Free Energy Formula

Singular Learning Theory Lecture 4

1 Introduction **2**

2 Bayesian learning **4**
2.1 Statistical models and parameter–distribution maps 4
2.2 Bayesian deep learning 6
2.3 Local partition functions and local free energies 7

3 The free energy formula **10**
3.1 Watanabe's free energy formula 10
3.2 Deriving the free energy formula 10
3.3 The generalised BIC 13
3.4 Internal model selection 14

4 Bayesian phase transitions **16**

References and further reading **18**

A Solutions **21**
A.1 Solutions to exercises from Section 2 22
A.2 Solutions to exercises from Section 3 27
A.3 Solutions to exercises from Section 4 28

1 Introduction

In Lectures 1 and 2, we studied the geometry of degeneracy in parameter space and introduced the local learning coefficient $\lambda(w^*)$ as a measure of the degree of degeneracy at a local minimum w^*. In this lecture we show that this geometry has consequences for learning (in the Bayesian setting). In particular, it creates a trade-off between model fit and model complexity that drives learning, providing a mechanism for *internal model selection*.

Contents. The lecture comprises the following three technical sections, each with definitions and guided exercises.

1. "Bayesian learning" moves from function approximation to statistical inference via parameter–distribution maps, and introduces Bayesian learning: the posterior, the partition function, and the (local) free energy.
2. "The free energy formula" states Watanabe's free energy formula connecting the local learning coefficient to Bayesian learning, sketches its derivation from the volume scaling of Lecture 2, and discusses its implications.
3. "Bayesian phase transitions" studies how the posterior's preferred region of parameter space can change as the amount of data grows.

Section A provides worked solutions for all exercises.

Recap: loss functions. We recall some notation from Lecture 1. Given an example $(x, y) \in \mathcal{X} \times \mathcal{Y}$, a **per-example loss** function $L_{x,y} : \mathcal{W} \to \mathbb{R}$ measures the deviation of $f_w(x)$ from the output y. For vector outputs $\mathcal{Y} \subset \mathbb{R}^m$, a typical choice is the **squared error** loss

$$L_{x,y}(w) = \|f_w(x) - y\|^2. \tag{1}$$

If $\mathcal{Y} = \Delta^{m-1}$ is a space of discrete distributions over m objects it is typical to use **cross entropy** loss

$$L_{x,y}(w) = -\sum_{i=1}^{m} y(i) \log f_w(i \mid x). \tag{2}$$

Given a data distribution $q \in \Delta(\mathcal{X} \times \mathcal{Y})$, the **population loss** is $L(w) = \mathbb{E}_{(x,y)\sim q}[L_{x,y}(w)]$, and given a data set of n examples sampled independently from q, the **empirical loss** is

$$L_n(w) = \frac{1}{n} \sum_{i=1}^{n} L_{x^{(i)},y^{(i)}}(w).$$

If the per-example loss is squared error, the empirical loss is known as the **mean squared error** objective,

$$L_n(w) = \frac{1}{n} \sum_{i=1}^{n} \left\| f_w(x^{(i)}) - y^{(i)} \right\|^2. \tag{3}$$

Recap: the local learning coefficient. As in Lecture 2, we assume that any population loss L is a real analytic function. Let $w^* \in \mathcal{W}$ be a local minimum of L, and let $B(w^*)$ be a sufficiently small closed ball centred on w^* such that $L(w) \geq L(w^*)$ for all $w \in B(w^*)$. Given $\epsilon > 0$, the **sublevel set** and its volume are

$$B(w^*, \epsilon) = \{w \in B(w^*) \mid L(w) - L(w^*) < \epsilon\}, \qquad V(\epsilon) = \int_{B(w^*,\epsilon)} dw. \tag{4}$$

Definition 1.1 (Local learning coefficient)**.** Let $w^* \in \mathcal{W}$ be a local minimum of the population loss L. There exists a unique rational number $\lambda(w^*) > 0$, a positive integer $\mu(w^*)$, and a constant $c > 0$ such that as $\epsilon \to 0$,

$$V(\epsilon) = c\,\epsilon^{\lambda(w^*)}(-\log \epsilon)^{\mu(w^*)-1} + o\big(\epsilon^{\lambda(w^*)}(-\log \epsilon)^{\mu(w^*)-1}\big)\,. \tag{5}$$

We call $\lambda(w^*)$ the **local learning coefficient (LLC)** at w^*, and $\mu(w^*)$ the **local multiplicity**.

Recall from Lecture 2 that if the Hessian $\nabla^2 L(w^*)$ is positive definite then $\lambda(w^*) = d/2$ and $\mu(w^*) = 1$, where $d = \dim \mathcal{W}$, and that in general $\lambda(w^*) \leq d/2$.

2 Bayesian learning

In this section, we move from viewing deep learning as function approximation to viewing it as statistical inference and introduce the Bayesian framework in which the main results of SLT are stated.

2.1 Statistical models and parameter–distribution maps

Lecture 1 introduced neural networks and supervised deep learning within a framework of function approximation. An alternative approach is to view supervised learning as statistical inference, and particularly Bayesian inference. This statistical framework is the setting for the main theorems of SLT which we will discuss in this lecture note.

The first step to upgrading our framework is to move from talking about functions to talking about probability distributions. For each neural network, we should be able to determine not just the output corresponding to each input, but a probability density measuring how likely each output is given each input.

Formally, let $\mathcal{D} \subseteq \Delta(\mathcal{Y})^{\mathcal{X}}$ represent a class of conditional distributions over $\mathcal{Y}$. A distributional neural network architecture comprises a **parameter–distribution map** $\Psi : \mathcal{W} \to \mathcal{D}$. It is sometimes convenient to refer to the conditional distribution $\Psi(w) : \mathcal{X} \to \Delta(\mathcal{Y})$ using the notation $p_w : \mathcal{X} \to \Delta(\mathcal{Y})$. The probability mass/density of a given output $y \in \mathcal{Y}$ conditional on a given input $x \in \mathcal{X}$ is typically denoted either $p_w(y \mid x)$ or $p(y \mid x, w)$.

Sometimes, defining a parameter–distribution map is simply a re-framing. For example, in image classification, our output space $\mathcal{Y}$ was a space of distributions over image classes. Let $\mathcal{C}$ represent the underlying class space, then $\mathcal{Y} = \Delta(\mathcal{C})$ and we can view our parameter–function map $\Phi : \mathcal{W} \to \mathcal{F}$ as a parameter–distribution map $\Psi : \mathcal{W} \to \mathcal{D}$ with $\mathcal{D} = \mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}} = \Delta(\mathcal{C})^{\mathcal{X}}$.

In other cases, we can easily construct a parameter–distribution map from a parameter–function map by introducing a **noise model**. For example, in regression with an output space $\mathcal{Y} = \mathbb{R}^m$, given a parameter–function map $w \mapsto f_w$ we can define a parameter–distribution map $\Psi : \mathcal{W} \to \Delta(\mathcal{Y})^{\mathcal{X}}$ such that, for $w \in \mathcal{W}$ and $x \in \mathcal{X}$,

$$p(y \mid x, w) = \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{\|y - f_w(x)\|^2}{2}\right). \tag{6}$$

The above **Gaussian noise model** amounts to modelling each output y as drawn given x, w according to the rule $y = f_w(x) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I_m)$.

As the following exercises explore, performing statistical inference according to the principle of maximum likelihood aligns with the loss-minimisation perspective (given an appropriate choice of loss function).

Exercise 2.1 (Maximum likelihood as MSE minimisation)**.** Consider a regression problem with $\mathcal{X} = \mathcal{Y} = \mathbb{R}^m$. Assume we have a neural network architecture with a parameter space $\mathcal{W}$. Form a corresponding statistical model using the Gaussian noise model of (6). Let $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathcal{X} \times \mathcal{Y}$ be a data set of n regression examples.

(a) Write an expression for the conditional probability of observing the labels $Y_n = (y^{(1)}, \ldots, y^{(n)})$ given the inputs $X_n = (x^{(1)}, \ldots, x^{(n)})$ and some fixed parameter $w \in$

$\mathcal{W}$, assuming the labels are sampled independently.

(b) Denote the resulting quantity by $p(Y_n \mid X_n, w)$. It is called the **likelihood** of the data set according to the model w. Compute and simplify the **negative log-likelihood**, given by $-\log p(Y_n \mid X_n, w)$.

(c) Show that if L_n is the mean squared error loss function from (3), then

$$\arg\max_{w \in \mathcal{W}} p(Y_n \mid X_n, w) = \arg\min_{w \in \mathcal{W}} L_n(w).$$

Exercise 2.2 (Maximum likelihood as cross entropy minimisation)**.** Consider a classification problem with $\mathcal{Y} = \Delta^{m-1}$. Assume we have a neural network architecture with a parameter space $\mathcal{W}$. As discussed, the parameter–function map in this case corresponds to a parameter–distribution map to conditional distributions over the underlying discrete class space $\mathcal{C} = \{1, \ldots, m\}$. Let $(x^{(1)}, c^{(1)}), \ldots, (x^{(n)}, c^{(n)}) \in \mathcal{X} \times \mathcal{C}$ be a data set of n labelled classification examples. For each $c^{(i)} \in \mathcal{C}$ define

$$y^{(i)} = \delta_{c^{(i)}} \in \Delta^{m-1} \tag{7}$$

to be the corresponding Dirac distribution (with probability mass 1 concentrated on $c^{(i)}$).

(a) Write an expression for the likelihood of the data set, $p(C_n \mid X_n, w)$, that is, the conditional probability of observing the labels $C_n = (c^{(1)}, \ldots, c^{(n)})$ given the inputs $X_n = (x^{(1)}, \ldots, x^{(n)})$ and some fixed parameter $w \in \mathcal{W}$. Assume the labels are sampled independently.

(b) Show that the negative log-likelihood

$$-\log p(C_n \mid X_n, w) = -\sum_{i=1}^{n} \log f_w(c^{(i)} \mid x^{(i)}).$$

(c) Using (7), show that if the per-example loss function for the optimisation problem is cross entropy loss (2), then

$$\arg\max_{w \in \mathcal{W}} p(C_n \mid X_n, w) = \arg\min_{w \in \mathcal{W}} L_n(w).$$

Exercise 2.3 (Maximum likelihood as loss minimisation, in general)**.** Consider a neural network architecture with a parameter space $\mathcal{W}$, parameter–function map Φ, and output space $\mathcal{Y} = \mathbb{R}^m$. Let $L_{x,y} : \mathcal{W} \to \mathbb{R}$ be a per-example loss function. Assume that for each $x \in \mathcal{X}$ and $w \in \mathcal{W}$, the function $y \mapsto \exp(-L_{x,y}(w))$ is integrable over $\mathcal{Y}$. Define a corresponding statistical model with parameter–distribution map Ψ mapping $w \in \mathcal{W}$ to $p_w : \mathcal{X} \to \Delta(\mathcal{Y})$ such that for all $y \in \mathcal{Y}$, $x \in \mathcal{X}$, and $w \in \mathcal{W}$, we have

$$p(y \mid x, w) = \frac{\exp(-L_{x,y}(w))}{Z(x, w)}, \qquad Z(x, w) = \int_{\mathcal{Y}} \exp(-L_{x,z}(w))\, dz.$$

We re-use the notation X_n, Y_n from Exercise 2.1, and assume that labels are sampled independently.

(a) Show that the negative log-likelihood decomposes as

$$-\log p(Y_n \mid X_n, w) = \sum_{i=1}^{n} L_{x^{(i)},y^{(i)}}(w) + \sum_{i=1}^{n} \log Z(x^{(i)}, w).$$

(b) Now suppose the loss function takes the form $L_{x,y}(w) = \ell(f_w(x) - y)$ for some integrable function $\ell : \mathbb{R}^m \to \mathbb{R}$. Show that $Z(x, w)$ is independent of w, and conclude that

$$\arg\max_{w \in \mathcal{W}} p(Y_n \mid X_n, w) = \arg\min_{w \in \mathcal{W}} L_n(w).$$

Remark 2.1 (Negative log-likelihood loss)**.** Exercises 2.1 to 2.3 nominally show that some maximum likelihood estimation problems have the same sets of ideal solutions as some carefully-chosen optimisation problems ($\arg\min = \arg\max$). In general, that the global optima for two optimisation problems coincide is insufficient to show that practical optimisation methods will always find the same solutions: practical optimisation methods don't always find global optima. However, in the above cases, your derivation likely reveals a stronger result: the optimisation landscapes for the function approximation problem are related to the optimisation landscape for the maximum likelihood estimation problem by a strictly monotonically decreasing transform (an affine transform of a logarithm).

2.2 Bayesian deep learning

The principle of maximum likelihood, explored in the preceding exercises, is one approach to statistical inference. An alternative approach is **Bayesian inference**, which treats w as a random variable and maintains a probability distribution over $\mathcal{W}$ reflecting our uncertainty about which parameter best explains the data. The majority of theoretical results in SLT have been developed within this setting (cf., Watanabe, 2009, 2018).

Bayesian learning begins with a **prior** distribution $\varphi \in \Delta(\mathcal{W})$, a probability density on $\mathcal{W}$ representing our beliefs about the parameter w before observing any data. Throughout these notes, we assume φ is positive and smooth on $\mathcal{W}$.

Definition 2.2 (Posterior distribution, partition function, free energy)**.** Given a data set of n input–output pairs $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathcal{X} \times \mathcal{Y}$, Bayes' rule updates the prior into the **posterior distribution** $\pi_n \in \Delta(\mathcal{W})$, a probability density on $\mathcal{W}$ defined by

$$\pi_n(w) = \frac{1}{Z_n} \varphi(w) \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}, w) \tag{8}$$

where $Z_n \in \mathbb{R}$ is a normalising constant called the **partition function** (or **marginal likelihood** or **model evidence**),

$$Z_n = \int_{\mathcal{W}} \varphi(w) \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}, w)\, dw. \tag{9}$$

We often work with the **(Bayesian) free energy**, the negative log of the model evidence,

$$F_n = -\log Z_n. \tag{10}$$

The posterior re-weights the prior by the likelihood of the observed data: parameters w under which the data is more probable receive higher posterior density, while parameters under which the data is improbable are down-weighted.

The partition function Z_n is the marginal probability of observing the data set, averaged over all parameters weighted by the prior. It quantifies how well the statistical model *as a whole* predicts the observed data. The free energy is an inverted measure of how well the model fits the data.

Exercise 2.4 (Bayesian posterior as a Gibbs distribution)**.** To interpret Bayesian inference in terms more similar to function approximation, we introduce a loss function based on the **negative log-likelihood**,

$$L_{x,y}(w) = -\log p(y \mid x, w), \tag{11}$$

$$L_n(w) = -\frac{1}{n}\sum_{i=1}^{n} \log p(y^{(i)} \mid x^{(i)}, w). \tag{12}$$

Here, $p(y \mid x, w)$ is the conditional density from the parameter–distribution map and $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathcal{X} \times \mathcal{Y}$ is a data set of n examples. Show that

$$\pi_n(w) = \frac{1}{Z_n}\varphi(w)\exp\bigl(-nL_n(w)\bigr) \qquad \text{and} \qquad Z_n = \int_{\mathcal{W}} \varphi(w)\exp\bigl(-nL_n(w)\bigr)\,dw.$$

The **Bayesian deep learning process** is the process of re-weighting each individual parameter vector in response to seeing an increasing amount of data, producing the sequence of belief distributions $\varphi, \pi_1, \pi_2, \ldots \in \Delta(\mathcal{W})$. Over time, the posterior will concentrate around parameters which provide a good fit for the data (though there is more to the story in the singular case, as we will see in later sections).

Bayesian deep learning is a *global search* in that all parameter vectors are considered in parallel. This makes exact Bayesian learning computationally intractable for non-trivial models, but more analytically tractable than SGD in general. Moreover, as we will explore in later sections, the dynamics of Bayesian learning still reflect the local geometry of the parameter–distribution map.

2.3 Local partition functions and local free energies

In preparation for exploring this connection between geometry and learning we finally introduce localised variants of the partition function and free energy, where the integral in (9) is restricted to a certain neighbourhood in parameter space.

Definition 2.3 (Local partition function and local free energy)**.** Given a neighbourhood $\mathcal{U} \subseteq \mathcal{W}$, the **local partition function** and **local free energy** are

$$Z_n(\mathcal{U}) = \int_{\mathcal{U}} \varphi(w) \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}, w)\,dw, \qquad F_n(\mathcal{U}) = -\log Z_n(\mathcal{U}). \tag{13}$$

The local partition function (free energy) measures how well (poorly) parameters within a given neighbourhood collectively explain the data. Note, $Z_n(\mathcal{W}) = Z_n$ and $F_n(\mathcal{W}) = F_n$.

Exercise 2.5 (Local posterior mass and local free energy)**.** Let $\mathcal{U}, \mathcal{V} \subseteq \mathcal{W}$ be two neighbourhoods of parameter space. Let $\pi_n(\mathcal{U}) = \int_{\mathcal{U}} \pi_n(w)\, dw$ denote the posterior mass of a neighbourhood. Show the following.

(a) $\pi_n(\mathcal{U}) = \dfrac{Z_n(\mathcal{U})}{Z_n}$ (the same holds for $\mathcal{V}$).

(b) $\log \dfrac{\pi_n(\mathcal{U})}{\pi_n(\mathcal{V})} = F_n(\mathcal{V}) - F_n(\mathcal{U})$.

In conclusion, the posterior odds ratio of the two regions $\pi_n(\mathcal{U})/\pi_n(\mathcal{V})$ depends exponentially on the difference between their local free energies.

Exercise 2.6 (Partition function of partitioned parameter space)**.** Let $\mathcal{U}_1, \dots, \mathcal{U}_k \subseteq \mathcal{W}$ be disjoint subsets with $\mathcal{W} = \mathcal{U}_1 \sqcup \cdots \sqcup \mathcal{U}_k$.

(a) Show that the partition function decomposes as $Z_n = \sum_{j=1}^{k} Z_n(\mathcal{U}_j)$, and hence

$$F_n = -\log \sum_{j=1}^{k} e^{-F_n(\mathcal{U}_j)}.$$

(b) Show that if $F_n(\mathcal{U}_1) < F_n(\mathcal{U}_j)$ for all $j \neq 1$, then

$$F_n(\mathcal{U}_1) - \log k < F_n < F_n(\mathcal{U}_1).$$

In conclusion, the overall free energy is dominated by the region(s) with the lowest local free energy.

Exercise 2.7 (Deriving the BIC via the Laplace approximation)**.** In this exercise we derive an asymptotic expansion of the free energy under a non-degeneracy assumption on the Hessian of the population loss.
Suppose the model is regular at a minimiser w^* of L, that is $H = \nabla^2 L(w^*)$ is positive definite, and let $\mathcal{U}$ be a neighbourhood of w^* containing no other minimiser of L. Assume L_n is smooth, and let $\hat{w}_n$ denote a minimiser of L_n in $\mathcal{U}$ with empirical Hessian $H_n = \nabla^2 L_n(\hat{w}_n)$. You may assume that $\hat{w}_n \to w^*$ and $H_n \to H$ as $n \to \infty$, so that for large n, $\hat{w}_n$ lies in the interior of $\mathcal{U}$ and H_n is positive definite.

(a) Expand $L_n(w)$ to second order and $\varphi(w)$ to zeroth order about $\hat{w}_n$ to show that, as $n \to \infty$,

$$Z_n(\mathcal{U}) \approx \varphi(\hat{w}_n)\, e^{-nL_n(\hat{w}_n)} \int_{\mathcal{U}} \exp\Big(-\frac{n}{2}(w - \hat{w}_n)^\top H_n (w - \hat{w}_n)\Big)\, dw.$$

(b) Argue that replacing $\mathcal{U}$ by $\mathbb{R}^d$ in the integral from part (a) introduces only an exponentially small error in n, and show that

$$\int_{\mathcal{U}} \exp\Big(-\frac{n}{2}(w - \hat{w}_n)^\top H_n (w - \hat{w}_n)\Big)\, dw \approx \left(\frac{2\pi}{n}\right)^{d/2} (\det H_n)^{-1/2}.$$

[*Hint:* You may use standard results about Gaussian integrals.]

(c) Deduce that as $n \to \infty$,

$$\begin{aligned} F_n(\mathcal{U}) &= nL_n(\hat{w}_n) + \frac{d}{2}\log n + \frac{1}{2}\log\det H_n - \frac{d}{2}\log 2\pi - \log\varphi(\hat{w}_n) + o_p(1) \\ &= nL_n(\hat{w}_n) + \frac{d}{2}\log n + O_p(1). \end{aligned}$$

Note that you do not need to rigorously justify the $o_p(1)$ remainder.

Challenge: repeat parts (a)–(c) tracking the remainder terms.

(d) Suppose instead that $\nabla^2 L(w^*)$ is not positive definite. Identify where the above derivation breaks down.

3 The free energy formula

Recall the Bayesian learning setup from Section 2.2. We have a parameter–distribution map $\Psi : \mathcal{W} \to \mathcal{D}$ sending each parameter w to a conditional distribution with density $p(y \mid x, w)$, a smooth positive prior φ on $\mathcal{W}$, and a data set of n i.i.d. examples from a data distribution q. Take the loss function to be the negative log-likelihood

$$L_n(w) = -\frac{1}{n}\sum_{i=1}^{n} \log p(y^{(i)} \mid x^{(i)}, w), \qquad L(w) = \mathbb{E}_{(x,y)\sim q}[-\log p(y \mid x, w)]\,.$$

The posterior π_n is given by Bayes' rule (8) and as we saw in Exercises 2.5 and 2.6 it concentrates around regions of parameter space with the lowest local free energy $F_n(\mathcal{U})$ (Definition 2.3).

3.1 Watanabe's free energy formula

We now state one of the main results of SLT: an asymptotic expansion of $F_n(\mathcal{U})$ as $n \to \infty$. Under certain technical assumptions on the statistical model (Watanabe, 2018; see Lau, 2025, §2.3.1 for a summary), we have the following theorem.

Theorem 3.1 (Local free energy formula)**.** *Suppose $\mathcal{U} \subseteq \mathcal{W}$ contains at least one minimiser of L, and let $w^* \in \mathcal{U}$ be any such minimiser. Define $\lambda_{\mathcal{U}}$ as the smallest local learning coefficient among all minimisers in $\mathcal{U}$, and let $\mu_{\mathcal{U}}$ be the maximum local multiplicity associated with $\lambda_{\mathcal{U}}$. Then, as $n \to \infty$,*

$$F_n(\mathcal{U}) = nL_n(w^*) + \lambda_{\mathcal{U}} \log n - (\mu_{\mathcal{U}} - 1)\log\log n + O_p(1). \tag{14}$$

Remark 3.2. Since $L_n(w^*)$ is a random variable, so is $F_n(\mathcal{U})$. The expansion (14) is an asymptotic statement about random variables where the remainder $O_p(1)$ denotes a term that is bounded in probability as $n \to \infty$.

3.2 Deriving the free energy formula

In this subsection we sketch where Theorem 3.1 comes from. The derivation splits the local free energy into two parts.

1. A *deterministic* part, depending only on the population loss and the prior. We analyse this part completely, using the volume scaling of Definition 1.1. This is where the LLC and the multiplicity enter.
2. A *stochastic* part, capturing the fluctuations of the empirical loss around the population loss. We treat this part as a black box and refer to Watanabe (2009) for full detail.

Throughout, fix $\mathcal{U}$ and w^* as in Theorem 3.1, and write $\lambda = \lambda_{\mathcal{U}}$ and $\mu = \mu_{\mathcal{U}}$. We assume w^* minimises L over $\mathcal{U}$, so that $L(w) \geq L(w^*)$ for all $w \in \mathcal{U}$ (as in Definition 1.1, we may take $\mathcal{U}$ to be a small closed neighbourhood of the minimisers).

Define the **excess population loss** and **excess empirical loss** relative to w^*,

$$K(w) = L(w) - L(w^*), \qquad K_n(w) = L_n(w) - L_n(w^*). \tag{15}$$

Note that $K \geq 0$ on $\mathcal{U}$ with $K(w^*) = 0$.

Using the Gibbs form of the posterior (Exercise 2.4) and factoring the minimum-loss term $e^{-nL_n(w^*)}$ out of the local partition function gives

$$Z_n(\mathcal{U}) = e^{-nL_n(w^*)} \int_{\mathcal{U}} e^{-nK_n(w)} \varphi(w)\, dw,$$

and so

$$F_n(\mathcal{U}) = nL_n(w^*) - \log \int_{\mathcal{U}} e^{-nK_n(w)} \varphi(w)\, dw. \tag{16}$$

Replacing empirical quantities by population quantities, we define the deterministic counterpart

$$F(n) = nL(w^*) - \log Y(n), \qquad Y(n) = \int_{\mathcal{U}} e^{-nK(w)} \varphi(w)\, dw, \tag{17}$$

where we suppress the dependence of $F(n)$ and $Y(n)$ on $\mathcal{U}$.

The stochastic part. The difference between (16) and (17) arises from the fluctuations of K_n around K. Controlling these fluctuations near a singularity is the technically hardest part of the theory. Watanabe (2009) does so by using resolution of singularities to put K into a normal crossing form in which the fluctuations can be analysed. We take the result as a black box and note that under the technical assumptions of Theorem 3.1,

$$F_n(\mathcal{U}) - nL_n(w^*) = F(n) - nL(w^*) + R_n, \tag{18}$$

where R_n converges in distribution to a random variable as $n \to \infty$ and in particular, $R_n = O_p(1)$.

It therefore remains to find the asymptotic behaviour of the deterministic integral $Y(n)$ as $n \to \infty$. This occupies the rest of the derivation.

The density of states. To connect $Y(n)$ to volume scaling, we weight volumes by the prior. Define the **prior-weighted volume** and the **density of states**

$$V_\varphi(\epsilon) = \int_{\{w \in \mathcal{U} \,:\, K(w) < \epsilon\}} \varphi(w)\, dw, \qquad v_\varphi(t) = \int_{\mathcal{U}} \delta\big(t - K(w)\big)\, \varphi(w)\, dw, \tag{19}$$

so that $V_\varphi(\epsilon) = \int_0^\epsilon v_\varphi(t)\, dt$. Here $v_\varphi(t)\, dt$ is the prior mass of the parameters whose excess loss lies in $[t, t + dt]$. With $\varphi \equiv 1$ and $\mathcal{U} = B(w^*)$, V_φ is the volume $V(\epsilon)$ from Definition 1.1, and v_φ is its derivative.

Following through the same proof as in the previous lecture for the expansion of the density of states we see that if the prior is positive and smooth then it gets absorbed into the $b(u)$ once we change variables. Therefore the same volume scaling formula holds albeit with a different multiplicative constant:

$$v_\varphi(t) \sim a\, t^{\lambda - 1} (-\log t)^{\mu - 1} \qquad \text{as } t \to 0. \tag{20}$$

Laplace Transform. Note that we can write

$$e^{-nK(w)} = \int_0^\infty e^{-nt}\,\delta\big(t - K(w)\big)\,dt.$$

Substituting this into $Y(n)$ and exchanging the order of integration,

$$\begin{aligned} Y(n) &= \int_{\mathcal{U}} e^{-nK(w)}\varphi(w)\,dw \\ &= \int_{\mathcal{U}} \left(\int_0^\infty e^{-nt}\,\delta\big(t - K(w)\big)\,dt\right)\varphi(w)\,dw \\ &= \int_0^\infty e^{-nt}\left(\int_{\mathcal{U}} \delta\big(t - K(w)\big)\,\varphi(w)\,dw\right)dt \\ &= \int_0^\infty e^{-nt}\,v_\varphi(t)\,dt. \end{aligned} \tag{21}$$

That is, the deterministic part of the evidence is the Laplace transform of the density of states.

Next, we use the small-t asymptotics of the density of states to find the large-n asymptotics of $Y(n)$. We keep track of the leading-order terms only. The argument below contains all the key ideas. A fully rigorous version (see Ogden (2026)) does not require further ingredients, it only tracks the error terms explicitly, which makes it harder to read.

First we split the integral into small and large t parts by fixing $0 < \delta < 1$ which gives

$$Y(n) = \int_0^\delta e^{-nt}\,v_\varphi(t)\,dt + \int_\delta^\infty e^{-nt}\,v_\varphi(t)\,dt.$$

In the second integral, $e^{-nt} \le e^{-n\delta}$, so

$$0 \le \int_\delta^\infty e^{-nt}\,v_\varphi(t)\,dt \le e^{-n\delta}\int_0^\infty v_\varphi(t)\,dt = e^{-n\delta}\int_{\mathcal{U}} \varphi(w)\,dw \le e^{-n\delta},$$

Therefore as $n \to \infty$

$$Y(n) = \int_0^\delta e^{-nt}\,v_\varphi(t)\,dt + O\big(e^{-n\delta}\big). \tag{22}$$

The tail is exponentially small, and we will see that it is negligible compared with the leading order asymptotics of the integral.

Large n corresponds to small t. Substituting $s = nt$ in (22) gives

$$Y(n) = \frac{1}{n}\int_0^{n\delta} e^{-s}\,v_\varphi\Big(\frac{s}{n}\Big)\,ds + O\big(e^{-n\delta}\big).$$

Note that as $n \to \infty$, $t = s/n \to 0$ so the large-n behaviour of Y depends only on the small-t behaviour of v_φ. Using our asymptotics (20)

$$\begin{aligned} v_\varphi\Big(\frac{s}{n}\Big) &\sim a\left(\frac{s}{n}\right)^{\lambda-1}(\log n - \log s)^{\mu-1} \\ &= a\left(\frac{s}{n}\right)^{\lambda-1}\Big[(\log n)^{\mu-1} + O\big((\log n)^{\mu-2}\big)\Big] && \text{(binomial theorem)} \\ &\sim a\,n^{1-\lambda}\,s^{\lambda-1}(\log n)^{\mu-1} && \text{as } n \to \infty. \end{aligned}$$

The Gamma integral. Substituting this into the integral for $Y(n)$ gives

$$\begin{aligned} Y(n) &\sim \frac{1}{n}\int_0^{n\delta} e^{-s}\, a\, n^{1-\lambda}\, s^{\lambda-1}(\log n)^{\mu-1}\, ds \\ &= a\, n^{-\lambda}(\log n)^{\mu-1}\int_0^{n\delta} e^{-s}\, s^{\lambda-1}\, ds. \end{aligned}$$

The $O(e^{-n\delta})$ term can be dropped because it decays faster than any power of n. As $n \to \infty$ the remaining integral converges to the gamma function $\int_0^\infty e^{-s}\, s^{\lambda-1}\, ds = \Gamma(\lambda)$. Hence

$$Y(n) \sim C\, n^{-\lambda}(\log n)^{\mu-1}, \qquad C = a\,\Gamma(\lambda) > 0, \qquad \text{as } n \to \infty. \tag{23}$$

Assembling the formula. Taking negative logarithms in (23) and substituting into (17),

$$F(n) = nL(w^*) + \lambda \log n - (\mu - 1)\log\log n - \log C + o(1).$$

In particular, $F(n) - nL(w^*) = \lambda \log n - (\mu - 1)\log\log n + O(1)$. Finally, restoring the stochastic part using (18),

$$F_n(\mathcal{U}) = nL_n(w^*) + \lambda_{\mathcal{U}} \log n - (\mu_{\mathcal{U}} - 1)\log\log n + O_p(1),$$

which is Theorem 3.1.

3.3 The generalised BIC

Suppose first that the model is regular at w^*, meaning the Hessian $\nabla^2 L(w^*)$ is non-degenerate, and let $\mathcal{U}$ be a neighbourhood small enough that w^* is the only minimiser of L it contains. Then, as we saw in Lecture 2, $\lambda(w^*) = d/2$ and $\mu(w^*) = 1$, so the $\log\log n$ term in (14) vanishes and the expansion collapses to

$$F_n(\mathcal{U}) = \underbrace{nL_n(w^*)}_{\text{data fit}} + \underbrace{\frac{d}{2}\log n}_{\text{model size}} + O_p(1), \tag{24}$$

in agreement with Exercise 2.7. Discarding the $O_p(1)$ remainder recovers the **Bayesian information criterion (BIC)** of classical statistics. Before explaining what this tells us about deep learning it is useful to understand how the BIC is used in classical statistics.

Recall that statistical models with a lower free energy assign a higher probability to the data. Assuming n is sufficiently large, the BIC is a good approximation to the free energy and we can use it to compare the global free energy of different statistical models (letting $\mathcal{U}$ be all of parameter space). We typically select the model with a smaller BIC which can trade off the model size and how well it fits the data.

In our set up of Bayesian learning, we are interested in where in parameter space the posterior concentrates. Rather than considering different statistical models and comparing their free energy, we have been considering the free energy associated to a particular region of parameter space and Exercise 2.5 shows that the posterior concentrates around regions of low free energy. If we consider two non-degenerate minima of the same model, both pay the same complexity penalty $\frac{d}{2}\log n$, so the BIC can only distinguish them by how well they fit the data.

We can use Theorem 3.1 to generalise the BIC when we relax the regularity assumptions on the Hessian. By truncating (14) after the first two terms we obtain a **generalised BIC**[1]

$$F_n(\mathcal{U}) \;\approx\; \underbrace{nL_n(w^*)}_{\text{data fit}} \;+\; \underbrace{\lambda_{\mathcal{U}} \log n}_{\text{complexity}}. \tag{25}$$

3.4 Internal model selection

Because the free energy is a local property, (25) lets us compare distinct minima of the *same* model. Let $w_1, w_2 \in \mathcal{W}$ be local minimisers of the population loss, and let $\mathcal{U}_1, \mathcal{U}_2$ be sufficiently small neighbourhoods of each. Suppose $L(w_1) = L(w_2)$ and w_1 is the more degenerate of the two so that $\lambda(w_1) < \lambda(w_2)$.[2]

By comparing the generalised BIC for each region we see that

$$F_n(\mathcal{U}_2) - F_n(\mathcal{U}_1) \;\approx\; \big(\lambda(w_2) - \lambda(w_1)\big) \log n \;>\; 0.$$

Since the posterior concentrates on regions of lowest free energy (Exercise 2.5), for all sufficiently large n it favours the neighbourhood of w_1.

This is **internal model selection**. Two parameters fit the data equally well, yet Bayesian updating prefers the more degenerate one. In this sense, Bayesian learning implements Occam's razor automatically.

Here we fixed n and assumed the two minima fit the data equally well. In the next section we relax both assumptions, and find that the trade-off between fit and complexity can reverse as n grows in a Bayesian phase transition.

Exercise 3.1 (Numerical demonstration of the free energy formula)**.** Consider the statistical model underlying the cubically-parameterised loss from Lecture 2. The parameter $\mu \in \mathbb{R}$ indexes the family of normal distributions

$$p(x \mid \mu) = (2\pi)^{-1/2} \exp\left(-\frac{1}{2}(x - \mu^3)^2\right).$$

Suppose the true distribution is

$$q(x) = (2\pi)^{-1/2} \exp\left(-\frac{1}{2}(x - \mu_0^3)^2\right)$$

for some fixed μ_0, and equip the model with a standard normal prior

$$\varphi(\mu) = (2\pi)^{-1/2} \exp\left(-\frac{1}{2}\mu^2\right).$$

[*Note:* The following exercises are intended to be attempted with a numerical programming framework such as Python or Julia, rather than by hand.]

[1] This two-term truncation should not be confused with Watanabe's *widely applicable* Bayesian information criterion (WBIC), which is an estimator of the free energy computed from a tempered posterior (Watanabe, 2013).

[2] Strictly, we also need the empirical losses at the two minima to agree up to $O_p(1)$, which holds for example when $p_{w_1} = p_{w_2}$. Equal population loss alone does not guarantee this: if $p_{w_1} \neq p_{w_2}$, the difference $nL_n(w_2) - nL_n(w_1)$ can fluctuate on the scale of $\sqrt{n}$, which dominates $\log n$.

(a) For $\mu_0 = 0$, numerically compute the free energy F_n for many values of n between $n = 100$ and $n = 10{,}000$. Compare the computed values against the asymptotic estimate $F_n \approx nL_n + \lambda \log n$, using the value of λ you computed for the cubically-parameterised loss at $\mu_0 = 0$ in Lecture 2.

(b) Repeat part (a) with $\mu_0 = 5$.

(c) Repeat part (a) with $\mu_0 = 0.1$. Note that $\mu_0 = 0.1 \neq 0$, so asymptotically $\lambda = 1/2$, as in part (b). However, observe how for moderate n the free energy behaves more like part (a) than part (b). This is another manifestation of the phenomenon observed for the cubically-parameterised loss in Lecture 2, where the finite-ϵ volume $V(\epsilon)$ is affected in a neighbourhood of $\mu_0 = 0$: the effects of the singularity at $\mu = 0$ extend into a neighbourhood around it.

4 Bayesian phase transitions

So far we have looked at the consequence of the free energy formula when considering n to be fixed. However, we also find it interesting to see what it tells us about changes in the posterior as n increases.

For the rest of this section, assume n is sufficiently large that (i) the free energy can be approximated by the first two terms in its expansion, and (ii) $L_n \approx L$ by the law of large numbers, so that the empirical loss can be treated as approximately constant in n.

Consider two disjoint regions $\mathcal{U}, \mathcal{V} \subseteq \mathcal{W}$, containing minimisers u^* and v^* of the population loss L with local learning coefficients $\lambda_{\mathcal{U}}$ and $\lambda_{\mathcal{V}}$ respectively. Applying the free energy formula to each region, the log posterior odds between the two regions satisfy

$$\log \frac{\pi_n(\mathcal{U})}{\pi_n(\mathcal{V})} = F_n(\mathcal{V}) - F_n(\mathcal{U}) \approx n\,\Delta L_n + \Delta\lambda \cdot \log n \tag{26}$$

where $\Delta L_n = L_n(v^*) - L_n(u^*)$ and $\Delta\lambda = \lambda_{\mathcal{V}} - \lambda_{\mathcal{U}}$.

Consider the case when one region is more accurate but less degenerate than the other, i.e.

$$L_n(u^*) > L_n(v^*) \qquad \text{and} \qquad \lambda_{\mathcal{U}} < \lambda_{\mathcal{V}}.$$

Then $\Delta L_n < 0$ and $\Delta\lambda > 0$, so the two terms in (26) have opposite signs. For smaller n, the $\Delta\lambda \cdot \log n$ term dominates so the log posterior odds is positive, meaning the posterior prefers $\mathcal{U}$ (the more degenerate, less accurate region). As n increases, the linear term $n\,\Delta L_n$ eventually dominates the logarithmic term, and the posterior shifts to prefer $\mathcal{V}$ (the less degenerate, more accurate region). Setting the leading terms equal, this change occurs at a critical sample size n^* satisfying

$$\frac{n^*}{\log n^*} \approx \frac{\Delta\lambda}{|\Delta L_n|}. \tag{27}$$

At $n = n^*$, the posterior undergoes a **Bayesian phase transition**: the region that dominates the posterior changes abruptly.

Exercise 4.1 (When do Bayesian phase transitions occur?)**.** We investigate when different terms in the local free energy can cause the posterior preference between two regions to change as the sample size increases.
Suppose $\mathcal{U}, \mathcal{V} \subseteq \mathcal{W}$ are disjoint regions whose local free energies satisfy

$$F_n(\mathcal{U}) \approx nL_{\mathcal{U}} + \lambda_{\mathcal{U}} \log n - (\mu_{\mathcal{U}} - 1) \log\log n + C_{\mathcal{U}},$$

and

$$F_n(\mathcal{V}) \approx nL_{\mathcal{V}} + \lambda_{\mathcal{V}} \log n - (\mu_{\mathcal{V}} - 1) \log\log n + C_{\mathcal{V}}.$$

Write $\Delta L = L_{\mathcal{V}} - L_{\mathcal{U}}$, and define $\Delta\lambda$, $\Delta\mu$, ΔC similarly.

(a) Show that

$$\log \frac{\pi_n(\mathcal{U})}{\pi_n(\mathcal{V})} \approx n\Delta L + \Delta\lambda \log n - \Delta\mu \log\log n + \Delta C.$$

(b) Suppose the two regions have the same loss, $\Delta L = 0$, but $\lambda_{\mathcal{U}} < \lambda_{\mathcal{V}}$.

(i) Ignoring the $\log\log n$ and constant terms, can a phase transition occur?

(ii) Now retain the $\log\log n$ term and suppose $\Delta C = 0$. Explain why a phase transition can only occur if $\Delta\mu > 0$.

(iii) Derive a necessary and sufficient condition, in terms of $\frac{\Delta\mu}{\Delta\lambda}$, for a phase transition to occur.

(iv) When this condition holds, how many critical sample sizes n^* are there? Are they all likely to correspond to phase transitions in practice?

Exercise 4.2 (Exploring phase transitions)**.** We show how *phase transitions* can occur when different local free energies change at different rates.

(a) Suppose we partition the overall parameter space into a disjoint union $\mathcal{W} = \mathcal{U} \sqcup \mathcal{V}$. Denote the overall free energy by F_n and the local free energies by $F_n(\mathcal{U})$ and $F_n(\mathcal{V})$, respectively. Show that the relationship

$$F_n = -\log(e^{-F_n(\mathcal{U})} + e^{-F_n(\mathcal{V})})$$

holds.

(b) Suppose $F_n(\mathcal{U}) \approx 0.3n + 20\log n$ and $F_n(\mathcal{V}) \approx 0.5n + 2\log n$. Plot (using a computer) the overall free energy F_n. Compare against $\min\{F_n(\mathcal{U}), F_n(\mathcal{V})\}$. What happens around $n = 570$?

(c) In statistical physics, a phase transition is traditionally defined as a discontinuity (or rapid change, for finite-size systems) in the derivatives of the free energy. Plot $\frac{d}{dn}F_n$ and explain why this justifies calling the phenomenon from (b) a phase transition.

(d) Change the coefficients of the terms in (b). How does this change when a phase transition occurs? Does a phase transition always occur?

References and further reading

Introductions to the free energy formula. For alternative introductions to the material in this lecture, see the following.

- Carroll (2023) "Distilling singular learning theory," a LessWrong sequence introducing Watanabe's free energy formula and discussing an example of a Bayesian phase transition in a small neural network.
- Lecture recordings from the *SLT & Alignment Summit, 2023*. In particular:
  - See the "SLT Low Road" lectures (Lau and Chen, 2023) for an outline of the derivation of Watanabe's free energy formula.
  - See the "SLT High Road" lectures (Murfet and Carroll, 2023) for a discussion of the free energy formula's implications including Bayesian phase transitions.
- Lau (2025, Chapter 2, "Singular Learning Theory Background"), a self-contained technical introduction to the main results of SLT with illustrative examples.

See Furman (2024) for a list of mathematical exercises. There is some overlap with exercises included in these notes, but there are also several additional exercises.

For the full derivation of the free energy formula, including the stochastic part that we treated as a black box, see Watanabe's two research monographs.

- Watanabe (2009) "Algebraic Geometry and Statistical Learning Theory," colloquially known as "the grey book." Derives the free energy formula in the realisable case along with other results concerning generalisation properties of Bayesian inference and maximum a posteriori inference.
- Watanabe (2018) "Mathematical Theory of Bayesian Statistics," colloquially known as "the green book." An alternative presentation of the free energy formula generalised to the non-realisable case, among other results. Compared to the grey book, the green book has an updated presentation of the main results, but does not contain all of the details of the proofs of the main results from the grey book.

Watanabe offers a set of lecture slides (Watanabe, 2023) which may serve as a useful guide to the "big picture" while working through the details in the books.

Bayesian phase transitions and stagewise development. As we have discussed, Watanabe's free energy formula suggests Bayesian deep learning should undergo Bayesian phase transitions under certain conditions. Carroll (2021) studied Bayesian phase transitions in small ReLU networks, and Chen et al. (2023) studied Bayesian phase transitions in a small feature autoencoder (the "toy model of superposition" from Elhage et al., 2022).

In practice, we use stochastic gradient-based optimisation, rather than Bayesian learning, to train neural networks. However, the Bayesian case serves as a model system from which we can derive empirically testable predictions. Chen et al. (2023) formulate the *Bayesian antecedent hypothesis*, the empirical conjecture that observed phase transitions in trained neural networks correspond to Bayesian phase transitions modelled by Watanabe's free

energy formula. Chen et al. (2023) study such *dynamical phase transitions* in their toy autoencoder and observe a temporal correspondence between phase changes and estimated LLC increases consistent with the free energy formula.

Wang et al. (2024); Hoogland et al. (2025) scale this methodology to transformers trained on natural language, finding similar *stagewise development* phenomena with changes in behaviour and internal structure accompanied by LLC changes. Panickssery and Vaintrob (2023); Hoogland et al. (2025); Carroll et al. (2025); Urdshals and Urdshals (2025) also study developmental stages in transformers trained on synthetic data. Elliott et al. (2026) extends this study to a case of goal misgeneralisation in deep reinforcement learning, generalising Watanabe's free energy formula from Bayesian inference to a non-stationary energy-based inference setting.

References.

Liam Carroll. *Phase Transitions in Neural Networks*. Master's thesis, School of Mathematics and Statistics, the University of Melbourne, 2021. Access via Daniel Murfet. Cited on page 18.

Liam Carroll. Distilling singular learning theory. 2023. Access via LessWrong. Cited on page 18.

Liam Carroll, Jesse Hoogland, Matthew Farrugia-Roberts, and Daniel Murfet. Dynamics of transient structure in in-context linear regression transformers. 2025. Preprint arXiv:2501.17745 [cs.LG]. Cited on page 19.

Zhongtian Chen, Edmund Lau, Jake Mendel, Susan Wei, and Daniel Murfet. Dynamical versus Bayesian phase transitions in a toy model of superposition. 2023. Preprint arXiv:2310.06301 [cs.LG]. Cited on pages 18 and 19.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. Cited on page 18.

Chris Elliott, Einar Urdshals, David Quarel, Matthew Farrugia-Roberts, and Daniel Murfet. Stagewise reinforcement learning and the geometry of the regret landscape. 2026. Preprint arXiv:2601.07524 [cs.LG]. Cited on page 19.

Zach Furman. Singular learning theory: Exercises. 2024. Access via LessWrong. Cited on page 18.

Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. Loss landscape degeneracy and stagewise development in transformers. *Transactions on Machine Learning Research*, 2025. Cited on page 19.

Edmund Lau. *A Singular Perspective on Machine Learning*. Ph.D. thesis, School of Mathematics and Statistics, the University of Melbourne, 2025. Cited on pages 10 and 18.

Edmund Lau and Zhongtian Chen. Singular learning theory: The low road. Lecture series at the SLT & Alignment Summit, 2023. Access via YouTube. Cited on page 18.

Daniel Murfet and Liam Carroll. Singular learning theory: The high road. Lecture series at the SLT & Alignment Summit, 2023. Access via YouTube. Cited on page 18.

Kai Ogden. Bayesian Free Energy and the Local Learning Coefficient in Singular Learning Theory. 2026. Access via web. Cited on page 12.

Nina Panickssery and Dmitry Vaintrob. Investigating the learning coefficient of modular addition. 2023. Access via AI Alignment Forum. Cited on page 19.

Einar Urdshals and Jasmina Urdshals. Structure development in list-sorting transformers. 2025. Preprint arXiv:2501.18666 [cs.LG]. Cited on page 19.

George Wang, Matthew Farrugia-Roberts, Jesse Hoogland, Liam Carroll, Susan Wei, and Daniel Murfet. Loss landscape geometry reveals stagewise development of transformers. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*. 2024. Access via OpenReview. Cited on page 19.

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009. Cited on pages 6, 10, 11, and 18.

Sumio Watanabe. A widely applicable Bayesian information criterion. *The Journal of Machine Learning Research*, 14(1):867–897, 2013. Cited on page 14.

Sumio Watanabe. *Mathematical Theory of Bayesian Statistics*. Chapman and Hall/CRC, 2018. Cited on pages 6, 10, and 18.

Sumio Watanabe. Singular learning theory, parts (1) and (2). 2023. Access via Sumio Watanabe (part 1, part 2). Cited on page 18.

A Solutions

Solution to Exercise 2.1 (Maximum likelihood as MSE minimisation) 22
Solution to Exercise 2.2 (Maximum likelihood as cross entropy minimisation) 22
Solution to Exercise 2.3 (Maximum likelihood as loss minimisation, in general) . . . 23
Solution to Exercise 2.4 (Bayesian posterior as a Gibbs distribution) 23
Solution to Exercise 2.5 (Local posterior mass and local free energy) 24
Solution to Exercise 2.6 (Partition function of partitioned parameter space) 24
Solution to Exercise 2.7 (Deriving the BIC via the Laplace approximation) 24

Solution to Exercise 3.1 (Numerical demonstration of the free energy formula) . . . 27

Solution to Exercise 4.1 (When do Bayesian phase transitions occur?) 28
Solution to Exercise 4.2 (Exploring phase transitions) 28

A.1 Solutions to exercises from Section 2

Solution to Exercise 2.1 (Maximum likelihood as MSE minimisation)**.**

(a) Since labels are sampled independently, the likelihood factorises as a product of Gaussian densities:

$$\begin{aligned} p(Y_n \mid X_n, w) &= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}, w) \\ &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{\|y^{(i)} - f_w(x^{(i)})\|^2}{2}\right). \end{aligned}$$

(b) Taking the negative logarithm of the above expression,

$$\begin{aligned} -\log p(Y_n \mid X_n, w) &= -\sum_{i=1}^{n} \left[-\frac{m}{2}\log(2\pi) - \frac{\|y^{(i)} - f_w(x^{(i)})\|^2}{2}\right] \\ &= \frac{nm}{2}\log(2\pi) + \frac{1}{2}\sum_{i=1}^{n} \|y^{(i)} - f_w(x^{(i)})\|^2. \end{aligned}$$

(c) From the mean squared error definition (3),

$$-\log p(Y_n \mid X_n, w) = \frac{nm}{2}\log(2\pi) + \frac{n}{2}L_n(w).$$

The right-hand side is a strictly increasing affine function of $L_n(w)$ (with positive coefficient $n/2$), and the additive constant $\frac{nm}{2}\log(2\pi)$ is independent of w. Therefore,

$$\arg\max_{w\in\mathcal{W}} p(Y_n \mid X_n, w) = \arg\min_{w\in\mathcal{W}} \left(-\log p(Y_n \mid X_n, w)\right) = \arg\min_{w\in\mathcal{W}} L_n(w). \qquad \square$$

Solution to Exercise 2.2 (Maximum likelihood as cross entropy minimisation)**.**

(a) Since labels are sampled independently, the likelihood factorises as

$$p(C_n \mid X_n, w) = \prod_{i=1}^{n} p(c^{(i)} \mid x^{(i)}, w) = \prod_{i=1}^{n} f_w(c^{(i)} \mid x^{(i)}).$$

(b) Taking the negative logarithm,

$$-\log p(C_n \mid X_n, w) = -\sum_{i=1}^{n} \log f_w(c^{(i)} \mid x^{(i)}).$$

(c) Since $y^{(i)} = \delta_{c^{(i)}}$ is the Dirac distribution on class $c^{(i)}$, we have $y^{(i)}(j) = 1$ if $j = c^{(i)}$ and $y^{(i)}(j) = 0$ otherwise. Substituting into the cross entropy loss (2),

$$L_{x^{(i)},y^{(i)}}(w) = -\sum_{j=1}^{m} y^{(i)}(j)\log f_w(j \mid x^{(i)}) = -\log f_w(c^{(i)} \mid x^{(i)}).$$

Therefore the empirical loss is

$$L_n(w) = \frac{1}{n}\sum_{i=1}^{n} L_{x^{(i)},y^{(i)}}(w) = -\frac{1}{n}\sum_{i=1}^{n} \log f_w(c^{(i)} \mid x^{(i)}) = \frac{1}{n}\left(-\log p(C_n \mid X_n, w)\right).$$

Since the negative log-likelihood is $n \cdot L_n(w)$, a positive multiple of the empirical loss, we conclude

$$\arg\max_{w\in\mathcal{W}} p(C_n \mid X_n, w) = \arg\min_{w\in\mathcal{W}} L_n(w). \qquad \square$$

Solution to Exercise 2.3 (Maximum likelihood as loss minimisation, in general)**.**

(a) Since labels are sampled independently, the likelihood factorises as

$$p(Y_n \mid X_n, w) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}, w) = \prod_{i=1}^{n} \frac{\exp\bigl(-L_{x^{(i)},y^{(i)}}(w)\bigr)}{Z(x^{(i)}, w)}.$$

Taking the negative logarithm,

$$\begin{aligned} -\log p(Y_n \mid X_n, w) &= -\sum_{i=1}^{n} \left[-L_{x^{(i)},y^{(i)}}(w) - \log Z(x^{(i)}, w)\right] \\ &= \sum_{i=1}^{n} L_{x^{(i)},y^{(i)}}(w) + \sum_{i=1}^{n} \log Z(x^{(i)}, w). \end{aligned}$$

(b) Suppose $L_{x,y}(w) = \ell(f_w(x) - y)$. Making the substitution $u = f_w(x) - z$ (so that $dz = du$) in the definition of the partition function,

$$Z(x, w) = \int_{\mathbb{R}^m} \exp(-\ell(f_w(x) - z))\, dz = \int_{\mathbb{R}^m} \exp(-\ell(u))\, du.$$

The right-hand side is a constant independent of both x and w. Therefore, by part (a), the negative log-likelihood is

$$-\log p(Y_n \mid X_n, w) = \sum_{i=1}^{n} L_{x^{(i)},y^{(i)}}(w) + C = n \cdot L_n(w) + C$$

where $C = n \log \int_{\mathbb{R}^m} \exp(-\ell(u))\, du$ is independent of w. Since the negative log-likelihood is an affine function of $L_n(w)$ with positive coefficient,

$$\arg\max_{w\in\mathcal{W}} p(Y_n \mid X_n, w) = \arg\min_{w\in\mathcal{W}} L_n(w). \qquad \square$$

Solution to Exercise 2.4 (Bayesian posterior as a Gibbs distribution)**.** By definition of the negative log-likelihood,

$$nL_n(w) = -\sum_{i=1}^{n} \log p(y^{(i)} \mid x^{(i)}, w).$$

Exponentiating both sides,

$$\exp\bigl(-nL_n(w)\bigr) = \exp\left(\sum_{i=1}^{n} \log p(y^{(i)} \mid x^{(i)}, w)\right) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}, w).$$

Substituting into the definitions of the posterior (8) and partition function (9) gives the stated expressions. $\square$

Solution to Exercise 2.5 (Local posterior mass and local free energy).

(a) Integrating the posterior (8) over $\mathcal{U}$,

$$\pi_n(\mathcal{U}) = \int_{\mathcal{U}} \pi_n(w)\, dw = \int_{\mathcal{U}} \frac{1}{Z_n}\varphi(w) \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}, w)\, dw = \frac{Z_n(\mathcal{U})}{Z_n}.$$

(b) By part (a),

$$\log \frac{\pi_n(\mathcal{U})}{\pi_n(\mathcal{V})} = \log \frac{Z_n(\mathcal{U})/Z_n}{Z_n(\mathcal{V})/Z_n} = \log Z_n(\mathcal{U}) - \log Z_n(\mathcal{V}) = F_n(\mathcal{V}) - F_n(\mathcal{U}). \qquad \square$$

Solution to Exercise 2.6 (Partition function of partitioned parameter space).

(a) Since $\mathcal{U}_1, \ldots, \mathcal{U}_k$ are disjoint and cover $\mathcal{W}$, the integral over $\mathcal{W}$ decomposes as a sum of integrals over each $\mathcal{U}_j$:

$$\begin{aligned} Z_n &= \int_{\mathcal{W}} \varphi(w) \exp\bigl(-nL_n(w)\bigr)\, dw \\ &= \sum_{j=1}^{k} \int_{\mathcal{U}_j} \varphi(w) \exp\bigl(-nL_n(w)\bigr)\, dw = \sum_{j=1}^{k} Z_n(\mathcal{U}_j). \end{aligned}$$

By definition, $Z_n(\mathcal{U}_j) = e^{-F_n(\mathcal{U}_j)}$, so $Z_n = \sum_{j=1}^{k} e^{-F_n(\mathcal{U}_j)}$. Taking $-\log$ of both sides,

$$F_n = -\log Z_n = -\log \sum_{j=1}^{k} e^{-F_n(\mathcal{U}_j)}.$$

(b) Since $F_n(\mathcal{U}_1) < F_n(\mathcal{U}_j)$ for all $j \neq 1$, the corresponding partition functions satisfy $Z_n(\mathcal{U}_1) > Z_n(\mathcal{U}_j)$ for all $j \neq 1$. The lower bound on Z_n is immediate:

$$Z_n = \sum_{j=1}^{k} Z_n(\mathcal{U}_j) > Z_n(\mathcal{U}_1).$$

For the upper bound, since each $Z_n(\mathcal{U}_j) < Z_n(\mathcal{U}_1)$,

$$Z_n = \sum_{j=1}^{k} Z_n(\mathcal{U}_j) < k \cdot Z_n(\mathcal{U}_1).$$

Taking $-\log$ (which reverses the inequalities) yields

$$F_n(\mathcal{U}_1) - \log k < F_n < F_n(\mathcal{U}_1).$$

The overall free energy is therefore determined by the region with the lowest local free energy, up to a correction of at most $\log k$. In particular, if the gap $F_n(\mathcal{U}_j) - F_n(\mathcal{U}_1)$ grows with n while k remains fixed, then $F_n - F_n(\mathcal{U}_1) \to 0$. $\square$

Solution to Exercise 2.7 (Deriving the BIC via the Laplace approximation).

(a) Since $\hat{w}_n$ is a minimiser of L_n in the interior of $\mathcal{U}$, we have $\nabla L_n(\hat{w}_n) = 0$. Taylor expanding about $\hat{w}_n$ to second order,

$$L_n(w) \approx L_n(\hat{w}_n) + \frac{1}{2}(w - \hat{w}_n)^\top H_n (w - \hat{w}_n),$$

and to zeroth order, $\varphi(w) \approx \varphi(\hat{w}_n)$. Substituting both into $Z_n(\mathcal{U}) = \int_{\mathcal{U}} \varphi(w)\, e^{-nL_n(w)}\, dw$ (Exercise 2.4) and pulling the constants out of the integral gives

$$Z_n(\mathcal{U}) \approx \varphi(\hat{w}_n)\, e^{-nL_n(\hat{w}_n)} \int_{\mathcal{U}} \exp\Big(-\frac{n}{2}(w - \hat{w}_n)^\top H_n (w - \hat{w}_n)\Big)\, dw.$$

(b) Let $\sigma > 0$ be a lower bound on the smallest eigenvalue of H_n for large n, so that $(w - \hat{w}_n)^\top H_n (w - \hat{w}_n) \geq \sigma \|w - \hat{w}_n\|^2$. Since $\hat{w}_n$ lies in the interior of $\mathcal{U}$, for large n some ball of fixed radius $r > 0$ about $\hat{w}_n$ is contained in $\mathcal{U}$. Outside this ball the integrand is at most $e^{-n\sigma r^2/2}$, and its integral over $\mathbb{R}^d \setminus \mathcal{U}$ is exponentially small in n. The error from replacing $\mathcal{U}$ by $\mathbb{R}^d$ is therefore exponentially small.

To evaluate the integral over $\mathbb{R}^d$, substitute $u = \sqrt{n}\,(w - \hat{w}_n)$, so that $dw = n^{-d/2}\, du$:

$$\int_{\mathbb{R}^d} \exp\Big(-\frac{n}{2}(w - \hat{w}_n)^\top H_n (w - \hat{w}_n)\Big)\, dw = n^{-d/2} \int_{\mathbb{R}^d} \exp\Big(-\frac{1}{2} u^\top H_n u\Big)\, du.$$

Write $H_n = Q^\top \Lambda Q$ with Q orthogonal and $\Lambda = \operatorname{diag}(\sigma_1, \ldots, \sigma_d)$, where each $\sigma_j > 0$. The substitution $v = Qu$ has unit Jacobian, and the integral factorises into one-dimensional Gaussian integrals:

$$\int_{\mathbb{R}^d} \exp\Big(-\frac{1}{2} u^\top H_n u\Big)\, du = \prod_{j=1}^{d} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sigma_j v_j^2}\, dv_j = \prod_{j=1}^{d} \sqrt{\frac{2\pi}{\sigma_j}} = (2\pi)^{d/2} (\det H_n)^{-1/2}.$$

Combining,

$$\int_{\mathcal{U}} \exp\Big(-\frac{n}{2}(w - \hat{w}_n)^\top H_n (w - \hat{w}_n)\Big)\, dw \approx \left(\frac{2\pi}{n}\right)^{d/2} (\det H_n)^{-1/2}.$$

(c) By parts (a) and (b),

$$Z_n(\mathcal{U}) \approx \varphi(\hat{w}_n)\, e^{-nL_n(\hat{w}_n)} \left(\frac{2\pi}{n}\right)^{d/2} (\det H_n)^{-1/2}.$$

Taking $F_n(\mathcal{U}) = -\log Z_n(\mathcal{U})$ gives

$$F_n(\mathcal{U}) = nL_n(\hat{w}_n) + \frac{d}{2}\log n + \frac{1}{2}\log\det H_n - \frac{d}{2}\log 2\pi - \log\varphi(\hat{w}_n) + o_p(1).$$

For the second equality, consider the last three terms. The term $\frac{d}{2}\log 2\pi$ is a constant. Since $H_n \to H$ and the determinant is continuous, $\log\det H_n \to \log\det H$, which is finite because H is positive definite. Since φ is smooth and positive and $\hat{w}_n \to w^*$, $\log\varphi(\hat{w}_n) \to \log\varphi(w^*)$, which is finite. All three terms therefore remain bounded as $n \to \infty$, so

$$F_n(\mathcal{U}) = nL_n(\hat{w}_n) + \frac{d}{2}\log n + O_p(1).$$

[*Note:* Under the further standard assumption $\sqrt{n}\,(\hat{w}_n - w^*) = O_p(1)$, a second-order Taylor expansion shows $nL_n(w^*) - nL_n(\hat{w}_n) = O_p(1)$ so the first term can be replaced with $nL_n(w^*)$.]

(d) The quadratic approximation of the loss relies on the Hessian at the minimum being positive definite, so that the integrand is approximately Gaussian near $\hat{w}_n$. If $\nabla^2 L(w^*)$ has a zero eigenvalue, the quadratic term in part (a) does not confine the integrand along the corresponding degenerate directions. The discarded higher-order terms then dominate there, and the approximation no longer applies. □

A.2 Solutions to exercises from Section 3

Solution to Exercise 3.1 (Numerical demonstration of the free energy formula)**.** This exercise is computational. We outline the procedure and key observations.

The model is $p(x \mid \mu) = (2\pi)^{-1/2} \exp\bigl(-\frac{1}{2}(x-\mu^3)^2\bigr)$ with prior $\varphi(\mu) = (2\pi)^{-1/2} \exp(-\mu^2/2)$ and true distribution $q(x) = (2\pi)^{-1/2} \exp\bigl(-\frac{1}{2}(x-\mu_0^3)^2\bigr)$.

For each n, sample $x^{(1)}, \ldots, x^{(n)} \sim q$ and compute

$$Z_n = \int_{-\infty}^{\infty} \varphi(\mu) \prod_{i=1}^{n} p(x^{(i)} \mid \mu)\, d\mu,$$

then set $F_n = -\log Z_n$.

The empirical loss is

$$L_n(\mu) = -\frac{1}{n} \sum_{i=1}^{n} \log p(x^{(i)} \mid \mu) = \frac{1}{2} \log(2\pi) + \frac{1}{2n} \sum_{i=1}^{n} (x^{(i)} - \mu^3)^2.$$

(a) For $\mu_0 = 0$, the LLC is $\lambda = 1/6$ (Lecture 2). Plot F_n against n and compare with the asymptotic estimate $\hat{F}_n = nL_n(0) + \frac{1}{6}\log n$. For large n, the two curves should agree up to an $O(1)$ offset: $F_n - \hat{F}_n$ should stabilise around a constant.

(b) For $\mu_0 = 5$, the LLC is $\lambda = 1/2$ (regular case). The asymptotic estimate is $\hat{F}_n = nL_n(5) + \frac{1}{2}\log n$.

(c) For $\mu_0 = 0.1$, the asymptotic LLC is $\lambda = 1/2$ (since $\mu_0 \neq 0$). However, the minimum of L is at $\mu = \mu_0 = 0.1$, very close to the singularity at $\mu = 0$ where $\lambda = 1/6$. For moderate n, the posterior is spread broadly enough that it "feels" the nearby singularity, and F_n behaves more like the $\lambda = 1/6$ regime from part (a). Only for sufficiently large n does the posterior concentrate tightly enough around $\mu_0 = 0.1$ for the regular $\lambda = 1/2$ asymptotics to take hold. This demonstrates that the singularity at $\mu = 0$ influences learning even when $\mu_0 \neq 0$, provided μ_0 is close to the singular point. □

A.3 Solutions to exercises from Section 4

Solution to Exercise 4.1 (When do Bayesian phase transitions occur?)**.**

(a) We have that $\log \frac{\pi_n(\mathcal{U})}{\pi_n(\mathcal{V})} = F_n(\mathcal{V}) - F_n(\mathcal{U})$ so substituting the two local free-energy expansions gives

$$\begin{aligned}\log \frac{\pi_n(\mathcal{U})}{\pi_n(\mathcal{V})} &\approx n(L_{\mathcal{V}} - L_{\mathcal{U}}) + (\lambda_{\mathcal{V}} - \lambda_{\mathcal{U}}) \log n - (\mu_{\mathcal{V}} - \mu_{\mathcal{U}}) \log\log n + (C_{\mathcal{V}} - C_{\mathcal{U}}) \\ &= n\Delta L + \Delta\lambda \log n - \Delta\mu \log\log n + \Delta C.\end{aligned}$$

(b) (i) Ignoring the $\log\log n$ and constant terms,

$$\log \frac{\pi_n(\mathcal{U})}{\pi_n(\mathcal{V})} \approx \Delta\lambda \log n.$$

For $n > 1$ this is positive, so the posterior always favours $\mathcal{U}$ and there is no phase transition.

(ii) Retaining the $\log\log n$ term and taking $\Delta C = 0$ gives

$$\log \frac{\pi_n(\mathcal{U})}{\pi_n(\mathcal{V})} \approx \Delta\lambda \log n - \Delta\mu \log\log n.$$

A phase transition can only occur when the log posterior odds changes sign. For large enough n, $\log n > 0$ and $\log\log n > 0$. Since $\Delta\lambda > 0$, the first term is positive so for a phase transition to exist we must have that $-\Delta\mu \log\log n < 0$ so $\Delta\mu > 0$. If $\Delta\mu \leq 0$, both terms favour $\mathcal{U}$, so the log posterior odds cannot change sign.

(iii) Let n^* be the critical sample size at which the phase transition occurs; then

$$\Delta\lambda \log n^* = \Delta\mu \log\log n^*.$$

Since $\Delta\lambda > 0$, divide through by $\Delta\lambda$ and define $r = \frac{\Delta\mu}{\Delta\lambda} > 0$. Writing $x = \log n$, the equation becomes $x = r \log x$. Define $g(x) = x - r\log x$ and notice that $g(1) = 1$ and $g \to \infty$ as $x \to \infty$. Therefore g has roots if and only if its turning point is below the x-axis. Calculate $g'(x) = 1 - \frac{r}{x}$ so g has a unique minimum at $x = r$. At this point, $g(r) = r - r\log r = r(1 - \log r)$, so we require

$$r(1 - \log r) < 0.$$

Since $r > 0$, this is equivalent to $\frac{\Delta\mu}{\Delta\lambda} > e$. This is the necessary and sufficient condition for the posterior preference to change.

At the boundary $r = e$, the minimum is exactly zero. There is a critical value at which the posterior masses are equal, but g does not change sign, so there is no phase transition.

(iv) By plotting $g(x)$ we see that it has two roots. This suggests two phase transitions $\mathcal{U} \to \mathcal{V} \to \mathcal{U}$ but in practice we would not expect the first transition due to lower order terms playing a role at small sample sizes. □

Solution to Exercise 4.2 (Exploring phase transitions)**.**

(a) By definition, the partition function is

$$Z_n = \int_{\mathcal{W}} \varphi(w) \prod_{i=1}^{n} p(x^{(i)} \mid w)\, dw.$$

Since $\mathcal{W} = \mathcal{U} \sqcup \mathcal{V}$ is a disjoint union, the integral splits additively:

$$Z_n = \int_{\mathcal{U}} \varphi(w) \prod_{i=1}^{n} p(x^{(i)} \mid w)\, dw + \int_{\mathcal{V}} \varphi(w) \prod_{i=1}^{n} p(x^{(i)} \mid w)\, dw = Z_n(\mathcal{U}) + Z_n(\mathcal{V}).$$

Recalling that $F_n = -\log Z_n$ and similarly for the local free energies,

$$Z_n = e^{-F_n(\mathcal{U})} + e^{-F_n(\mathcal{V})},$$

and taking the negative logarithm gives

$$F_n = -\log Z_n = -\log(e^{-F_n(\mathcal{U})} + e^{-F_n(\mathcal{V})}).$$

(b) The curves F_n and $\min(F_n(\mathcal{U}), F_n(\mathcal{V}))$ are nearly identical; both track whichever local free energy is smaller. However, F_n is smooth whereas $\min(F_n(\mathcal{U}), F_n(\mathcal{V}))$ has a kink at $n \approx 570$, where $F_n(\mathcal{U}) = F_n(\mathcal{V})$. For $n \lesssim 570$ the simpler region $\mathcal{V}$ (lower λ) dominates; for $n \gtrsim 570$ the lower loss region $\mathcal{U}$ (lower loss coefficient) dominates.

(c) Differentiating the log-sum-exp from (a),

$$\frac{d}{dn} F_n = \frac{\frac{d}{dn} F_n(\mathcal{U})\, e^{-F_n(\mathcal{U})} + \frac{d}{dn} F_n(\mathcal{V})\, e^{-F_n(\mathcal{V})}}{e^{-F_n(\mathcal{U})} + e^{-F_n(\mathcal{V})}}.$$

Plotting this reveals a steep step around $n \approx 570$, transitioning from $\frac{d}{dn} F_n(\mathcal{V}) \approx 0.5$ to $\frac{d}{dn} F_n(\mathcal{U}) \approx 0.3$. This rapid change in the derivative of F_n is a smooth approximation of a discontinuity, justifying the term *phase transition*.

(d) Write $F_n(\mathcal{U}) \approx L_{\mathcal{U}}\, n + \lambda_{\mathcal{U}} \log n$ and $F_n(\mathcal{V}) \approx L_{\mathcal{V}}\, n + \lambda_{\mathcal{V}} \log n$, and set $\Delta L = L_{\mathcal{U}} - L_{\mathcal{V}}$ and $\Delta\lambda = \lambda_{\mathcal{U}} - \lambda_{\mathcal{V}}$. The critical sample size n^* where the phase transition occurs satisfies

$$\Delta L \cdot n^* = -\Delta\lambda \cdot \log n^*, \qquad \text{i.e.} \qquad \frac{n^*}{\log n^*} = -\frac{\Delta\lambda}{\Delta L}.$$

Since $n^*/\log n^*$ is positive, a solution exists if and only if $-\Delta\lambda/\Delta L > 0$, i.e. ΔL and $\Delta\lambda$ have *opposite signs*. A phase transition occurs precisely when one region fits better while the other is simpler (ΔL and $\Delta\lambda$ have opposite signs). If both fit and complexity favour the same region, that region dominates for all n and no phase transition occurs.

$n^*/\log n^*$ is monotonically increasing for $n > e$ so for such n, increasing $|\Delta\lambda|$ increases n^* and increasing $|\Delta L|$ decreases n^*. □