What Is Singular Learning Theory?

Rumi Salazar (University of Melbourne)

Learning at the Singularity, 24 November 2026

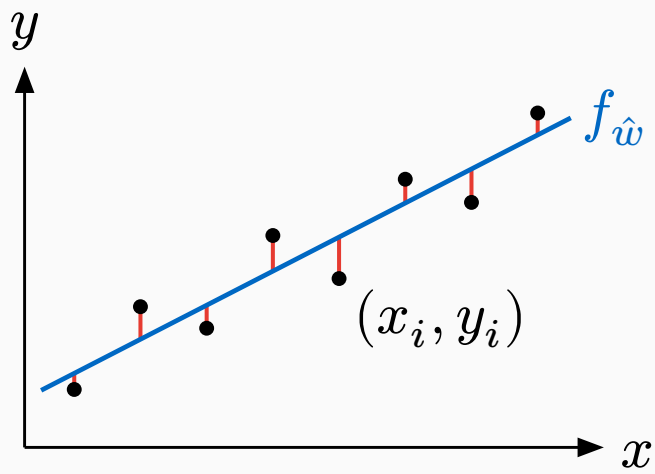
What is learning?

- Data space: inputs $x \in X$ and outputs $y \in Y$. Suppose given a dataset $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- Parameter space $\mathcal{W} \subseteq \mathbb{R}^d$: each parameter w gives a function $f_w : X \to Y$ with some structure (e.g. a neural network).
- Learning (training): find $w \in \mathcal{W}$ minimising a loss function, usually by stochastic gradient descent. E.g. the mean squared error

$$\ell_n(w) = \frac{1}{n} \sum_{i=1}^{n} \big(f_w(x_i) - y_i\big)^2.$$

Example: line of best fit

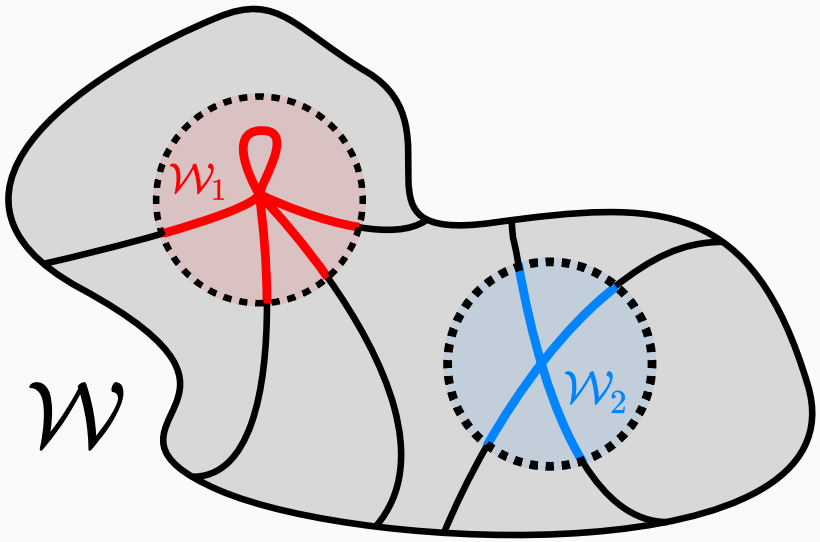
- $X = Y = \mathbb{R}, \quad \mathcal{W} = \mathbb{R}^2$
- $f_w(x) = w_2 x + w_1$
- red: residuals $f_w(x_i) - y_i$



Motivating questions

For a neural network, many parameters approximate the same function to be learned (almost) equally well: these parameters form whole regions of parameter space, not necessarily a single point, and there can be several such regions.

- Which regions does the learning algorithm select?
- Is the region of highest accuracy always selected?
- Does this change as the model sees more data?



Degeneracy

Statistical models

Definition

A statistical model (or parameter-distribution map) is a map

$$\Psi : \mathcal{W} \to \Delta(Z), \qquad w \mapsto p_w = p(\cdot|w)$$

from an open set $\mathcal{W} \subseteq \mathbb{R}^d$ called the parameter space.

We also fix a probability distribution φ on $\mathcal{W}$, called the prior.

Given a true distribution $q \in \Delta(Z)$, the population loss (Kullback-Leibler divergence/relative entropy) is

$$K(w) := D_{\mathrm{KL}}(q \,\|\, p_w) := \int_Z q(z) \log \frac{q(z)}{p(z|w)} \, dz \;\geq 0.$$

Realisable case: $q = p_{w^*}$ for some $w^* \in \mathcal{W}$. SLT studies the set of optimal parameters $\mathcal{W}_0 := \Psi^{-1}(q) = K^{-1}(0) \subseteq \mathcal{W}$, together with *how* K vanishes along it, not just *where*.

Regular and singular models

Definition [Watanabe 2009]

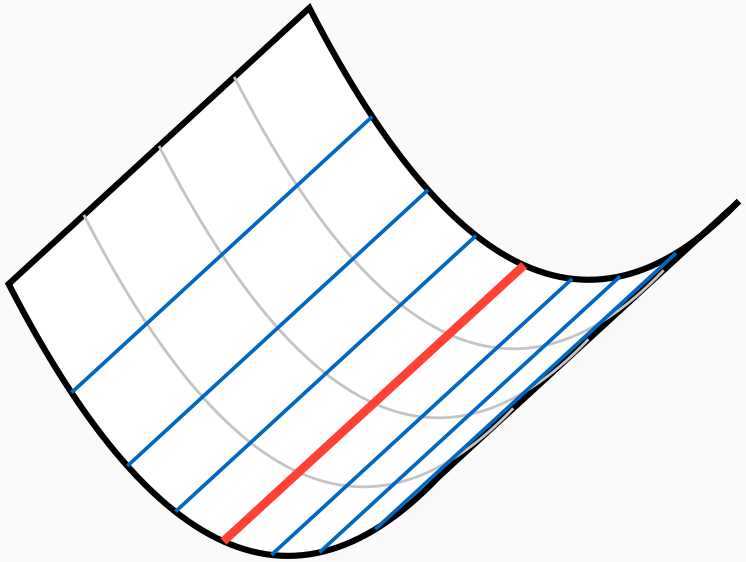
A statistical model $\Psi : \mathcal{W} \to \Delta(Z), w \mapsto p_w$ is regular if it is injective and its derivative is injective:

- Ψ injective (identifiability): $p_w = p_{w'} \Rightarrow w = w'$;
- $D\Psi(w)$ injective for every w: $D_v\, p_w \not\equiv 0$ for $v \neq 0$.

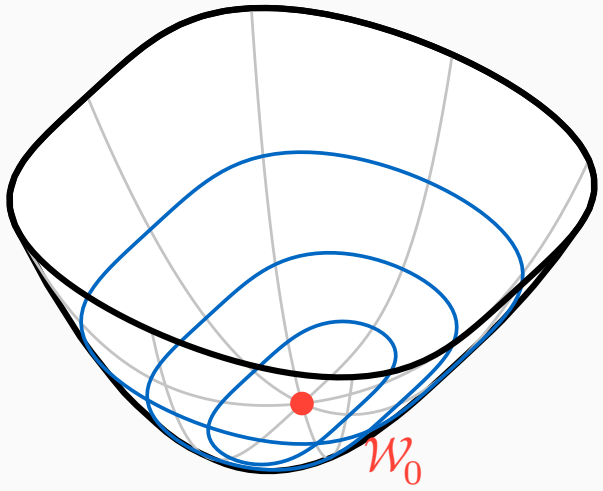
Otherwise it is (strictly) singular.

- Regular, realisable: $\mathcal{W}_0 = \{w^*\}$ and invertible Hessian $\nabla^2 K(w^*)$.
- Singular: $\mathcal{W}_0$ can be an analytic set with singularities, with $\nabla^2 K$ singular on it. We call parameters for which the Hessian is not invertible degenerate.

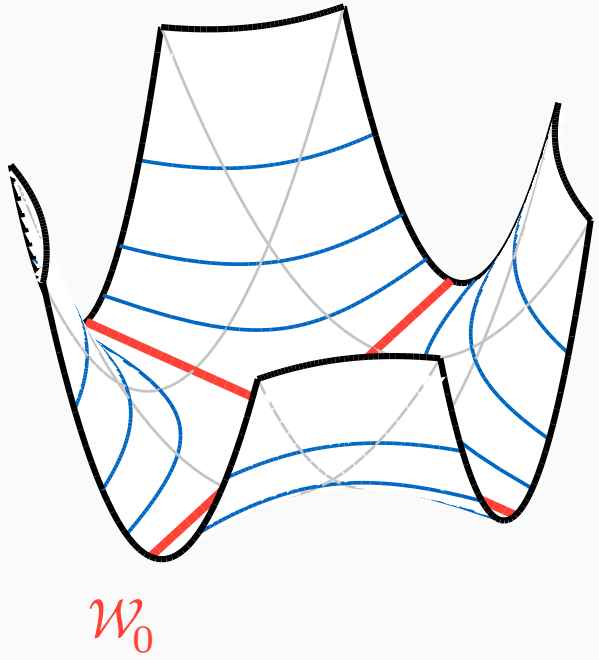
Example: Three degenerate minima



$$K_1(w_1, w_2) = w_1^2$$
$$\nabla^2 K_1(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$



$$K_2(w_1, w_2) = w_1^2 + w_2^4$$
$$\nabla^2 K_2(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$



$$K_3(w_1, w_2) = w_1^2 w_2^2$$
$$\nabla^2 K_3(0,0) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Local learning coefficient: measuring degeneracy by volume

Definition [Watanabe 2009, Lau et al. 2025]

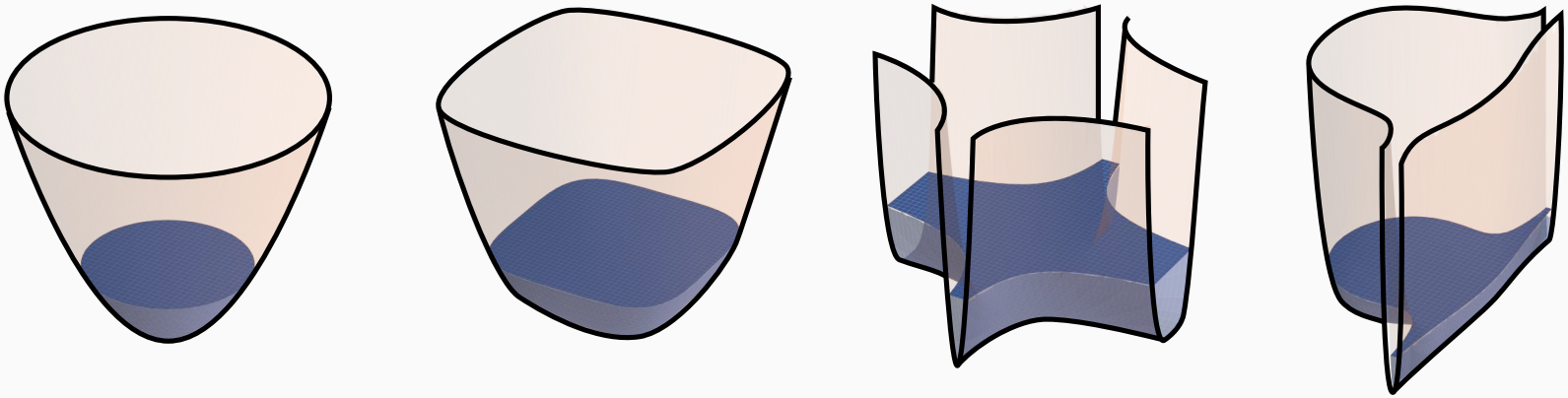
For $w^* \in \mathcal{W}_0$ and a small neighbourhood $\mathcal{U}$ of w^*, the prior-weighted volume of near-optimal parameters is

$$V(\varepsilon) = \int_{\{w \in \mathcal{U} : K(w) < \varepsilon\}} \varphi(w)\, dw,$$
$$\sim c\, \varepsilon^{\lambda(w^*)} (-\log \varepsilon)^{m(w^*)-1} \quad (\varepsilon \to 0).$$

$\lambda(w^*) \in \mathbb{Q}_{>0}$ is the local learning coefficient (LLC) and $m(w^*) \in \mathbb{Z}_{>0}$ its multiplicity.

- In algebraic geometry, $\lambda(w^*)$ is the real log canonical threshold (RLCT) of K at w^*, computed by Hironaka's resolution of singularities.
- Nondegenerate minimum: $\lambda(w^*) = d/2$. Degenerate: $\lambda(w^*) < d/2$.
- Will's talk: an algorithmic interpretation in setting of Turing machines.

Example: Local learning coefficient

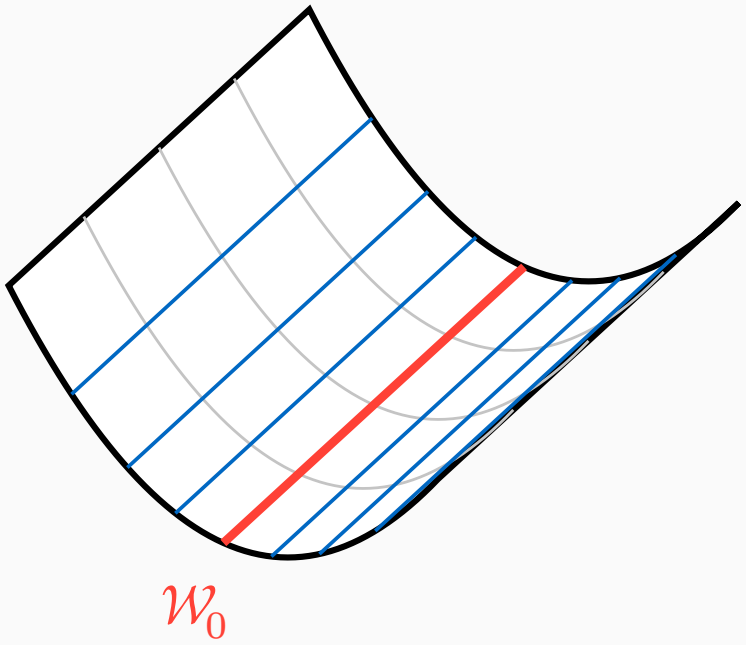


$V(\epsilon) \propto \epsilon^{\lambda}$ ← Local learning coefficient

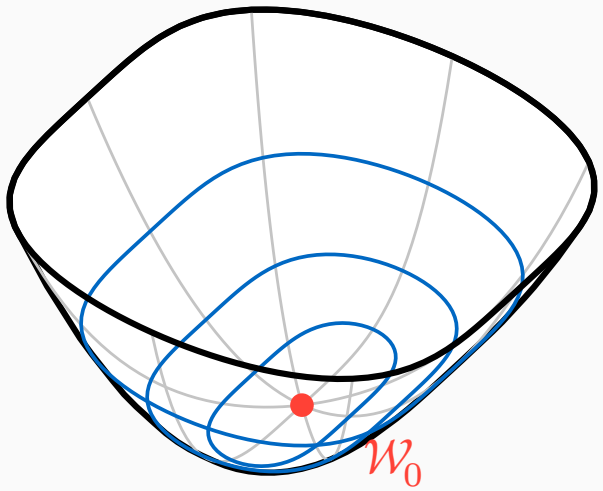
$K(x,y) = x^2 + y^2$	$K(x,y) = x^4 + y^4$	$K(x,y) = x^2 y^4$	$K(x,y) = (x-1)^2(x^2+y^2)^4$
$\lambda(0,0) = 1$	$\lambda(0,0) = \frac{1}{2}$	$\lambda(x,0) = \frac{1}{4}$	$\lambda(0,0) = \frac{1}{4} \quad \lambda(1,y) = \frac{1}{2}$

Image credit: Jesse Hoogland

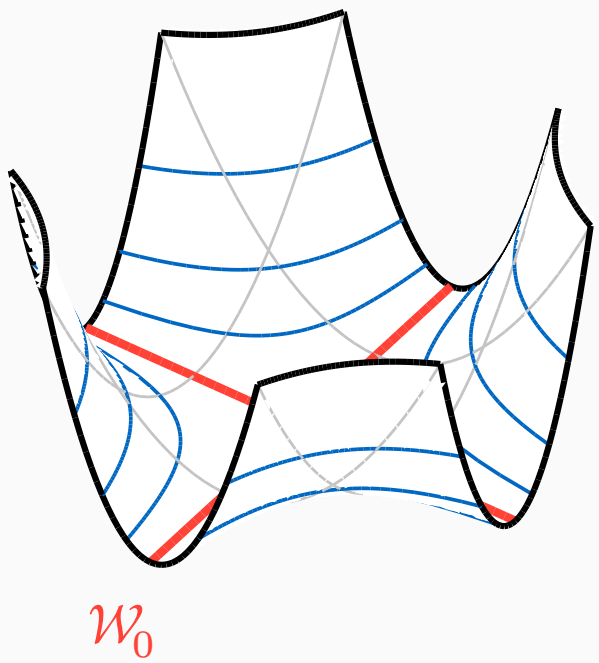
Example revisited: Three degenerate minima



$$K_1(w_1, w_2) = w_1^2$$
$$\nabla^2 K_1(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$
$$\lambda(0, w_2) = 1/2$$
$$m(0, w_2) = 1$$



$$K_2(w_1, w_2) = w_1^2 + w_2^4$$
$$\nabla^2 K_2(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$
$$\lambda(0, 0) = 3/4$$
$$m(0, 0) = 1$$



$$K_3(w_1, w_2) = w_1^2 w_2^2$$
$$\nabla^2 K_3(0,0) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$
$$\lambda(0, 0) = 1/2$$
$$m(0, 0) = 2$$

The (Bayesian) learning problem

The Bayesian posterior

For n i.i.d. samples $D_n = \{z_1, \ldots, z_n\}$ drawn from the (unknown) true distribution $q \in \Delta(Z)$, the likelihood is $p(D_n|w) := \prod_i p(z_i|w)$, and

$$K_n(w) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(z_i)}{p(z_i|w)}$$

is the empirical KL divergence; $K_n(w) \to K(w)$ as $n \to \infty$.

Definition

The posterior and the posterior probability of a region $\mathcal{U} \subseteq \mathcal{W}$:

$$p(w|D_n) := \frac{p(D_n|w)\,\varphi(w)}{p(D_n)} = \frac{e^{-nK_n(w)}\,\varphi(w)}{Z_n(\mathcal{W})},$$
$$p(\mathcal{U}|D_n) := \int_{\mathcal{U}} p(w|D_n)\,dw = \frac{Z_n(\mathcal{U})}{Z_n(\mathcal{W})},$$

where $Z_n(\mathcal{U}) := \int_{\mathcal{U}} e^{-nK_n(w)}\,\varphi(w)\,dw$ is the local partition function.

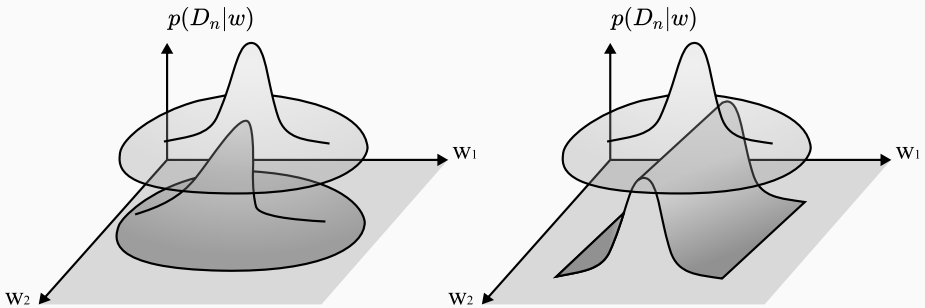
Classical statistical learning theory: regular models

Free energy $F_n(\mathcal{W}) := -\log Z_n(\mathcal{W})$.

Theorem (classical)

Suppose the model is regular and that the truth is realisable by the model (i.e $q \in \text{image}(\Psi)$). As $n \to \infty$:

- **Laplace approximation**: the posterior is asymptotically normal around $w^* \in \mathcal{W}_0$, with covariance $\left(n\nabla^2 K(w^*)\right)^{-1}$;
- **Free energy**: $F_n = \dfrac{d}{2}\log n + O_p(1)$.



Regular Singular

Image credit: Jesse Hoogland

The free energy formula for singular models

[Recall: $F_n(\mathcal{W}) = -\log \int_{\mathcal{W}} e^{-nK_n(w)}\varphi(w)\,dw$]

Main theoretical result of SLT:

Theorem [Watanabe 2009]

In the realisable setting, if $\mathcal{W}$ is compact and $\log p(z|w)$ is real analytic in w (plus integrability conditions), then the **free energy formula** is

$$F_n = \lambda \log n - (m-1)\log\log n + O_p(1).$$

Here $\lambda = \min_{w\in\mathcal{W}_0} \lambda(w)$ is the (global) learning coefficient and m its multiplicity.

So λ takes the place of $d/2$.

- Regular models: $\lambda = d/2$, $m = 1$. In general $\lambda \leq d/2$.
- Also $\log\log n$ term from the multiplicity.

Internal model selection and (Bayesian) phase transitions

Internal model selection

Partition parameter space into phases $\mathcal{W} = \mathcal{W}_1 \sqcup \cdots \sqcup \mathcal{W}_k$, with local minima $w_j^* \in \mathcal{W}_j$ of K:

$$Z_n(\mathcal{W}) = \sum_j Z_n(\mathcal{W}_j), \qquad F_n(\mathcal{W}) = -\log \sum_j e^{-F_n(\mathcal{W}_j)} \approx \min_j F_n(\mathcal{W}_j).$$

The posterior over phases is another Boltzmann distribution/softmax:

$$p(\mathcal{W}_i|D_n) = \frac{Z_n(\mathcal{W}_i)}{Z_n(\mathcal{W})} = \frac{e^{-F_n(\mathcal{W}_i)}}{\sum_j e^{-F_n(\mathcal{W}_j)}}.$$

Free energy formula in each phase (unrealisable setting, i.e. $q \notin \Psi(\mathcal{W}_j)$):

$$F_n(\mathcal{W}_j) \approx \underbrace{K_n(w_j^*)}_{\text{accuracy}} \cdot n + \underbrace{\lambda(w_j^*)}_{\text{simplicity}} \cdot \log n.$$

- The posterior concentrates on the phase of lowest free energy.
- Accuracy versus simplicity: a form of Occam's razor.

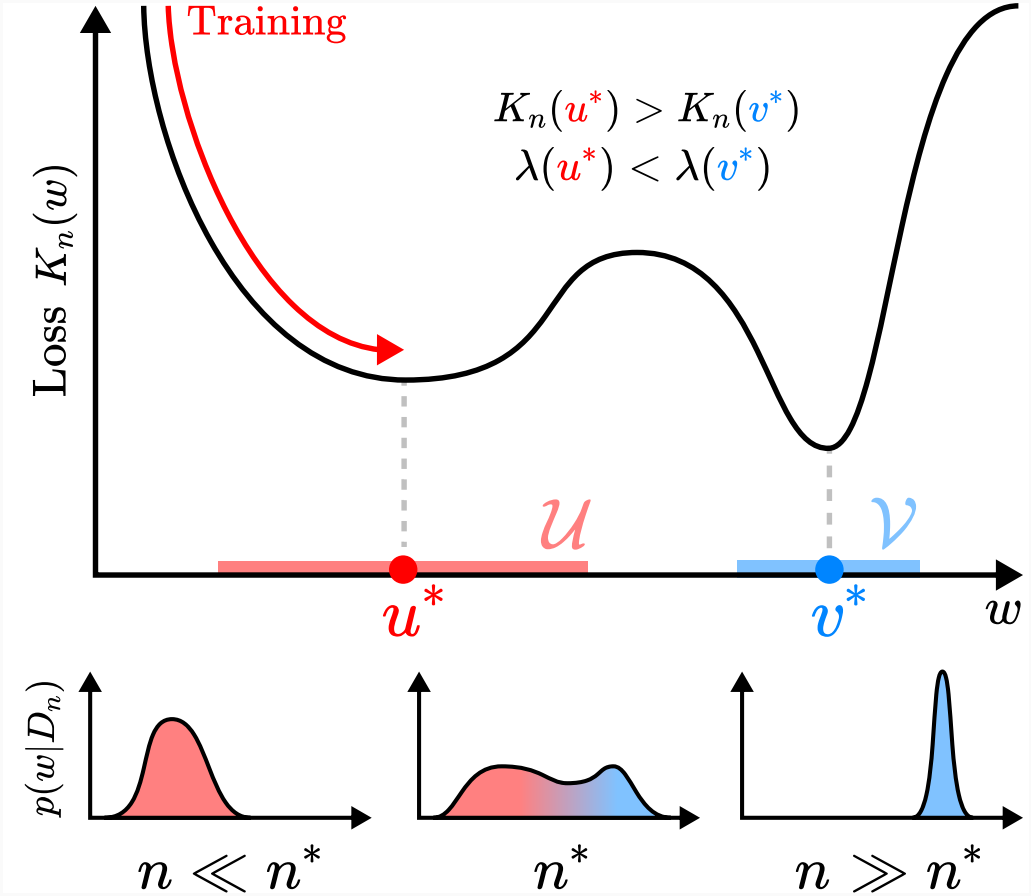
Perfect specification is not enough: phase transitions

[Recall: $p(\mathcal{U}|D_n) = Z_n(\mathcal{U})/Z_n(\mathcal{W})$, $F_n(\mathcal{U}) := -\log Z_n(\mathcal{U}) \approx K_n(u^*)\cdot n + \lambda(u^*)\cdot \log n$.]

$$\log \frac{p(\mathcal{U}|D_n)}{p(\mathcal{V}|D_n)} = F_n(\mathcal{V}) - F_n(\mathcal{U}) \approx \Delta K_n \cdot n + \Delta\lambda \cdot \log n,$$

$$\Delta K_n := K_n(v^*) - K_n(u^*) < 0, \qquad \Delta\lambda := \lambda(v^*) - \lambda(u^*) > 0.$$

Phase transition at n^*: $\dfrac{n^*}{\log n^*} = -\dfrac{\Delta\lambda}{\Delta K_{n^*}} > 0.$



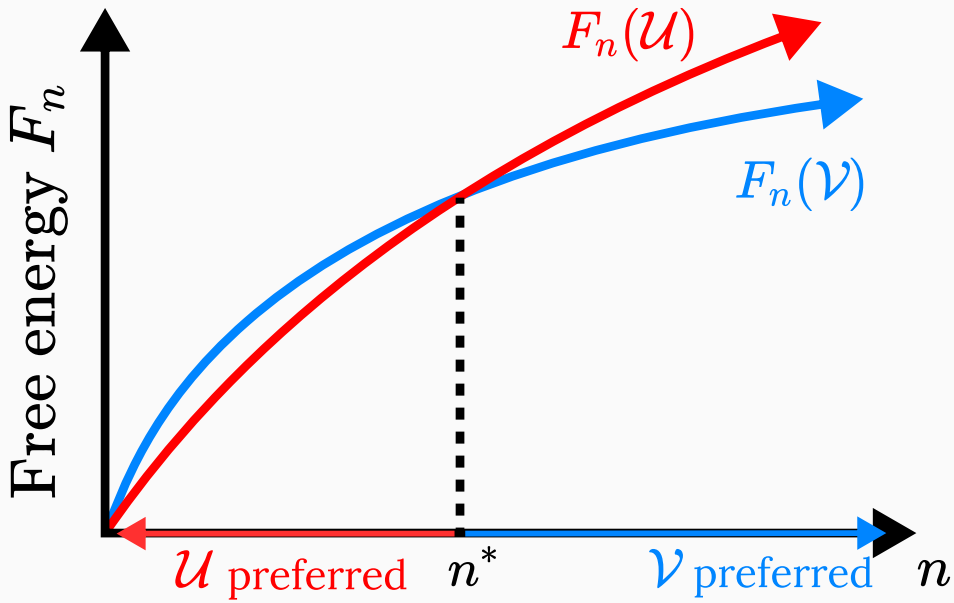
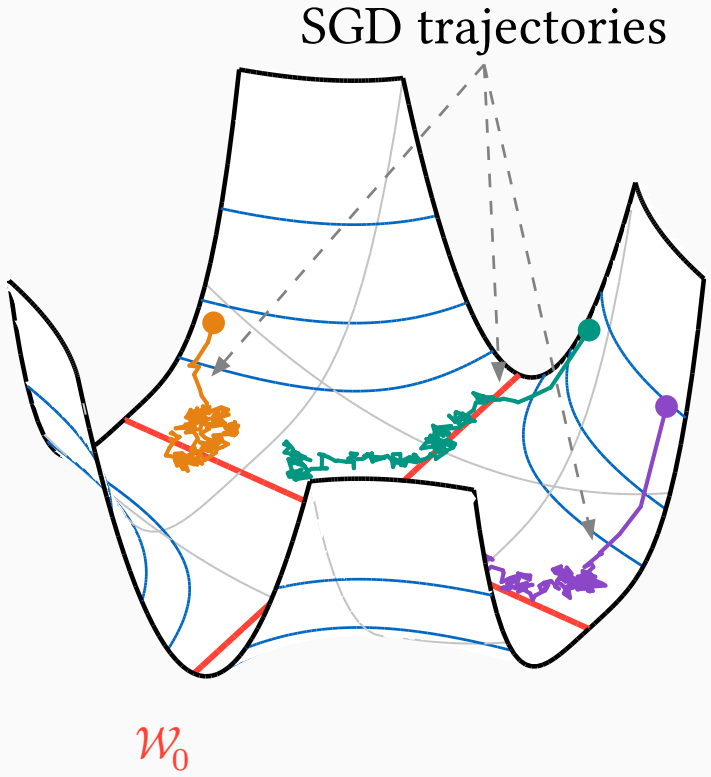


Image credit: Jesse Hoogland

SGD vs Bayes

- Real-world neural networks are trained by stochastic gradient descent (SGD).



Summary

- The local learning coefficient $\lambda(w^*)$ measures degeneracy through the volume of near-optimal parameters, which the Hessian can't detect. It is a measure of model complexity.
- The Bayesian posterior prefers phases that minimise the free energy: low loss first, then small local learning coefficient.
- The preference depends on n: for small n the posterior can prefer a simpler phase with higher loss; for large n it prefers lower loss.
- AI alignment: even with a perfect specification, learning the aligned phase may require a lot more data.

Thank you!

References i


Z. Chen, E. Lau, J. Mendel, S. Wei, D. Murfet.
Dynamical versus Bayesian Phase Transitions in a Toy Model of Superposition. Preprint, 2023.


H. Hironaka.
Resolution of Singularities of an Algebraic Variety over a Field of Characteristic Zero: I, II. Annals of Mathematics 79, 1964.


E. Lau, Z. Furman, G. Wang, D. Murfet, S. Wei.
The Local Learning Coefficient: A Singularity-Aware Complexity Measure. AISTATS 2025.

References ii

- S. Pepin Lehalleur, J. Hoogland, M. Farrugia-Roberts, S. Wei, A. Gietelink Oldenziel, G. Wang, L. Carroll, D. Murfet.
  You Are What You Eat: AI Alignment Requires Understanding How Data Shapes Structure and Generalisation. Preprint, 2025.
- S. Watanabe.
  Algebraic Geometry and Statistical Learning Theory. Cambridge University Press, 2009.
- S. Watanabe.
  Mathematical Theory of Bayesian Statistics. CRC Press, 2018.

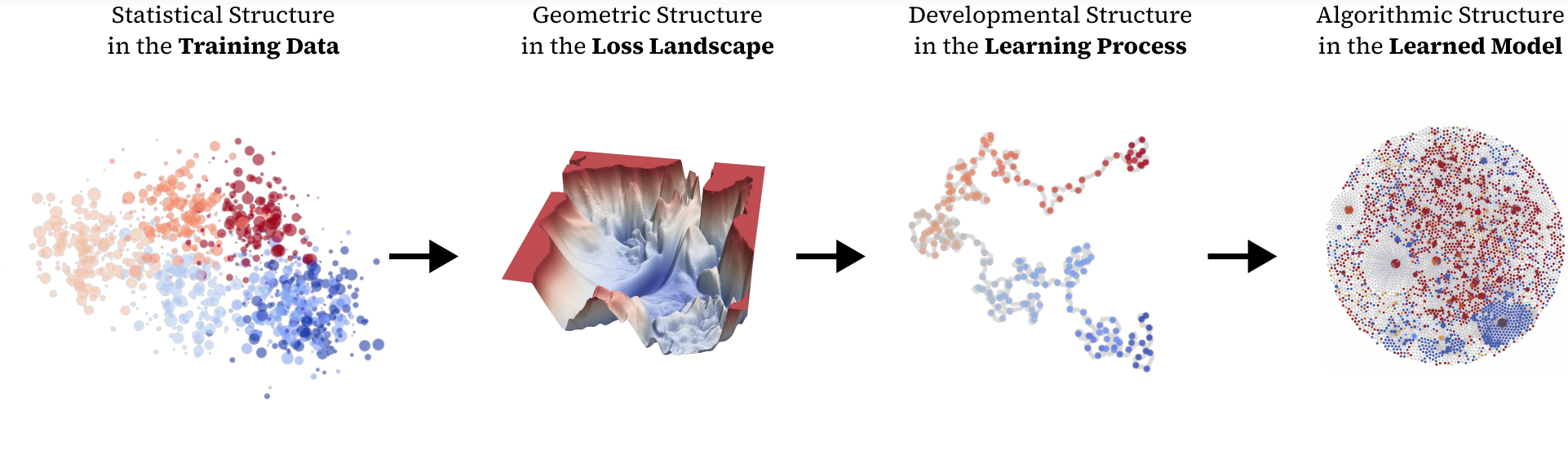
Extras

What Is Singular Learning Theory?

Some possible answers:

- A mathematical theory of Bayesian statistics: Title of Watanabe's second book on SLT.
- The physics of intelligence: Inspiring, but SLT doesn't talk about training dynamics.
- The equilibrium thermodynamics of intelligence: Closer to the truth.
- Algebraic geometry applied to machine learning: There is also algebraic statistics and neuroalgebraic geometry.

How does structure in the data shape structure in the model?



- Loss landscape: the graph of the loss function $\ell_n : \mathcal{W} \to \mathbb{R}$.
- Learning process: e.g. the trajectory of stochastic gradient descent through the loss landscape during training, or how the best-fitting parameters change as the dataset size increases.
- Algorithmic structure: e.g. circuits for specialised tasks inside a large neural network.

Example: regression with Gaussian noise

Take $Z = X \times Y$ with $Y = \mathbb{R}^N$, inputs drawn from $q(x)$, and Gaussian noise with covariance $\sigma^2 I$ around f_w:

$$p(y|x,w) = (2\pi\sigma^2)^{-N/2} \exp\Big(-\frac{1}{2\sigma^2}\|y - f_w(x)\|^2\Big).$$

Suppose the true distribution $q(x,y)$ comes from a target function $f^* = f_{w^*}$ with the same noise. Then the population loss is

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

Local free energy formula: a heuristic argument

Definition

The local free energy of a region $\mathcal{U}$ is $F_n(\mathcal{U}) := -\log Z_n(\mathcal{U})$.

Recall: For $w^* \in \mathcal{W}_0$ and $\mathcal{U}$ a small neighbourhood of w^*,

$$V(\varepsilon) = \int_{\{w \in \mathcal{U} \,:\, K(w) < \varepsilon\}} \varphi(w)\, dw \;\sim\; c\, \varepsilon^{\lambda(w^*)} (-\log \varepsilon)^{m(w^*)-1} \quad (\varepsilon \to 0).$$

For large n, $K_n \approx K$, and $e^{-nK(w)}$ is close to 1 where $K(w) \lesssim 1/n$ and close to 0 elsewhere. So, setting $\varepsilon = 1/n$,

$$\begin{aligned} Z_n(\mathcal{U}) &= \int_{\mathcal{U}} e^{-nK_n(w)}\, \varphi(w)\, dw \\ &\approx V(1/n) \sim c\, n^{-\lambda(w^*)} (\log n)^{m(w^*)-1}, \end{aligned}$$

$$F_n(\mathcal{U}) \approx \lambda(w^*) \log n - \big(m(w^*) - 1\big) \log\log n + \text{const.}$$

So $\lambda(w^*)$ takes the place of $d/2$.

Open problems

- Bridging SGD and Bayesian learning: SLT currently describes the Bayesian posterior, not the training dynamics of SGD.
- Computing learning coefficients analytically for neural network architectures (MLPs, CNNs, ResNets, transformers).
- Probably intractable in full generality, so results will need assumptions on the true distribution q.