β. When $\beta = 1$, this term is zero, which gives us $O_p(n^{-2})$.

$$CvL(n) - \mathrm{WAIC}(n) = O_p(n^{-3/2}) \quad \text{for arbitrary } \beta, \qquad CvL(n) - \mathrm{WAIC}(n) = O_p(n^{-2}) \text{ when } \beta = 1. \tag{1}$$

This is ultimately the crux of Watanabe's Theorem 1: WAIC and leave-one-out cross-validation are asymptotically equivalent as random variables. Their difference converges to zero faster than $1/n$.

Birational invariants: the real log-canonical threshold

In a regular model, the expected log-likelihood $\mathbb{E}_X[\log p(X|w)]$ has a unique maximum at w_0, and near w_0, the KL divergence $K(w) = \mathbb{E}_X[f(X, w)] - \mathbb{E}_X[f(X, w_0)]$ behaves quadratically:

$$K(w) \approx \frac{1}{2}(w - w_0)^T I(w_0)(w - w_0),$$

where $I(w_0)$ is the non-singular Fisher Information Matrix. The integral of $e^{-nK(w)}$ over a neighborhood of w_0 is quite important for large n asymptotics (Laplace approximation). This integral scales like

$\int e^{-n\frac{1}{2}\|w-w_0\|^2_{I(w_0)}}dw \sim n^{-d/2}$. This $n^{-d/2}$ scaling of volume near the minimum gives us the $d/2 \log n$ term in the asymptotic expansion of the negative log marginal likelihood:

$$\mathbb{E}[-\log p(D_n|\text{model})] \approx n\mathbb{E}_X[f(X, w_0)] + \frac{d}{2}\log n.$$

As aformentioned, for deep networks, the parameter-to-distribution map is non-identifiable, and the set $W_0 = \{w \mid K(w) = 0\}$ is a singular analytic variety. $K(w)$ does not vanish quadratically near W_0. To analyze the large n asymptotics of the posterior integral $\int \prod p(X_i|w)^\beta \varphi(w)dw$, which depends on the behavior of the likelihood function near W_0, we use algebraic geometry.

KL divergence $K(w)$ quantifies the discrepancy from the true distribution (pretty well known). What Wantanabe does that is particularly interesting, is learn the zeta-function $\zeta(z) = \int_W K(w)^z \varphi(w)\,dw$. For a regular model with $K(w) \sim \|w - w_0\|^2$, the volume $\{w \mid K(w) < \epsilon\}$ scales like $\epsilon^{d/2}$, leading $\zeta(z)$ to have poles related to $-d/2$.

For a singular model, the volume scaling is more interesting, say like $\epsilon^\lambda(\log(1/\epsilon))^{\theta-1}$. This scaling implies $\zeta(z)$ has poles at $z = -\lambda$ with multiplicity θ.

But this resolution of singularities means that we can analyze $\zeta(z)$. It maps a smooth space U to W ($\pi : U \to W$) such that $K(w)$ becomes simpler in u coordinates, locally $\prod u_j^{k_j}$. $\int K(w)^z\varphi(w)dw$ transforms to $\int K(\pi(u))^z\varphi(\pi(u))|\det D\pi(u)|du$. The exponents k_j of $K(\pi(u))$ and h_j of the Jacobian term $|\det D\pi(u)|$ determine the poles $z = -(h_j+1)/(2k_j)$. The real log-canonical threshold (λ) is defined as the maximum of these values:

$$\lambda = \max_j \frac{h_j + 1}{2k_j}.$$

λ is a birational invariant, independent of parameterization or prior. This is very interesting because it gives us a way of measuring the severity of the singularity and replacing $d/2$ as the volume scaling exponent near W_0. The singular fluctuation $\nu(\beta)$ is a second invariant! This means specific posterior spread, defined via $V(n)$ is captured by it.

Watanabe proves that these invariants govern the asymptotic expectation of errors:

$$\lim_{n\to\infty} n\,\mathbb{E}[\mathrm{Bg}(n)] = \frac{\lambda - \nu(\beta)}{\beta} + \nu(\beta), \qquad \lim_{n\to\infty} n\,\mathbb{B}{\approx}(n)] = \frac{\lambda - \nu(\beta)}{\beta} - \nu(\beta). \tag{2}$$

$n(\mathrm{Bg}(n) + \mathrm{Bt}(n)) + V(n) \to_p 2\lambda/\beta$ also holds.

Cross-validation mirrors generalisation through λ

Combining the CvL/WAIC equivalence (Eq. 1) and the error asymptotics (Eq. 2) yields Theorem 2, we get:

$$\mathrm{Bg}(n) + \mathrm{CvL}(n) = \frac{(\beta-1)V(n)}{n} + \frac{2\lambda}{\beta n} + o_p(n^{-1});$$

$$\text{in particular, when } \beta = 1,\ \mathrm{Bg}(n) + \mathrm{CvL}(n) = \frac{2\lambda}{n} + o_p(n^{-1}). \tag{3}$$

This essentially means CvL(n) is an asymptotically unbiased estimator for $\mathrm{Bg}(n)$ up to $O(1/n)$, with the bias term controlled by λ and ν. For standard Bayesian learning ($\beta = 1$), it is the relationship is particularly simple. The key takeaway is that λ's directly influences the $1/n$ bias term.

Corollaries: For $\beta = 1$, $Bg(n)$ and $CvL(n)$ have asymptotically the same variance, making CvL/WAIC optimal estimators in a sense. Since $\lambda \ll d/2$ for deep nets, the $2\lambda/n$ bias is much smaller than d/n or $(d\log n)/(2n)$. This is why BIC over-penalizes singular models.

Sketch of the proof of Theorem 3

1. WAIC-CvL Equivalence (Theorem 1): This follows from using the cumulant expansions of the functional cumulant generating function $F(\alpha)$. We expand $F(\alpha)$ around $\alpha = 0$ up to order n^{-2} and compare the expansions for WAIC and CvL. This gives us $O_p(n^{-3/2})$ or $O_p(n^{-2})$.

2. Asymptotic Expansion of the Free Energy: We start with the expected negative log marginal likelihood, $\mathbb{E}[-\log Z_{n,\beta}]$:

   $$Z_{n,\beta} = \int \prod p(X_i|w)^\beta \varphi(w) dw.$$

   We taylor expand $\sum f(X_i, w)$ around w_0 and relate the integral to $\int e^{-n\beta K(w)}\varphi(w)dw$ (plus terms for empirical fluctuations), We can then claim that SLT shows that the expectation has an asymptotic expansion:

   $$\mathbb{E}[-\log Z_{n,\beta}] = n\mathbb{E}_X[f(X, w_0)] + \lambda \log(n\beta) - (\theta - 1)\log\log(n\beta) + \mathcal{O}(1).$$

   We get λ from the $n^{-\lambda}\beta^{-\lambda}$ scaling of the integral $\int e^{-n\beta K(w)}\varphi(w)dw$. This scaling happens because the volume of parameter space where $K(w) < \epsilon$ scales like ϵ^λ, and the integral $\int e^{-CK(w)}d\mu$ scales like $C^{-\lambda}$ for large C, with $C = n\beta$.

3. Relating Errors to Free Energy: The expected training and generalization errors are related to the free energy through physics thermodynamic-like identities, which I admittedly have very little intuition for. But, for instance, the expected value of the total log-likelihood $\mathbb{E}[n \cdot \mathrm{BtL}(n)]$ is related to the derivative of the expected free energy with respect to inverse temperature:

   $$\mathbb{E}[n \cdot \mathrm{BtL}(n)] = \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}_w[f(X_i, w)]\right] = -\frac{\partial}{\partial \beta}\mathbb{E}[-\log Z_{n,\beta}].$$

   Now we just take derivatives of the asymptotic expansion from Step 2 with respect to β. This gives us the asymptotic forms in Eq. (2) involving λ and ν. For instance, the derivative of $\lambda \log(n\beta)$ with respect to β is $\lambda \frac{1}{n\beta} n = \lambda/\beta$. After some more tinkering involving some expectations over data sets, we see that it accounts for the λ terms in Eq. (2).

4. Combining Results and Variance Control: Now we just need to combine the asymptotic equivalence of CvL and WAIC (Step 1) with the relationships derived in Step 3. We have the term $(\beta - 1)V(n)/n$ in Eq. (3) from the difference between WAIC and expected training loss. By simple controls over the fluctuations of things like $V(n)$ around their expectations (Lemma 1), we can convert these expectation-based equalities (like Eq. 2) into high-probability statements (like Eq. 3).

The Generic Singularity of Deep Networks

Section 1 introduced that SLT applies when a model's set of optimal parameters forms a singular analytic variety, leading to a singular Fisher Information Matrix or KL Hessian. Murfet et al. show that deep neural networks generically do in fact have this property. They then try to outline why their Fisher Information matrix is degenerate.

The argument is that the problem largely has to do with the inherent non-identifiability and parameter redundancy in deep networks, similar to what was outlined in Section 1. Many different parameter values w produce the exact same function $f(x, w)$. This redundancy implies that the gradient vectors $\{\partial_{w_i} f(x, w)\}$ are linearly dependent in function space. It is easy to see that the Fisher Information and the KL Hessian (at a true parameter) are Gram matrices of these gradients. Therefore, their singularity is equivalent to this linear dependency: there exists a non-zero vector v such that

$$\sum_{i=1}^{d} v_i \partial_{w_i} f(x, w_0) = 0 \quad \text{for } q\text{-almost all } x.$$

Murfet et al. prove the existence of such v for ReLU and SiLU networks (Lemma 2). I will give a quick sketch down below.

The core of the argument for ReLU networks comes from a differential equation satisfied by the network function f. Lemma 1 (Appendix A.1) shows that for any hidden ReLU neuron j in layer l, the network function f satisfies:

$$\left(\sum_k w^l_{jk} \frac{\partial}{\partial w^l_{jk}} + b^l_j \frac{\partial}{\partial b^l_j} - \sum_i w^{l+1}_{ij} \frac{\partial}{\partial w^{l+1}_{ij}} \right) f(x, w) = 0$$

Obviously, this holds for inputs x where none of the pre-activations u are zero. The terms w^l_{jk} are weights *into* neuron j from layer $l-1$, b^l_j is the bias of neuron j, and w^{l+1}_{ij} are weights *out of* neuron j to layer $l+1$.

Much of the proof is spent computing the partial derivatives $\frac{\partial f}{\partial w^l_{jk}}$, $\frac{\partial f}{\partial b^l_j}$, and $\frac{\partial f}{\partial w^{l+1}_{ij}}$ using the chain rule and the properties of the ReLU derivative, and showing that this specific linear combination sums to zero. The implication is essentially that this corresponds to a direction in parameter space (defined by the vector v with components $w^l_{jk}, b^l_j, -w^{l+1}_{ij}$) where the function $f(x, w)$ does not change locally.

Lemma 2 uses result to prove the degeneracy of the Fisher and Hessian. The Fisher Information Matrix $I(w)$ has entries $I(w)_{rs} = \int \langle \frac{\partial f(x,w)}{\partial w_r}, \frac{\partial f(x,w)}{\partial w_s} \rangle q(x) dx$. If we form a linear combination of the *rows* of this matrix using the coefficients v_s from the differential equation, we get:

$$\sum_s v_s I(w)_{rs} = \sum_s v_s \int \left\langle \frac{\partial f}{\partial w_r}, \frac{\partial f}{\partial w_s} \right\rangle q(x) dx = \int \left\langle \frac{\partial f}{\partial w_r}, \sum_s v_s \frac{\partial f}{\partial w_s} \right\rangle q(x) dx.$$

Since $\sum_s v_s \frac{\partial f}{\partial w_s} = 0$ (from Lemma 1), $\sum_s v_s I(w)_{rs} = 0$. The implication is that the rows of $I(w)$ are linearly dependent for any w with a non-trivial hidden neuron. Therefore, $I(w)$ is degenerate!

For the Hessian of the KL divergence, $D^2K(w)$, its general form is given by $D^2K(w)_{rs} = \int \langle \frac{\partial f}{\partial w_r}, \frac{\partial f}{\partial w_s} \rangle q(x)dx + \int \langle f(x, w) - f(x, w_0), \frac{\partial^2 f}{\partial w_r \partial w_s} \rangle q(x) dx$. At a true parameter $w_0 \in W_0$, $f(x, w_0) - f(x, w_0) = 0$. Therefore,

$$D^2K(w_0)_{rs} = \int \left\langle \frac{\partial f(x, w_0)}{\partial w_r}, \frac{\partial f(x, w_0)}{\partial w_s} \right\rangle q(x) dx.$$

This shows that at a true parameter w_0, the Hessian $D^2K(w_0)$ is proportional to the Fisher Information Matrix $I(w_0)$. Since $I(w_0)$ is degenerate (as shown above), $D^2K(w_0)$ is also degenerate at any true parameter w_0 involving non-trivial hidden neurons.

Basically the above construction is a rigorous construction that shows that deep networks satisfy the singularity condition in SLT i.e. the Hessian at the true parameter is degenerate.

Empirical Estimation of λ via Tempered Posteriors

In section 1, we established that the theoretical invariant λ governs the asymptotic behavior of Bayesian quantities. For instance, it appears in expansions of the marginal likelihood and expected errors. Murfet et al. use one of these asymptotic results (Theorem 4 in Watanabe [2013], related to the free energy expansion in Section 1) about the expected total training loss evaluated under tempered posteriors. The theory essentially states that for large n and small temperature $T = 1/\beta$:

$$\mathbb{E}_{D_n} \left[\mathbb{E}_{w \sim \pi_{n,1/T}} \left[nL_n(w) \right] \right] \approx n\mathbb{E}_X[f(X, w_0)] + \lambda T.$$

This linear relationship between expected total training loss and temperature T, with slope λ, is how we construct the empirical estimation method. For a fixed dataset D_n, we approximate the expectation over datasets by sampling w from the tempered posterior $\pi_{n,1/T}$ for various temperatures T_j. The procedure works as follows:

1. Loop through each training dataset $\mathcal{D}_n$ in the set $\mathcal{T}$. I think about this in robustness terms i.e. because λ is a property of the model class and the true distribution, averaging estimates over multiple datasets provides a more robust result than relying on a single dataset's specific realization of noise.

2. For the current dataset $\mathcal{D}_n$, enter an inner loop that iterates through the specified range of inverse temperatures β_j (or temperatures $T_j = 1/\beta_j$).

3. Inside this inner loop, for each specific inverse temperature β_j, obtain a set of R approximate samples $\{w_1, \ldots, w_R\}$ from $p^{\beta_j}(w|\mathcal{D}_n)$. They use a NUTS variant of Hamiltonian Monte Carlo (HMC) to get these samples. Without getting into the weeds, it is apparently great at "exploring complex parameter spaces". In their experiments, they collected $R = 20,000$ samples for each β_j and dataset, discarding the first 1000 as burn-in.

4. Using these samples, approximate the expected total training loss under the posterior for this β_j and $\mathcal{D}_n$:

$$\widehat{E_j} = \frac{1}{R}\sum_{r=1}^{R} nL_n(w_r).$$

   The $\widehat{E_j}$ essentially provides a data point $(T_j, \widehat{E_j})$ corresponding to the theoretical relationship $\mathbb{E}_{D_n}[\mathbb{E}_{\pi_{n,1/T}}[nL_n(w)]] \approx n\mathbb{E}_X[f(X, w_0)] + \lambda T$.

5. Once the inner loop is complete (i.e., $\widehat{E_j}$ has been computed for all T_j in $\mathcal{D}_n$), perform a linear regression on the pairs $\{(T_j, \widehat{E_j})\}$ to fit the linear model. The slope essentially gives you an estimate $\hat{\lambda}(\mathcal{D}_n)$ for the true λ based on this single dataset(they use generalized least squares to get this). In their Table 2 experiments, they used 5 different inverse temperatures β_j.

6. After looping through all datasets in $\mathcal{T}$ and obtaining an estimate $\hat{\lambda}(\mathcal{D}_n)$ for each, the algorithm outputs the average of these estimates:

$$\hat{\lambda} = \frac{1}{|\mathcal{T}|}\sum_{\mathcal{D}_n \in \mathcal{T}} \hat{\lambda}(\mathcal{D}_n).$$

   The final average is the empirical estimate of the real log canonical threshold λ.

Their experiments on small networks (Table 2) had a $\hat{\lambda} \approx 0.55$, which is way lower than the classical $d/2 = 10.5$. Therefore, the excellent linear fits ($R^2 > 0.99$) empirically validated the predicted asymptotic linear behavior. So now we have a validation for the lower effective complexity interpretation/explanation.

λ, Flat Minima, and Generalization Error

Now that we have found $\lambda \ll d/2$ from empirical estimation, we can actually say some things about some of the key concepts we are interested in(generalization, minima etc). For one, we have a theoretical explanation for why classical "flat minima" heuristics, often using the Hessian determinant, fail. As discussed in Section 1, the singular geometry means the volume of parameters near the true function scales with $\epsilon^{\lambda}(\log(1/\epsilon))^{\theta-1}$, not $\epsilon^{d/2}$. λ is the correct exponent governing this volume scaling, relevant for Bayesian model evidence, not the local Hessian determinant.

Another thing we can verify is the direct link between λ and generalization error predicted by SLT (Theorem 2, Section 1, implying $\mathbb{E}_n G(n) \approx \lambda/n$) is also empirically verified. Murfet et al. plot $n \cdot G(n)$ against n. As expected if $G(n) \approx \lambda/n$, their plots (Figure 1) show $nG(n)$ converging to a value that matches the independently estimated $\hat{\lambda}$ (Table 1). In essence, this is evidence that λ controls the asymptotic generalization rate. By extension, architecture choices impacting λ (even if increasing d) can improve generalization.

Finally, they apply $\hat{\lambda}$ to Bayesian model selection. Classical BIC uses a $d \log n$ penalty, which is heir to the classic $n^{-d/2}$ marginal likelihood scaling in regular models. For singular models, theory predicts an $n^{-\lambda}$ scaling i.e. a $\lambda \log n$ penalty (WBIC). Murfet et al. essentially show that WBIC using their estimated $\hat{\lambda}$ successfully identifies best-generalizing models.

KL Risk for Linear Feed-Forward Networks

Now we are on to our final survey paper. Before we get to the result, we need to set up some preliminaries.

Let us consider a fully linear feed-forward network with L hidden layers. We denote the dimension of layer s as $H^{(s)}$, for $s = 1, \dots, L+1$, where $H^{(L+1)}$ is the input dimension and $H^{(1)}$ is the output dimension. The network function $h(x, w)$ is a composition of linear maps $F^{(s)}(x) = A^{(s)}x + B^{(s)}$, where $A^{(s)}$ is a weight matrix of size $H^{(s)} \times H^{(s+1)}$ and $B^{(s)}$ is a bias vector of size $H^{(s)}$. The network function for an input $x \in \mathbb{R}^{H^{(L+1)}}$ is given by the composition $h(x, A, B) = F^{(1)} \circ F^{(2)} \circ \cdots \circ F^{(L)}(x)$.

Expanding this composition, we can write $h(x, A, B)$ as a linear function of x:

$$h(x, A, B) = \left(\prod_{s=1}^{L} A^{(s)} \right) x + \sum_{S=2}^{L} \left(\prod_{s=1}^{S-1} A^{(s)} \right) B^{(S)} + B^{(1)}.$$

Here, the product $\prod_{s=1}^{S-1} A^{(s)}$ is interpreted as the identity matrix if $S - 1 < 1$. Let $w = (\{A^{(s)}\}, \{B^{(s)}\})$ denote the set of parameters and $w^* = (\{A^{*(s)}\}, \{B^{*(s)}\})$ be the true parameters.

Assuming a Gaussian noise model for the output, the KL divergence $K(w)$ is given by:

$$K(w) = \frac{1}{2} \int_X \|h(x, w) - h(x, w^*)\|^2 q(x) dx,$$

where $q(x)$ is the probability density function of the input x.

The key result we will be focusing on is Aoyagi's Theorem 3, which essentially simplifies the seemingly hard problem of analyzingp the singularity of $K(w)$. It states that for a polynomial function $h(x, w)$ in x (which our linear network function is) and a positive continuous function $q(x)$ on its support $X \subset \mathbb{R}^N$ with $\int_X q(x)dx > 0$, there exist positive constants α_1, α_2 such that the L^2 $\int_X h^2(x, w)q(x)dx$ is equivalent to the squared Euclidean norm of the polynomial's coefficients. Another way to think about it is, if $v(w)$ is the vector of coefficients of $h(x, w)$ as a polynomial in x, then $K(w)$ is equivalent to $\|v(w) - v(w^*)\|^2$.

$$\alpha_1 \|v(w)\|^2 \leq \int_X h^2(x, w)q(x)dx \leq \alpha_2 \|v(w)\|^2.$$

The proof is fairly straightforward. We begin by showing $\int_X h^2(x, w)q(x)dx = v(w)^T \left(\int_X C(x)q(x)dx \right) v(w)$, where $C(x)$ is a positive semidefinite matrix whose elements are products of monomials in x. Since $q(x) > 0$ on X, $\int_X C(x)q(x)dx$ is positive definite. This gives us the equivalence with $\|v(w)\|^2$.

Applying this to$h(x, w) - h(x, w^*)$, which is also a linear function of x, means that analyzing the singularity of $K(w)$ near the set of true parameters $W_0 = \{w \mid h(x, w) = h(x, w^*) \text{ for all } x\}$ is equivalent to analyzing the singularity of the squared Euclidean norm of the vector of coefficients of $h(x, w) - h(x, w^*)$.

We can actually write $h(x, w) - h(x, w^*)$ explicitly as:

$$h(x, w) - h(x, w^*) = \left(\prod_{s=1}^{L} A^{(s)} - \prod_{s=1}^{L} A^{*(s)} \right) x + \left(\sum_{S=2}^{L} \left(\prod_{s=1}^{S-1} A^{(s)} B^{(S)} - \prod_{s=1}^{S-1} A^{*(s)} B^{*(S)} \right) + (B^{(1)} - B^{*(1)}) \right).$$

Theorem 4 proves that the singularity of $K(w)$ is equivalent to the singularity of the sum of the squared Frobenius norm of the matrix product difference and the squared Euclidean norm of the overall bias difference term:

$$c_{w^*}(K(w)) = c_{w^*}(\| \prod_{s=1}^{L} A^{(s)} - \prod_{s=1}^{L} A^{*(s)} \|^2 + \|\text{BiasTerms}(w) - \text{BiasTerms}(w^*)\|^2).$$

where $\text{BiasTerms}(w) = \sum_{S=2}^{L} \left(\prod_{s=1}^{S-1} A^{(s)} \right) B^{(S)} + B^{(1)}$.

Her analysis shows that the term $\|\text{BiasTerms}(w) - \text{BiasTerms}(w^*)\|^2$ can be related to the norm squared of a vector in $\mathbb{R}^{H^{(1)}}$, which contributes $H^{(1)}/2$ to λ independently. Much of it focuses on $\| \prod_{s=1}^{L} A^{(s)} - \prod_{s=1}^{L} A^{*(s)}\|^2$. The singularity arises from the constraint that $\prod A^{(s)} = \prod A^{*(s)}$, especially when the rank r of the true product $\prod A^{*(s)}$ is less than the minimum dimension of the matrices involved. Unsurprisingly, the geometry of this singularity is complex and depends on the dimensions of all layers.

Resolving Singularities using Blow-ups

To analyze the rate at which the term $\| \prod_{s=1}^{L} A^{(s)} - \prod_{s=1}^{L} A^{*(s)}\|^2$ vanishes near the singular set where the product is fixed to $\prod A^{*(s)}$, Aoyagi uses "resolution of singularities" via toric blow-ups". This is where she does some algebraic geometry machinery to handle the geometry near W_0.

As we discussed earlier, a blow-up provides a coordinate change $\pi : U \to W$ that simplifies the local structure of a singular set or function. For singularities related to matrix rank, a sequence of blow-ups transforms the parameters w into new local coordinates $u = (u_1, \ldots, u_m)$ such that the function we are analyzing (the matrix product difference norm squared) takes a simplified form near $u = 0$:

$$\| \prod_{s=1}^{L} A(\pi(u)) - \prod_{s=1}^{L} A^*(\pi(u))\|^2 \approx \prod_{j=1}^{m} u_j^{2k_j}$$

for positive integer exponents k_j. Simultaneously, the transformation of the prior density weighted by the Jacobian determinant also takes a monomial-like form:

$$|\det D\pi(u)|\varphi(\pi(u)) \approx \prod_{j=1}^{m} u_j^{h_j} b(u),$$

where h_j are integers and $b(u)$ is a non-zero analytic function at $u = 0$.

The interesting result is that the learning coefficients λ and θ are determined by these exponents. The integral $\int K(w)^z \varphi(w) dw$ transforms to $\int K(\pi(u))^z |\det D\pi(u)|\varphi(\pi(u)) du$. Near the singularity ($u = 0$), this integral behaves like $\int (\prod u_j^{2k_j})^z (\prod u_j^{h_j}) b(u) du = \int \prod u_j^{2zk_j + h_j} b(u) du$. The poles of this integral, which determine the poles of $\zeta(z)$(the learning function), occur when $2zk_j + h_j = -1$ for any index j, leading to poles at $z = -(h_j + 1)/(2k_j)$.

So why go through all this machinery. Ultimately, by transforming the singular integral into an integral involving monomials, we can explicitly read off the exponents k_j and h_j. These exponents determine the poles of the learning zeta function $\zeta(z)$, and thus the real log canonical threshold λ and its multiplicity θ. As established in Section 1 (Watanabe's theory), λ is defined by these poles and directly governs the asymptotic behavior of Bayesian generalization error and model evidence. What Aoyagi gives us is the math to "compute" these exponents for specific models like linear networks by explicitly identifying how these exponents combine based on the network dimensions and rank.

Exact Learning Coefficients for Linear Networks

Definition 4 introduces parameters ℓ, M, a based on the analysis of the excess widths $M^{(s)} = H^{(s)} - r$. Let $\{S_1, \ldots, S_{\ell+1}\}$ be the indices of the $\ell + 1$ smallest excess widths $M^{(s)}$ for $s = 1, \ldots, L + 1$. These parameters are then used in the formula for $\lambda(H^{(1)}, \ldots, H^{(L+1)}, r)$ and $\theta(H^{(1)}, \ldots, H^{(L+1)}, r)$. These essentially represent the contributions from the product part of the singularity. The formula for the product part contribution to λ is given by:

$$\lambda(H^{(1)}, \ldots, H^{(L+1)}, r) = \frac{-r^2 + r(H^{(1)} + H^{(L+1)})}{2} + \frac{Ma + (M-1)\sum_{j=1}^{\ell+1} M^{(S_j)}}{4\ell} - \frac{1}{4}\sum_{j=1}^{\ell+1}(M^{(S_j)})^2.$$

The first two terms, $\frac{-r^2+r(H^{(1)}+H^{(L+1)})}{2}$, come from the base geometry of the space of matrices of rank r. The remaining terms come from some tinkering with the exponents yielded by the blow-up centered on the singularity related to the excess dimensions in the layers.

Now we move to theorem 4 where we ultimately get a representation for the learning coefficients in a linear network. For this, we just combine the contribution from the product part and the bias term $B^{(1)}$. As mentioned earlier, the bias $B^{(1)}$ contributes $H^{(1)}/2$ to λ and 1 to θ independently. Therefore:

$$\lambda = \frac{H^{(1)}}{2} + \lambda(H^{(1)}, H^{(2)}, \ldots, H^{(L+1)}, r)$$
$$\theta = \theta(H^{(1)}, H^{(2)}, \ldots, H^{(L+1)}, r)$$

where $\lambda(\ldots, r)$ and $\theta(\ldots, r)$ follow from Defn 4.

We can actually prove this. We just have to show equivalence between the singularity of $K(w)$ and the sum of the squared norms of the matrix product difference.

Recall that the KL divergence is $K(w) = \frac{1}{2}\int_X \|h(x,w) - h(x,w^*)\|^2 q(x)dx$. As shown by Theorem 3, $K(w)$ is equivalent to the squared norm of the vector of coefficients of $h(x,w) - h(x,w^*)$ as a polynomial in x. The difference $h(x,w) - h(x,w^*)$ can be rewritten as

$$h(x,w) - h(x,w^*) = \left(\prod_{s=1}^{L} A^{(s)} - \prod_{s=1}^{L} A^{*(s)}\right) x + \left(\sum_{S=2}^{L}\left(\prod_{s=1}^{S-1} A^{(s)}B^{(S)} - \prod_{s=1}^{S-1} A^{*(s)}B^{*(S)}\right) + (B^{(1)} - B^{*(1)})\right).$$

Let $P_L(w) = \prod_{s=1}^{L} A^{(s)}$ and $B_{terms}(w) = \sum_{S=2}^{L}(\prod_{s=1}^{S-1} A^{(s)})B^{(S)} + B^{(1)}$. The expression is $(P_L(w) - P_L(w^*))x + (B_{terms}(w) - B_{terms}(w^*))$.

Now we use Lemma 1(2) in the paper which states that the log canonical threshold of a sum of squared norms of vectors of functions is equivalent to the log canonical threshold of the sum of the squared norms if the sets of variables are independent. We are interested in this because it allows us to treat the singularity arising from the matrix product difference and the singularity arising from the bias terms separately.

$$\|P_L(w) - P_L(w^*)\|^2 + \|B_{terms}(w) - B_{terms}(w^*)\|^2.$$

$\|P_L(w) - P_L(w^*)\|^2$ tells you a lot about the entries of the matrix product difference. If we analyze the singularity using blow-ups, we get $\lambda(H^{(1)}, \ldots, H^{(L+1)}, r)$ given in Definition 4.

Now we can do change of variable on $\|B_{terms}(w) - B_{terms}(w^*)\|^2$, entries of the composite bias vector.involves the entries of the composite bias vector. It is easy to see that this is $\|B^{(1)} - B^{*(1)}\|^2$. We don't have to worry about the other terms in the difference because they do not end up introducing singularities worse than the $B^{(1)}$ term when considering their contribution to the overall singularity at w^*. $\|B^{(1)} - B^{*(1)}\|^2$ involves $H^{(1)}$ independent squared differences of parameters related to the final bias. Therefore, its singularity contributes $H^{(1)}/2$ to λ and 1 to θ.

Since the singularity of $K(w)$ is equivalent to the sum of the singularities of $\|P_L(w) - P_L(w^*)\|^2$ and $\|B^{(1)} - B^{*(1)}\|^2$, their contributions to λ add up, and their contributions to θ combine through an additive rule for multiplicity. Theorem 4 says that the final result of this combination: λ is the sum of $H^{(1)}/2$ (from the final bias) and $\lambda(H^{(1)}, \ldots, H^{(L+1)}, r)$ (from the matrix product), while θ is the same as $\theta(H^{(1)}, \ldots, H^{(L+1)}, r)$ (as the bias term has multiplicity 1).

Additive Rule for ReLU Networks

For neural networks with ReLU units, the function $h_+(x, A, B)$ is piecewise linear. The input space is partitioned into regions where the activation patterns are constant. Within each region, the network function is equivalent to a linear network.

I will now present some peripheral theories.

Theorem 5 essentially provides an additive rule for combining singularities. If a function's singularity can be decomposed into independent components (think different variables or different branches of a singular set), the learning coefficients of the combined function can be calculated by summing or combining the coefficients of the components.

Theorem 6 is set up as follows. Let the $H^{(2)}$ hidden units be divided into $k^{(2)}$ groups, corresponding to different linear regions active near w_0. Let r_i be the rank of the linear map in region i. Theorem 6 states that the λ and θ for the three-layer ReLU network are sums/combinations of the coefficients $\lambda(H_i^{'(1)}, H_i^{(2)}, H^{(3)} + 1, r_i)$ and $\theta(H_i^{'(1)}, H_i^{(2)}, H^{(3)} + 1, r_i)$ calculated for the corresponding linear sub-models in each region:

$$\lambda = \sum_{i=1}^{k^{(2)}} \lambda(H_i'^{(1)}, H_i^{(2)}, H^{(3)} + 1, r_i)$$

$$\theta = \sum_{i=1}^{k^{(2)}} (\theta(H_i'^{(1)}, H_i^{(2)}, H^{(3)} + 1, r_i) - 1) + 1.$$

$H_i^{'(1)}$ is the number of active output units affected by hidden group i, and $H_i^{(2)}$ is the number of hidden units in group i. The result essentially implies that ReLU does not create fundamentally new types of singularities but rather combines existing linear ones.

Example Computation

Just to make these formulations concrete, let us consider a depth-two linear regressor (one hidden layer) with input dimension $d = H^{(3)}$, hidden width $m = H^{(2)}$, scalar output $1 = H^{(1)}$, and true rank $r = 1$.

The dimensions are $H^{(1)} = 1, H^{(2)} = m, H^{(3)} = d$. The rank of the product $A^{(1)}A^{(2)}$ is $r = 1$.

The excess widths are $M^{(1)} = H^{(1)} - r = 1 - 1 = 0$, $M^{(2)} = H^{(2)} - r = m - 1$, $M^{(3)} = H^{(3)} - r = d - 1$.

The bias term $B^{(1)}$ has dimension $H^{(1)} = 1$, which contributes $H^{(1)}/2 = 1/2$ to λ. Using the formulas from Definition 4 and Theorem 4 with these values, the calculation for λ and θ gives us:

$$\lambda = \frac{d}{2} + \frac{m-1}{4}$$

$$\theta = 1$$

This essentially shows us that λ scales linearly with the input dimension d and hidden dimension m for this specific network structure. For instance, increasing input dimension d by 1 increases λ by 1/2, while increasing hidden dimension m by 1 increases λ by only 1/4.

Impact on Learning Curves and Model Selection

These exact computations of λ and θ are incredibly helpful. This is mostly because they provide the precise constants needed for Watanabe's asymptotic formulas (as discussed in Section 1):

$$\mathbb{E}[\text{generalization error at } n] \approx \frac{\lambda}{n} + o(n^{-1})$$

$$-\log p(D_n \mid \text{model}) \approx nL_n(w_0) + \lambda \log n - (\theta - 1) \log \log n + \mathcal{O}(1).$$

We see that λ dictates the dominant $1/n$ rate of decay of the generalization error, and (λ, θ) determine the dominant terms in the asymptotic expansion of the marginal likelihood.

This is really exciting because it provides rigorous support for Wantanabe's predictions. Replacing the classical BIC penalty $\frac{d}{2} \log n$ with $\lambda \log n$ gives us criteria (like WBIC) that are asymptotically unbiased for deep networks, using the correct complexity measure λ instead of the misleading parameter count d.

Conclusion

In conclusion, this paper demonstrates how Singular Learning Theory provides a much more reasonable mathematical framework for understanding generalization in deep neural networks, where classical methods fail due to the singular geometry of their parameter spaces. Watanabe's theory establishes the RLCT(λ) as the important invariant that controls Bayesian generalization error and model evidence. Murfet et al. give good evidence for the singularity of DNNs, show how λ can be estimated in practice. Aoyagi's exact calculations for specific architectures like linear and ReLU networks further confirm the theory, and essentially give us precise values for λ and θ. Some potential avenues for work that I see include extending this framework to more complex architecture such as transformers and also thinking about the preciseness of lambda for other activations which are not exactly linear.

References

[1] Aoyagi, M. (2025). Singular leaning coefficients and efficiency in learning theory. *arXiv preprint*, arXiv:2501.12747.

[2] Murfet, D., Wei, S., Gong, M., Li, H., Gell-Redman, J., & Quella, T. (2020). Deep Learning is Singular, and That's Good. *arXiv preprint*, arXiv:2010.11560.

[3] Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11, 3571–3594.